ISSN 0515-0361
Volume 30, No. 1 May 2000

ASTIN BULLETIN

The Journal of the ASTIN and AFIR Sections of the International Actuarial Association

CONTENTS

Ceuterick

EDITORIAL POLICY

ASTIN BULLETIN started in 1958 as a journal providing an outlet for actuarial studies in non-life insurance. Since then a well-established non-life methodology has resulted, which is also applicable to other fields of insurance. For that reason *ASTIN BULLETIN* has always published papers written from any quantitative point of view – whether actuarial, econometric, engineering, mathematical, statistical, etc. – attacking theoretical and applied problems in any field faced with elements of insurance and risk. Since the foundation of the AFIR section of IAA, i.e. since 1988, *ASTIN BULLETIN* has opened its editorial policy to include any papers dealing with financial risk.

We especially welcome papers opening up new areas of interest to the international actuarial profession.

ASTIN BULLETIN appears twice a year (May and November), each issue consisting of at least 80 pages.

Details concerning submission of manuscripts are given on the inside back cover.

MEMBERSHIP

ASTIN and AFIR are sections of the International Actuarial Association (IAA). Membership is open automatically to all IAA members and under certain conditions to non-members also. Applications for membership can be made through the National Correspondent or, in the case of countries not represented by a national correspondent, through a member of the Committee of ASTIN.

Members of ASTIN receive *ASTIN BULLETIN* free of charge. As a service of ASTIN to the newly founded section AFIR of IAA, members of AFIR also receive *ASTIN BULLETIN* free of charge.

SUBSCRIPTION AND BACK ISSUES

ASTIN BULLETIN is published and printed for ASTIN by Ceuterick s.a., Brusselsestraat 153, B-3000 Leuven, Belgium – E-mail: info@ceuterick.be, fax: +32 16 20 45 33

All queries and communications concerning subscriptions, including claims and address changes, and concerning back issues should be sent to Ceuterick.

The current subscription or back issue price per volume of 2 issues including postage is BEF 2 500.

Back issues up to issue 10 (= up to publication year 1979) are available for half of the current subscription price.

INDEX TO VOLUMES 1-27

The Cumulative Index to Volumes 1-27 is also published for ASTIN by Ceuterick at the above address and is available for the price of BEF 400.

EDITORIAL CHANGE

Readers will have been delighted to hear of the appointment last year of Henrik Ramlau-Hansen as editor of ASTIN Bulletin in succession to Harry Reid. Henrik has subsequently been promoted to the position of chief executive of his company in Denmark and has felt it necessary to step down as editor. We all congratulate Henrik on his promotion and express our regret that he was not able to have a longer term as editor.

EQUITY AND EXACT CREDIBILITY

By

S. David Promislow and Virginia R. Young

Abstract

We consider an alternative to the usual credibility premium that arises from squared-error loss, namely, a so-called *equitable* credibility premium (Promislow and Young, 1999). We derive formulas for the credibility weight in certain cases and give sufficient conditions for exact credibility.

1. Introduction

When setting premiums for insurance, inequities will necessarily arise when, due to imperfect information, some policyholders are charged more than they should be and others less. By building on the previous work of Promislow (1987, 1991), we deal with the problem of choosing premiums to minimize this inequity (Promislow and Young, 1999). Much of our work parallels classical credibility theory, but in place of the traditional squared-error loss functions, we use the family of entropy loss functions. This is a familiar family that has frequently appeared in the economics literature for the purpose of measuring income inequality. We obtain formulas for the optimal premiums, and in certain cases, we obtain explicit formulas for the best affine approximation to the optimal premiums. A natural question, then, is to ask how good the affine approximations are. A basic result of the classical squared-error approach is that they are often exact. This occurs (given certain regularity conditions) when probability distributions are chosen from the linear exponential family with conjugate priors (Jewell, 1974a,b). The purpose of this note is to investigate conditions of exactness for a particular case of an entropy loss function.

In Section 2, we set our notation and assumptions and briefly review previous work in credibility theory, including some of our work in Promislow and Young (1999). We consider a specific case of our equitable credibility estimator. In Section 3, we study the case for which the equitable credibility premium is constrained to be an affine function of the claim data.

ASTIN BULLETIN, Vol. 30, No. 1, 2000, pp. 3-11

For the special case investigated in this paper, we have an explicit expression for the credibility weight and determine a sufficient condition for exact credibility.

2. BACKGROUND

Assume that the total claims of a policyholder, or risk, in the i^{th} policy period, is a random variable X_i whose distribution depends on $\theta, i = 1, 2, ...,$ in which θ varies across policyholders and may be vector-valued. Assume that the X_i are independent (conditional on θ) and identically distributed. The value of θ completely determines the claim distribution of the policyholder. Assume that the value θ is fixed for a given risk, although it is generally unknown and unobservable. Denote the probability (density) function of θ by $\pi(\theta)$, also called the *structure function* (Bühlmann, 1970).

One goal of credibility theory is to calculate a premium for period $n+1$ of a policyholder, given that the policyholder's claim experience in the first n periods is $\boldsymbol{X}_n = \boldsymbol{x}_n = (x_1, x_2, ..., x_n) \in (\mathbf{R}^+)^n$, or more generally given any information, such as a demographic data. Consider general credibility estimators, denoted by Y, in which Y is a real-valued function on the information given, such as $(\mathbf{R}^+)^n$, if the information is prior claim data. We use a capital letter to denote the credibility premium Y to emphasize that it is a random variable. If we constrain Y to be a linear function of the claim data x, then we write L for Y.

If one knew the value of θ that determines the claim distribution of a policyholder, then $\mathrm{E}(X_{n+1}|\theta)$ would be the most equitable premium for period $n+1$, or more simply $\mathrm{E}(X|\theta)$. Let $\mu(\theta)$ denote $\mathrm{E}(X|\theta)$; also, let μ denote $\mathrm{E}X$. The inequity of any other premium is measured relative to this most equitable premium. A general procedure is to select an appropriate loss (or unfairness) function U and then to choose $Y(\boldsymbol{x}_n)$ to minimize

$$\mathrm{E}U[(Y(\boldsymbol{x}_n), \mu(\theta))].$$

In Bühlmann's classical theory (1967, 1970), the loss function U is taken to be the traditional squared error. That is,

$$U[(Y(\boldsymbol{x}_n), \mu(\theta))] = (Y(\boldsymbol{x}_n) - \mu(\theta))^2. \tag{2.1}$$

The resulting credibility premium is the posterior expected value of the conditional mean

$$Y(\boldsymbol{x}_n) = \int \mathrm{E}(X_{n+1}|\theta)\pi(\theta|\boldsymbol{x}_n)d\theta. \tag{2.2}$$

By restricting the form of the credibility premium L to be a linear combination of prior claims, and by assuming that the claims are conditionally independent and identically distributed, one deduces the credibility estimator

$$L(\mathbf{x}_n) = (1 - Z)\mu + Z\bar{x}, \tag{2.3}$$

in which $\mu = \mathrm{E}X = \mathrm{E}[\mu(\theta)]$ is the overall, or grand, mean; $\bar{x}$ is the sample mean,

$$Z = \frac{n}{n+k}, \tag{2.4}$$

in which

$$k = \frac{\mathrm{E}[\mathrm{Var}(X|\theta)]}{\mathrm{Var}[\mu(\theta)]} \tag{2.5}$$

is the ratio of the expected process variance to the variance of the hypothetical means.

In certain cases, the predictive mean (2.2) is an affine function of the sample mean and, thus, equals the linear credibility estimator (2.3). Jewell (1974a,b) verifies conditions under which this *exact credibility* occurs: Under certain regularity conditions, exact credibility occurs for probability distributions from the linear exponential family when one uses the conjugate prior.

Promislow (1987, 1991) and Promislow and Young (1999) argue that squared error is inappropriate for measuring unfairness, and they justify using the entropy family in its place. Squared error is a function of the *absolute difference* between the charged premium and the true premium, while unfairness should depend on the *relative difference* between these two quantities. For example, we consider an individual who should be charged 1 unit but is actually charged 10 units to be treated more unfairly than an individual who should be charged 1001 units but is actually charged 1010 units.

Promislow (1987, 1991) and Promislow and Young (1999) show that appropriate loss functions to meet this objective are of the form

$$U(Y,\ \mu(\theta)) = \mu(\theta)g(r),$$

in which r denotes the ratio $Y/\mu(\theta)$. That is, loss is expressed as a function of the relative difference, weighted by the true premiums. It is shown, moreover, that the function g should be convex and satisfy $g(1) = 0$. In this paper, we will deal with the case in which $g(r) = r^2 - 1$, which leads to the loss function

$$U_2(Y,\ \mu(\theta)) = \frac{Y^2}{\mu(\theta)} - \mu(\theta) \tag{2.6}$$

In place of g, we could take the function h given by $h(r) = (r-1)^2$. Since h differs additively from g by a multiple of $(r-1)$, it is not difficult to see that there will be no effect on the result when we compute expectations. It is of interest to note that the classical squared error loss can be expressed in a similar form but at the cost of distorting the weights. We can write (2.1) in the form

$$U(Y,\ \mu(\theta)) = \mu(\theta)^2 h(r),$$

in which $h(r) = (r-1)^2$. The weights now are the squares of the true premiums, which give much higher weight than before to the high cost situations. Also, we can also compare this with squared percentage error, where the loss function is

$$U(Y,\ \mu(\theta)) = h(r).$$

In this case, the weights are distorted by being independent of the true cost.

Note that the loss function in (2.6) equals $U_c(Y,\ \mu(\theta))$ from Promislow and Young (1999) in the special case for which $c = 2$. We will restrict our attention to this case for the remainder of this paper. In place of formula (2.2), one now gets an optimal premium of

$$Y_2(\boldsymbol{x}_n) = \frac{\mu}{\mathrm{E}\left(\left(\mathrm{E}_{\theta|\boldsymbol{x}_n}\left[\mu(\theta)^{-1}\right]\right)^{-1}\right)}\left(\mathrm{E}_{\theta|\boldsymbol{x}_n}\left[\mu(\theta)^{-1}\right]\right)^{-1}. \tag{2.7}$$

There is a convenient analogue of formulas (2.3) through (2.5). Indeed, (2.3) holds with Z replaced by z_2 given as follows:

$$z_2 = \frac{n}{n + \dfrac{J}{\mu W(\mu(\theta))}}, \tag{2.8}$$

in which $J = \mathrm{E}\left[\dfrac{X^2}{\mu(\theta)}\right] - \mu = \mathrm{E}\left[\dfrac{\mathrm{Var}(X|\theta)}{\mu(\theta)}\right]$, and $W(A) = \mathrm{E}(A)\ \mathrm{E}(A^{-1}) - 1$, for any positive random variable A (Promislow and Young, 1999). Note that z_2 approaches 1 as n goes to infinity.

The expression in (2.8) is similar to the formula for Z, given by (2.4) and (2.5), with J replacing the expected process variance $\mathrm{E}[\mathrm{Var}(X|\theta)]$ as a measure of the variability in X given a value of θ, and with $\mu W(\mu(\theta))$ replacing the variance of the hypothetical means $\mathrm{Var}(\mu(\theta))$ as a measure of the heterogeneity of the risks. See Promislow and Young (1999) for further discussion of the "variance" measures J and W and for the derivation of z_2.

3. Conditions for exactness

For the linear exponential family and conjugate priors, we derive a sufficient condition for exact credibility. By exact credibility, we mean that the equitable credibility estimator $Y_2(\boldsymbol{x}_n)$ given by (2.7) is an affine function of the sample mean and, therefore, equals the credibility estimator $L_2(\boldsymbol{x}_n) = (1 - z_2)\mu + z_2\bar{x}$, with z_2 given by (2.8).

Suppose $X|\theta$ is distributed according to a distribution from a linear exponential family. Specifically, the pf or pdf of $X|\theta$ is of the form

$$f(x|\theta) = \frac{p(x)e^{-x\theta}}{q(\theta)},$$

for $x \geq 0$ and for θ taking values in an interval (θ_0, θ_1), where $-\infty \leq \theta_0 < \theta_1 \leq \infty$. Note that q is the Laplace transform of p because $q(\theta) = \int_0^\infty p(x)e^{-x\theta}dx$. The conditional mean of $X|\theta$ is given by

$$\mu(\theta) = -\frac{q'(\theta)}{q(\theta)},$$

(Klugman et al., 1998). We concentrate on linear exponential families because if the sample mean is a sufficient statistic for θ and if the support of the pdf of the continuous random variable $X|\theta$ is independent of θ, then the distribution of $X|\theta$ comes from a linear exponential family (Lehmann, 1991, Theorem 5.4).

The natural conjugate prior of θ has the form

$$\pi(\theta) = \frac{\{q(\theta)\}^{-k}e^{-\mu k\theta}}{c(\mu, k)}, \quad \theta_0 < \theta < \theta_1,$$

for some μ and $k > 0$. The value $c(\mu, k)$ is a normalizing constant for given values of μ and k. Assume that $\pi(\theta_0) = \pi(\theta_1)$. It follows that $\mathrm{E}X = \mu$, the posterior density of θ given $\boldsymbol{x}_n$ is of the same form as the prior with $k^* = n + k$ and $\mu^* = \dfrac{\mu k + n\bar{x}}{k + n}$, and the predictive mean equals

$$\mathrm{E}[X_{n+1}|\boldsymbol{x}_n] = \mu^* = \frac{k}{n+k}\mu + \frac{n}{n+k}\bar{x},$$

in which $k = \mathrm{E}[Var(X|\theta)]/Var[\mu(\theta)]$. Thus, we get exact credibility for the predictive mean.

To obtain exact credibility for the equitable estimator Y_2, assume that $\pi(\theta_0) = \pi(\theta_1)$ and that $\pi(\theta_0)\ v(\theta_0) = \pi(\theta_1)\ v(\theta_1)$, in which

$$v(\theta) = -\frac{q(\theta)}{q'(\theta)}$$

is the multiplicative inverse of the conditional mean. We next prove the following result for exact credibility.

Theorem 3.1 *Suppose that $\{f(X|\theta)\}$ is a linear exponential family and that the natural conjugate prior satisfies the regularity conditions on its boundary given above. If v satisfies the differential equation*

$$v'' = av'$$

for some constant a, then the equitable credibility estimator Y_2 is exact. Specifically,

$$Y_2(\boldsymbol{x}) = (1 - z_2)\mu + z_2\bar{x}.$$

Moreover, $z_2 = \dfrac{n}{n + k - a/\mu}$.

Proof:
$\mathrm{E}[\mu(\theta)|\boldsymbol{x}]^{-1} = \int_{\theta_0}^{\theta_1} v(\theta)\pi(\theta|\bar{x})d\theta = c(k^*, \mu^*)^{-1} \int_{\theta_0}^{\theta_1} \nu(\theta)q(\theta)^{-k^*} e^{-a\theta}e^{-(\mu^* k^* - a)\theta}d\theta.$

Note that $\left[q^{-k*}ve^{-a\theta}\right]' = -k^*q^{-k^*-1}q've^{-a\theta} + q^{-k^*}v'e^{-a\theta} - aq^{-k^*}ve^{-a\theta}$. By using the definition of v and the fact that $v' - av$ is a constant, we deduce that

$$\left[q^{-k^*}ve^{-a\theta}\right]' = Kq^{-k^*}ve^{-a\theta},$$

for some constant K.

We next integrate by parts and obtain

$$\mathrm{E}[\mu(\theta)|\boldsymbol{x}]^{-1} = \frac{K}{\mu^*k^* - a}.$$

Since $\mu^*k^* = \mu k + n\bar{x}$, we have that

$$Y_2(\boldsymbol{x}) = \frac{\mu}{\mu(n+k) - a}(\mu k + n\bar{x} - a),$$

and the result follows. □

Remarks:

(1) Note that z_2 will be equal to, greater than, or less than the corresponding Bühlmann credibility weight Z, according as a is zero, positive, or negative.

(2) The possibilities for v are limited. If $a = 0$, then $v(\theta) = c_1\theta + c_2$, for some constants c_1 and c_2. If $a \neq 0$, then $v(\theta) = c_1e^{a\theta} + c_2$, for some constants c_1 and c_2. After the following examples, we determine the functions q and p that correspond to these forms of v.

Example 3.2 (Gamma-Gamma) Let $X|\theta \sim \text{Gamma}(\gamma, \theta)$ with conditional mean $\frac{\gamma}{\theta}$ and conditional variance $\frac{\gamma}{\theta^2}$, in which the shape parameter $\gamma > 0$ is known, and let $\theta \sim \text{Gamma}(\alpha, \beta)$. The differential equation of Theorem 3.1 holds with $a = 0$, $z_2 = \dfrac{n}{n + \dfrac{\alpha - 1}{\gamma}} = Z$, $Y_2(\boldsymbol{x}) = (1 - z_2) \cdot \dfrac{\gamma\beta}{\alpha - 1} + z_2 \cdot \bar{x} = \mathrm{E}[X_{n+1}|\boldsymbol{x}_n]$. □

Example 3.3 (Poisson-Gamma) Let $X|\lambda \sim \text{Poisson}(\lambda)$ with conditional mean λ, and let $\lambda \sim \text{Gamma}(\alpha, \beta)$. To put this in standard form, let $\theta = -\ln(\lambda)$. Then, $v(\theta) = e^{\theta}$, and the differential equation for v holds with $a = 1$. The credibility weight z_2 equals $\dfrac{n}{n + \beta - \beta/\alpha}$ versus $Z = \dfrac{n}{n + \beta}$, and $Y_2(\boldsymbol{x}) = (1 - z_2) \cdot \dfrac{\alpha}{\beta} + z_2 \cdot \bar{x}$. Note that $Z < z_2$, so that the equitable premium Y_2 gives more weight to the policyholder's experience than in the Bühlmann credibility estimator. □

Example 3.4 (Binomial-Beta) Let $X|p \sim \text{Binomial}(r, p)$ with conditional mean rp, and let $p \sim \text{Beta}(\alpha, \beta)$. To put this in standard form, let $\theta = -\ln(p/(1 - p))$. Then, $v(\theta) = (1 + e^{\theta})/r$, and the differential equation for v holds with $a = 1$. The credibility weight z_2 equals $\dfrac{n}{n + \dfrac{(\alpha + \beta)(\alpha - 1)}{r\alpha}}$ versus $Z = \dfrac{n}{n + \dfrac{\alpha + \beta}{r}}$, and $Y_2(\boldsymbol{x}) = (1 - z_2) \cdot \dfrac{r\alpha}{\alpha + \beta} + z_2 \cdot \bar{x}$. Note that $Z < z_2$ because $a > 0$, as in Example 3.3. □

Example 3.5 (Negative Binomial-Beta) Let $X|p \sim \text{Negative Binomial}(r, p)$ with probability function

$$f(x|p) = \binom{r + x - 1}{x} p^r (1 - p)^x, \quad x = 0,\ 1,\ 2,\ \ldots,$$

in which $r > 0$ is known, and let $p \sim \text{Beta}(\alpha, \beta)$. To put this in standard form, let $\theta = -\ln(1 - p)$. Then, $v(\theta) = (e^{\theta} - 1)/r$, and the differential equation for v holds with $a = 1$. The credibility weight z_2 equals $\dfrac{n}{n + \dfrac{(\alpha - 1)(\beta - 1)}{r\beta}}$ versus $Z = \dfrac{n}{n + \dfrac{\alpha - 1}{r}}$, and $Y_2(\boldsymbol{x}) = (1 - z_2) \cdot \dfrac{r\beta}{\alpha - 1} + z_2 \cdot \bar{x}$.

Note that $Z < z_2$ because $a > 0$, as in Examples 3.3 and 3.4. □

It is not always the case for the linear exponential family with conjugate prior, that the equitable premium Y_2 is an affine function of the sample mean $\bar{x}$, as we see in the next example.

Example 3.6 Let $X|\theta$ have pdf $f(x|\theta) = \dfrac{(x+1)\theta^2 e^{-x\theta}}{\theta+1}$, for $x > 0$, a member of the linear exponential family, and let θ have pdf proportional to $\dfrac{\theta^2 e^{-\theta}}{\theta+1}$, for $\theta > 0$, the natural conjugate prior for θ. After some tedious calculation, one finds that for $n = 1$,

$$Y_2(x) \propto c\left[\frac{3}{8b^4} - \frac{3}{4b^3} + \frac{7}{4b^2} - \frac{15}{2b} - e^{2b}\int_1^\infty \frac{e^{-2bu}}{u}du + 32e^{2b}\int_2^\infty \frac{e^{-2bu}}{u}du\right]^{-1},$$

in which $c = \left[\dfrac{1}{4b^3} - \dfrac{1}{2b^2} + \dfrac{3}{2b} - 4e^{2b}\displaystyle\int_1^\infty \frac{e^{-2bu}}{u}du + e^{2b}\int_1^\infty \frac{e^{-2bu}}{u^2}du\right]\mu$, and $b = \dfrac{x+1}{2}$. Via numerical calculation, one can verify that Y_2 is not linear in x. □

Now, we return to the problem of determining which distributions of $X|\theta$ lead to $v'' = av'$, for some constant a. We consider the following cases:

(1) $a = 0 \Rightarrow v(\theta) = c_1\theta + c_2$, for some constants c_1 and c_2, not both 0.
   (a) $c_1 = 0 \Rightarrow q(\theta) = c_3 e^{-\theta/c_2}$, for some constant c_3. Because p is the inverse Laplace transform of q, we have that $p(x)$ is a point mass at $x = 1/c_2$. It follows that $f(x|\theta) = 1$ if $x = 1/c_2$ and 0 otherwise.
   (b) $c_1 \neq 0 \Rightarrow q(\theta) = c_3(c_1\theta + c_2)^{-1/c_2}$, for some constant c_3. It follows that $p(x)$ is proportional to $x^{\gamma-1}e^{-\lambda x}$, in which $\gamma = 1/c_1$ and $\lambda = c_2/c_1$. Thus, $f(x|\theta)$ is proportional to $x^{\gamma-1}e^{-x(\lambda+\theta)}$, from which it follows that $X|\theta$ is distributed according to Gamma(γ, $\lambda + \theta$). In Example 3.2, we saw a special instance of this case in which $\lambda = 0$.

(2) $a \neq 0 \Rightarrow v(\theta) = c_1 e^{a\theta} + c_2$, for some constants c_1 and c_2, not both 0.
   (a) $c_2 = 0 \Rightarrow q(\theta) = c_3\exp\left(\dfrac{1}{ac_1}e^{-a\theta}\right)$, for some constant c_3. It follows that $p(x)$ is proportional to

$$\delta(x) + \frac{1}{ac_1}\delta(x-a) + \frac{1}{2(ac_1)^2}\delta(x-2a) + \ldots + \frac{1}{n!(ac_1)^n}\delta(x-na) + \ldots,$$

   in which δ is the Dirac delta function. Thus, $f(x|\theta) \propto \dfrac{e^{-x\theta}}{(ac_1)^{x/a}(x/a)!}$, for $x = 0, a, 2a, \ldots$. If $a = c_1 = 1$, then we have that $X|\theta$ is distributed according to the Poisson distribution with conditional mean $e^{-\theta}$, as in Example 3.3.

(b) $c_2 \neq 0 \Rightarrow q(\theta) = c_3(c_1 + c_2 e^{-a\theta})^{1/c_2}$, for some constant c_3. It follows that $p(x)$ is proportional to $\sum_{m=0}^{\infty} \binom{1/c_2}{m} (c_1/c_2)^{(1/c_2 - m)} \delta(x - am)$, under appropriate regularity conditions. For example, if $1/c_2 = r$, a positive integer, if $c_1 = c_2$, and if $a = 1$, then we have the binomial distribution, as in Example 3.4. If $1/c_2 = r$, a negative integer, if $c_1 = c_2$, and if $a = 1$, then we have the negative binomial distribution, as in Example 3.5.

We see that in some sense, Examples 3.2 through 3.5 cover the simplest of the interesting cases for the distribution of $X|\theta$ for which the conditions of Theorem 3.1 hold. Also, note from the above discussion that if $\mu(\theta) \geq 0$, then $a \geq 0$ from which it follows that $z_2 \geq Z$.

ACKNOWLEDEGEMENTS

We thank the Casualty Actuarial Society and the Committee for Knowledge Extension Research of the Society of Actuaries for financially supporting our project. We thank the Actuarial Education and Research Fund for providing administrative support for our work. Finally, we thank two anonymous referees for their helpful comments.

REFERENCES

BÜHLMANN, H. (1967) Experience rating and credibility, *ASTIN Bulletin* **4**, 199-207.

BÜHLMANN, H. (1970) *Mathematical Models in Risk Theory*, Springer-Verlag, New York.

JEWELL, W.S. (1974a) Credibility is exact Bayesian for exponential families, *ASTIN Bulletin* **8**, 77-90.

JEWELL, W.S. (1974b) Regularity conditions for exact credibility, *ASTIN Bulletin* **8**, 336-341.

KLUGMAN, S., PANJER, H., and WILLMOT, (1998). *Loss Models: From Data to Decisions*, Wiley, New York.

LEHMANN, E.L. (1991) *Theory of Point Estimation*, Wadsworth, Pacific Grove, California.

PROMISLOW, S.D. (1987) Measurement of equity, *Transactions of the Society of Actuaries* **39**, 215-256.

PROMISLOW, S.D. (1991) An axiomatic characterization of some measures of unfairness, *Journal of Economic Theory* **53**, 345-368.

PROMISLOW, S.D. and YOUNG, V.R. (1999) Equity and credibility, *Scandinavian Actuarial Journal*, to appear.

VENTER, G. (1990) *Credibility, Foundations of Casualty Actuarial Science*, Casualty Actuarial Society, New York.

PROF. V.R. YOUNG
Actuarial Science, School of Business
University of Wisconsin-Madison
975 University Avenue
Madison, Wisconsin 53706-1323
USA

A NOTE ON CHRISTOFIDES' CONJECTURE REGARDING WANG'S PREMIUM PRINCIPLE [1]

BY

WANG JING-LONG

East China Normal University, Shanghai, China

1. INTRODUCTION

Young (1999) discussed the conjecture proposed by Christofides (1998) regarding the premium principle of Wang (1995, 1996). She shows that this conjecture is true for location-scale families and for certain other families, but false in general. In addition Young (1999) states that it remains an open problem to determine under what circumstances Wang's premium principle reduces to the standard deviation (SD) premium principle.

In this paper we will provide further discussion of this problem. We will show that, for a fixed distortion, the natural set on which Wang's premium principle can reduce to the SD premium principle is and only is the union of location-scale families which satisfies some condition. Furthermore, it will be shown that the natural set is and only is a location-scale family if Wang's premium principle can be reduced to the SD premium principle for any distortion.

2. RESULTS

As we all know, the standard deviation premium principle applied to a random variable X gives the premium

$$E(X) + \lambda\sqrt{Var(X)}$$

for some $\lambda > 0$, and Wang's (1995, 1996) premium principle gives the premium

$$H_g(x) = \int_{-\infty}^{0} \{g[S_X(t)] - 1\}dt + \int_{0}^{\infty} g[S_X(t)]dt$$

[1] Project 19831020 Supported by National Natural Science Foundation of China.

where $S_X(t) = P(X > t)$ is the decumulative distribution function (ddf) of X, and the distortion g is a non-decreasing function from [0, 1] onto itself. Suppose that Wang's premium principle reduces to the SD premium principle for a set F of distributions, i.e. for any $X \in F$, we have

$$E(X) + \lambda\sqrt{Var(X)} = H_g(x)$$

Hence for any $X, Y \in F$, we have

$$\frac{H_g(X) - E(X)}{\sqrt{Var(X)}} = \frac{H_g(Y) - E(Y)}{\sqrt{Var(Y)}}$$

Further, such a set is called the natural set for the given g, if for any $X \in F$ and $Y \notin F$, the following condition is satisfied:

$$\frac{H_g(X) - E(X)}{\sqrt{Var(X)}} \neq \frac{H_g(Y) - E(Y)}{\sqrt{Var(Y)}}$$

As Young (1999) shows, Wang's premium principle reduces to the SD premium principle on the location-scale family $\Pi = \{X = \mu + \sigma \cdot Z : \mu \in \mathbf{R}, \sigma > 0\}$, where Z is a random variable. We call Z the underlying distribution of this location-scale family Π. In fact, any a distribution of Π can be regarded as the underlying distribution of Π. Obviously, if F is a natural set, $\Pi \subseteq F$ for $Z \in F$. That means the location-scale family with the underlying distribution being a element of the natural set F for which Christofides conjecture is true is a subset of F.

Christofides (1998) conjectures that for a parametric family of distributions with constant skewness Wang's premium principle reduces to the SD premium principle. Young (1999) shows that this conjecture is false in general. Otherwise, if Wang's premium principle reduces to the SD premium principle for a parametric family of distribution, is this family with constant skewness? It is also not true in general. See the following examples.

Following and example in Young (1999), let X be a random variable have a two-sided exponential distribution with parameters $\alpha = 1$, $\beta = 2.27466$ and $w = 0.1$. Its ddf and skewness are, respectively:

$$S_X(t) = \begin{cases} w + (1 - w)(1 - e^{\beta t}), & t < 0; \\ we^{-\alpha t}, & t \geq 0. \end{cases}$$

$$SkewX = \frac{E\left\{[X - E(X)]^3\right\}}{[Var(X)]^{\frac{3}{2}}} = 1.84166$$

Let the distortion $g(p) = p^{0.5}$. Then

$$\frac{H_g(X) - E(X)}{\sqrt{Var(X)}} = 1.02386$$

Let Y have a Pareto distribution with parameters $\alpha = 43.41704$ and β. Its ddf, expectation, variance and skewness are the follows respectively:

$$S_Y(t) = P(Y > t) = \left(\frac{\beta}{1+\beta}\right)^{\alpha}$$

$$E(Y) = \frac{\beta}{\alpha - 1}, \quad Var(Y) = \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$$

$$Skew(Y) = \frac{2(\alpha+1)\sqrt{\alpha-2}}{\sqrt{\alpha}(\alpha-3)} = 2.146777 \neq Skew(X)$$

Let the distortion be the same as above, then

$$\frac{H_g(Y) - E(Y)}{\sqrt{Var(Y)}} = \left(\frac{\alpha}{\alpha-2}\right)^{\frac{1}{2}} = 1.02386 = \frac{H_g(X) - E(X)}{\sqrt{Var(X)}}$$

Hence, for the distortion $g(p) = p^{0.5}$, Wang's premium principle can reduce to the SD premium principle for the union $\Pi_1 \cup \Pi_2$, where

$$\Pi_1 = \{\mu + \sigma \cdot X \ : \ \mu \in \mathbf{R}, \ \sigma > 0\}$$
$$\Pi_2 = \{\nu + \tau \cdot Y \ : \ \nu \in \mathbf{R}, \ \tau > 0\}$$

are the location-scale families with underlying distributions X and Y respectively. Thus, Wang's premium principle reduces to the SD premium principle for a parametric family of distribution $\Pi_1 \cup \Pi_2$ whose members do not all have the same skewness.

From the preceeding discussion we get the following proposition.

Proposition 1.
For a fixed distortion g, the natural set on which Wang's premium principle reduces to the standard deviation premium principle is and only is the union of location-scale families:

$$\cup_{i \in I} \Pi_i$$

where I is an index set, for any $i \in I$, Π_i is a location-scale family: $\Pi_i = \{\mu_i + \sigma_i \cdot X_i \ : \ \mu_i \in \mathbf{R}, \ \sigma_i > 0\}$, and their underlying distributions $X_i (i \in I)$ satisfy the following condition: for any $i, j \in I$ we have

$$\frac{H_g(X_i) - E(X_i)}{\sqrt{Var(X_i)}} = \frac{H_g(X_j) - E(X_j)}{\sqrt{Var(X_j)}}$$

Furthermore, the natural set is and only is a location-scale family if Wang's premium principle can reduce to the SD premium principle for any distortion.

Proposition 2.
The natural set on which Wang's premium principle can reduce to the standard deviation premium principle for any distortion is and only is a location-scale family.

Proof:
Suppose that for any distortion Wang's premium principle can reduce to the SD premium principle on the set F of distributions. Then for any $X, Y \in F$ and for any distortion g, we have

$$\frac{H_g(X) - E(X)}{\sqrt{Var(X)}} = \frac{H_g(Y) - E(Y)}{\sqrt{Var(Y)}} \tag{1}$$

Let

$$U = \frac{X - E(X)}{\sqrt{Var(X)}}, \quad V = \frac{Y - E(Y)}{\sqrt{Var(Y)}}$$

From equation (1), we have

$$H_g(U) = H_g(V) \tag{2}$$

because H_g is location and scale equivariant. Denote the decumulative distribution functions of U and V by $S_U(t)$ and $S_V(t)$ respectively. Firstly, using proof by contradiction, we will show that $S_U(t) = S_V(t)$ when $t \geq 0$. Assume that there is a $t_0 \geq 0$ so that $S_U(t_0) > S_V(t_0)$. Let $\alpha = S_V(t_0)$, $0 \leq \alpha < 1$.

- Case 1.

Suppose that $\{t \;:\; S_U(t) = \alpha\} \neq \emptyset$
Let

$$u_1 = \inf\{t \;:\; S_U(t) = \alpha\}, \quad v_1 = \inf\{t \;:\; S_V(t) = \alpha\}$$

Because the non-increasing ddf is a right continuous function, $S_U(u_1) = S_V(v_1) = \alpha$, which implies that $v_1 \leq t_0 < u_1$. Let the distortion

$$g(w) = \begin{cases} 1, & \alpha < w \leq 1; \\ 0, & 0 \leq w \leq \alpha. \end{cases}$$

Then $H_g(U) = u_1$, and for $v_1 \in (-\infty, \infty)$, it can be proved that $H_g(V) = v_1$. Hence $H_g(U) \neq H_g(V)$, contradicting equation (2).

- Case 2.

Suppose that $\{t \;:\; S_U(t) = \alpha\} = \emptyset$.
Then there is a u_0 so that $S_U(u_0 - 0) \geq \alpha > S_U(u_0)$ and $S_U(t) > \alpha$ when $t < u_0$. Obviously, $t_0 < u_0$. Let the distortion be the same as above, then $H_g(U) = u_0$, $H_g(V) = v_1$, which implies $H_g(U) \neq H_g(V)$, again contradicting equation (2).

Because a ddf is a non-increasing function, the set consisting of points at which the ddf is discontinuous (that is, $\{t$: ddf isn't continuous at point $t\}$) is either a finite set or a countable set. Therefore, it is proved that $S_U(t) = S_V(t)$ when $t \geq 0$.

Now, we will show $S_U(t) = S_V(t)$ when $t < 0$. Let $\tilde{U} = -U$ and $\tilde{V} = -V$. Then the decumulative distribution functions of $\tilde{U}$ and $\tilde{V}$ are $S_{\tilde{U}}(t) = 1 - S_U(-t-0)$ and $S_{\tilde{V}}(t) = 1 - S_V(-t-0)$ respectively. From equation (2), we have $H_g(\tilde{U}) = H_g(\tilde{V})$. According to the above result, $S_{\tilde{U}}(t) = S_{\tilde{V}}(t)$ when $t \geq 0$, which implies $S_U(t-0) = S_V(t-0)$ when $t \leq 0$. Hence $S_U(t) = S_V(t)$ when $t \leq 0$.

To sum up, $S_U(t) \equiv S_V(t)$. Hence, $U = V$ almost surely. The proposition is proved.

In fact, the assumption of this Proposition can be weakened: that is, the natural set on which Wang's premium principle can reduce to the standard deviation premium principle for any two-step-up distortion is and only is a location-scale family. Is and only is the natural set a location-scale family for any power distortion? This is a subject of future study.

Acknowledgements

I wish to thank the editors and referees for their valuable comments and corrections, which resulted in considerable improvement of my original paper.

References

Christofides, S. (1998) Principle for risk in financial transactions. *Proceedings of the GISG/ASTIN Joint Meeting in Glasgow, Scotland.* October, 1998, **2**, 63-109.

Wang, S.S. (1995) Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics* **17**, 43-54.

Wang, S.S. (1996) Premium calculation by transforming the layer premium density. *ASTIN Bulletin* **26**, 71-92.

Young, V.R. (1999) Discussion of Christofides' conjecture regarding Wang's premium principle. *ASTIN Bulletin* **29**, 191-195.

Wang Jing-Long
Department of Statistics
East China Normal University
3663 Zhongshan Road (Northern)
Shanghai 200062
P.R. China

SOME NOTES ON THE DYNAMICS AND OPTIMAL CONTROL OF STOCHASTIC PENSION FUND MODELS IN CONTINUOUS TIME

BY

ANDREW CAIRNS [1, 2]

Heriot-Watt University, United Kingdom

ABSTRACT

This paper discusses the modelling and control of pension funds.

A continuous-time stochastic pension fund model is proposed in which there are n risky assets plus the risk-free asset as well as randomness in the level of benefit outgo. We consider Markov control strategies which optimise over the contribution rate and over the range of possible asset-allocation strategies.

For a general (not necessarily quadratic) loss function it is shown that the optimal proportions of the fund invested in each of the risky assets remain constant relative to one another. Furthermore, the asset allocation strategy always lies on the capital market line familiar from modern portfolio theory.

A general quadratic loss function is proposed which provides an explicit solution for the optimal contribution and asset-allocation strategies. It is noted that these solutions are not dependent on the level of uncertainty in the level of benefit outgo, suggesting that small schemes should operate in the same way as large ones. The optimal asset-allocation strategy, however, is found to be counterintuitive leading to some discussion of the form of the loss function. Power and exponential loss functions are then investigated and related problems discussed.

The stationary distribution of the process is considered and optimal strategies compared with dynamic control strategies.

Finally there is some discussion of the effects of constraints on contribution and asset-allocation strategies.

KEYWORDS

Continuous time; stochastic differential equation; asset-allocation; contribution strategy; Bellman equation; optimal control; constraints.

[1] E-mail: A.Cairns@ma.hw.ac.uk

[2] WWW: http://www.ma.hw.ac.uk/~andrewc/

1. Introduction

The analysis and control of pension fund dynamics is becoming increasingly important as members start to pay more attention to the security of promised benefits and as sponsoring employers become more concerned about the timing and stability of cashflows.

This paper discusses some current problems in the analysis and control of defined benefit pension funds. Under a pure defined benefit pension fund the benefits payable to an individual member depend only upon his or her salary and length of past service.

The principal alternative to a defined benefit scheme is a defined contribution occupational pension scheme. Here the benefits are defined by the level of contributions which are paid into an individual member's fund or 'pot' and by the investment returns which are achieved over the period up to retirement. Since the pot is used to purchase an annuity at the time of retirement the level of pension is also determined by the annuity rate which prevails at the date of retirement and, in particular, the term structure of interest rates on that date. Generally the rates of contribution by the sponsor and by the member are fixed. All of the investment risk is borne by the member and there is no opportunity for the member to smooth out the effects of adverse investment returns. Existing literature on defined contribution problems typically deals with the case where the terminal utility is a function of the fund size at retirement (for example, see Merton, 1990, Gerber & Shiu, 2000, and Deelstra *et al.*, 1999). The case where the terminal utility is a function of pension purchased at retirement (that is, fund divided by annuity rate) in a stochastic interest-rate environment has been considered by Cairns *et al.* (2000).

Under a defined benefit scheme the sponsoring employer has no ability to vary the timing or amount of the benefits payable. In contrast to this and to a defined contribution scheme the rate at which contributions are paid into the fund are (within limits) very flexible. Typically this flexibility rests fully with the fund sponsor while individual members contribute a fixed percentage of their salaries.

Increasingly, we also see schemes which provide elements of both defined benefit and defined contribution. Most common are schemes which allow for discretionary increases to pensions in payment with the size of the increase depending upon recent investment returns. Other 'hybrid' schemes provide a pension which is equal to the maximum of a defined benefit pension and a defined contribution pension.

Within the pure defined benefit framework there is considerable scope for freedom:

- in how the variable contribution rate should be varied;
- in the choice of asset allocation strategy.

1.1. Contributions

By-and-large, the fund sponsor has considerable freedom in how the contribution rate can be varied. The basic principle underlying how the contribution rate is set is that it should take account of the amount of surplus or deficit (that is, the excess of assets over liabilities). Thus, in some sense, the contribution rate can be reduced during periods of surplus and increased above the normal rate when the scheme is in deficit. The role of the actuary is to take account of the needs of the sponsor and of the members before recommending to what extent surplus or deficit should affect the contribution rate.

The overall level of flexibility may be restricted by the presence of certain constraints:

- There may be a legal requirement to keep the funding level (the asset/liability ratio) above a certain minimum level (the method of calculation of which can take a number of forms). If the funding level drops below this minimum the sponsor may be compelled to make up the deficit immediately.
- Similarly there may be a restriction on the maximum size of the fund. This may require refunds to the sponsor or improvements to the benefits (although, in the latter case, the fund would cease then to be a 'pure' defined benefit scheme).
- The fund sponsor may wish to keep the contribution rate below a certain level (for example, twice the normal rate).
- Regulations or plan rules may prevent refunds to the employer, or perhaps refunds are only permitted when the funding level is sufficiently high.

1.2. Assets

A pension fund will normally fall under the responsibility of a group of *trustees or managers* who must act in the best interests of the fund members. Within this remit they can choose how to invest the assets of the fund. Appropriate investment strategies will take account of:

- prudence;
- requirements to
  - maximise returns;
  - minimise risk;
  - diversify;
  - avoid self-investment;
- immediate cashflow requirements;
- security;
- the tax status of the fund and of the various potential assets.

Besides taking the advice of their fund managers, trustees may also seek the advice of the fund actuary before deciding upon an appropriate strategy. How the funds available should be allocated presents an interesting problem for the actuary. The solution to such a problem must take account of many things:

- the balance between the conflicting interests of the members and the sponsor;
- the expected returns on the various assets and the associated risks and dependencies (both between individual assets and through time);
- the current level of funding;
- constraints on short selling of assets.

1.3. Objectives

For an actuary to set an optimal contribution rate and asset allocation strategy it is necessary to use a well defined objective function with appropriate constraints. Objective functions must be sufficiently precise to avoid ambiguous or non-sensical solutions. For example, the imprecise objective *minimise variance* leads to various outcomes which minimise the variance of the funding level and/or the contribution rate.

Other apparently precise objectives lead to optimal solutions which do not entirely make sense. In such circumstances it may be necessary to revise the objective function.

1.4. Types of model

A basic question which must be answered first is should we use a deterministic or a stochastic model. Deterministic models are adequate for cashflow projections and valuations but little else. Stochastic models, on-the-other-hand, allow us to investigate fully the dynamics of the fund through time and, for example, devise suitable control strategies. Here we consider stochastic models only.

A separate question is whether models should be kept simple or be made very realistic. The answer here depends on the reasons for modelling. In a more academic study we are looking for the major drivers of pension fund dynamics. Simple models allow detailed study of these factors. Often it is possible to derive analytical results which can then be used to provide specific links between causes and effects. A more complex model, on the other hand, may be required if the modeller has in mind a specific pension fund with a very specific benefit structure. As models become more complex we input more and more factors and find that more detail comes in the output from each simulation. It then becomes very difficult to identify why certain effects are evident. However, simple models provide the backup in the analysis of complex models. Such models give pointers to what we should be investigating. Thus we may be able empirically to observe the same links between causes and effects as were found analytically in the simple model. More-often-than-not such comparisons can explain, with ease, the majority of the variation in the dynamics of a complex model.

In some problems the aim may be to devise an optimal control strategy. As we show here it is possible using simple models to derive precisely an optimal control. This then gives us the starting point for further study and optimisation within a more complex model.

This paper has a number of aims. First, it will pull together some recent results in continuous-time pension fund modelling (O'Brien, 1986, 1987, Dufresne, 1990, Boulier *et al.*, 1995, 1996, and Cairns, 1996, 1997). Fresh proofs of these results will be presented as appropriate along with further discussion of their implications. Second, some new avenues will be developed to show how this earlier work can be modified to consider some generalisations and to pull the results closer to current practice. Third, the paper will discuss some open problems.

Within this framework the paper will proceed as follows. Section 2 introduces the continuous-time stochastic model for the dynamics of a pension fund in its most general form which will be used in the majority of the paper.

Section 3 considers dynamic stochastic control of the model by making reference to a value function which discounts exponentially future random values of a quadratic loss function. The section proceeds by looking at various cases both constrained and unconstrained. The advantages and disadvantages of the quadratic loss function are discussed in detail here. Finally, power and exponential loss functions are considered with problems similar to those under the quadratic loss function identified.

In Section 4 we take the longer-term view and consider the stationary distribution of the process (although the distribution of the model nears its stationary form within 10 to 15 years usually). This includes a look at the continuous proportion portfolio insurance approach to asset allocation introduced by Black and Jones (1988) and compares this with a static investment strategy. Section 5 compares the results of dynamic versus stationary optimisation derived in Sections 3 and 4 and shows how sensitive these results are to changes in the control parameters.

Finally Section 6 discusses how the model and value function might be developed in the future to come closer to reality.

2. A GENERAL MODEL

In this paper we consider continuous-time stochastic models for pension fund dynamics which allow for n risky assets and for noise in the level of benefit outgo. The general form of this simple model is:

$$dX(t) = X(t) \cdot d\delta_X(t, X(t)) + c(t) \cdot dt - B \cdot dt - \sigma_b \cdot dZ_b(t) \qquad (1)$$

$$\begin{aligned} \text{where } X(t) &= \text{fund size at } t \\ d\delta_X(t, X(t)) &= \text{instantaneous return on assets between } t \text{ and } t + dt \\ c(t) &= c(t, X(t)) \\ &= \text{contribution rate} \\ B &= \text{expected rate of benefit outgo} \\ \text{and } \sigma_b &= \text{volatility in benefit outgo} \end{aligned}$$

Discrete-time models have been considered by Cairns (1995), Cairns & Parker (1997), Dufresne (1988, 1989, 1990) and Haberman & Sung (1994). Such models have yielded a number of useful analytical results with wider applications. Continuous-time models, which are, in some ways, more idealised, yield further analytical results (for example, see Dufresne, 1990, Boulier *et al.*, 1995, and Cairns, 1996). Similar results can then be sought empirically in discrete-time models.

The contribution rate, $c(t)$, is a predictable process and provides us with one of the means of controlling the dynamics of the pension fund. Dufresne (1990) and Cairns (1996) considered continuous-time models in which the contribution rate was a linear function of the current fund size, $X(t)$. Boulier *et al.* (1995) considered more general forms for $c(t)$ but found that the optimal solution to a simple control problem was that the contribution rate should indeed be linear in $X(t)$. These results are discussed in detail in Sections 3 and 4 of this paper. O'Brien (1987) considered a similar objective function where the contribution rate only was controllable and where there was a stochastic reserve (in contrast to the constant target x_p relative to salary roll used in Section 3 of this paper). He found that the optimal contribution rate was linear in the amount of surplus. However, other aspects of the model used by O'Brien (1987) were unrealistic even for a simple pension scheme, making a fresh start here appropriate.

The other means of control is through the asset-allocation strategy. First we may allow for the possibility of a risk-free asset (or *cash*) which has a value at time t of $R_0(t) = R_0(0)\exp(\delta_0 t)$. There are, in addition, n risky assets, the prices of which (including reinvestment of dividend income) we assume follow correlated geometric brownian motion: that is,

$$\frac{dR_i(t)}{R_i} = d\delta_i(t) = \delta_i \cdot dt + \sum_{j=1}^{n} \sigma_{ij} \cdot dZ_j(t) \tag{2}$$

$$\text{or } d\delta(t) = \delta \cdot dt + S \cdot dZ \tag{3}$$

$$\begin{aligned} \text{where } d\delta(t) &= (d\delta_1(t), \ldots, d\delta_n(t))^T \\ \delta &= (\delta_1, \ldots, \delta_n)^T \\ S &= (\sigma_{ij})_{i,j=1}^{n} \\ dZ &= (dZ_1, \ldots, dZ_n)^T \end{aligned}$$

and $Z(t)$ is standard n-dimensional Brownian motion. We assume that $Z(t)$ and $Z_b(t)$ are independent.

For convenience later on, we define $D = SS^T$ (the instantaneous covariance matrix) and $\lambda = (\lambda_1, \ldots, \lambda_n)^T$ where $\lambda_i = \delta_i - \delta_0$ is the risk premium attached to asset i.

Let us assume that $\delta_i > \delta_0$ for all $i \geq 1$ (that is, investors are rewarded with higher expected returns for taking on some risk). No assumption is made about the level of correlation between the returns on the various stocks including, for example, the benefits (or otherwise) of diversification. The proportion of the assets invested in asset i ($i = 0, 1, \ldots, n$) is denoted by $p_i(t, X(t))$. It follows that $\sum_{i=0}^{n} p_i(t, X(t)) = 1$. In the development below we write $p = p(t, X(t)) = (p_1(t, X(t)), \ldots, p_n(t, X(t)))^T$. The instantaneous rate of return on the fund is then:

$$\left(1 - \sum_{i=1}^{n} p_i\right)\delta_0 \cdot dt + \sum_{i=1}^{n} p_i d\delta_i(t) = \delta_0 \cdot dt + p^T\lambda \cdot dt + p^T S dZ \tag{4}$$

In this paper we will consider a range of constraints on the proportions invested in each asset. These include the possibility that we hold no cash (or a fixed percentage of the fund in cash) and that there shall be no short-selling of assets.

We allow for more than one risky asset for two reasons. First, it allows for a degree of realism without complicating substantially the analysis. Second, the experience of the UK pension funding scene is that pension funds only use cash for short-term liquidity rather than as a serious asset. Instead funds use government bonds (fixed interest and index linked) as low-risk (but non-zero-risk) assets. This situation is modelled in Section 3.3.

3. Optimal dynamic stochastic control

3.1. The general quadratic case

We consider first the case where there is no constraint on the amount invested in cash. Following Boulier *et al.* (1995) we define the value function for a general controlled pension fund process

$$W(t,x)(c,p) = E\left[\int_t^\infty \exp(-\beta s) L(s, c(s, X(s)), X(s)) ds | X(t) = x\right] \qquad (5)$$

Here $\exp(-\beta s)$ is a discount function and $L(s,c,x)$ is a loss function given that at time s, $X(s) = x$. This value function is also a function of the chosen, Markov contribution strategy $c(s, X(s))$ and investment strategy $p(s, X(s))$ which we abbreviate, where appropriate to c and p respectively.

Let $V(t,x) = \inf_{(c,p)} W(t,x)(c,p) = W(t,x)(c^*,p^*)$ assuming that such optimal control strategies c^* and p^* exist. Then $V(t,x)$ satisfies the Hamilton-Jacobi-Bellman equation (for example, see Merton, 1990, Øksendal, 1998, or Fleming & Rishel, 1975):

$$0=\inf_{c,p}\left(e^{-\beta t}L(t,c,x)+V_t+\left[(\delta_0+p^T\lambda)x+c-B\right]V_x+\frac{1}{2}V_{xx}\left(x^2p^TDp+\sigma_b^2\right)\right)$$

$$\begin{aligned}\text{where } V_t &\equiv \partial V/\partial t\\ V_x &\equiv \partial V/\partial x\\ V_{xx} &\equiv \partial^2 V/\partial x^2\end{aligned} \qquad (6)$$

We differentiate the expression in brackets with respect to c and p to find that:

$$\frac{\partial}{\partial c}(\cdot) = e^{-\beta t}L_c + V_x = 0, \qquad \text{where } L_c = \partial L/\partial c \qquad (7)$$

$$\Rightarrow c^*(t,x) = L_c^{-1}\left(-e^{\beta t}V_x\right) \qquad (8)$$

$$\text{and } \frac{\partial}{\partial p}(\cdot) = \lambda x V_x + Dpx^2 V_{xx} = 0 \qquad (9)$$

$$\Rightarrow p^*(t,x) = -\left(\frac{V_x}{xV_{xx}}\right)D^{-1}\lambda \qquad (10)$$

We see from the form of p^* that the amounts invested in each of the risky assets always stay in the same proportion. Thus we may define a special portfolio, A, which is a mixture of assets 1 to n in the same proportions (in market value terms) as $D^{-1}\lambda$. Then for any x we hold a proportion $\tilde{p}(x)$ (which depends upon $V(t,x)$) in portfolio A and $1 - \tilde{p}(x)$ in cash. This result has obvious parallels in modern portfolio theory where the combination here of cash and portfolio A mimics movement along the capital market line. However, here we have not yet specified any form for the loss function $L(t,c,x)$ whereas modern portfolio theory (which works in discrete time) relies upon the use of a quadratic loss function. Further consideration of the model shows that portfolio A (which is efficient in the sense of minimising the value function) is also efficient in the sense of modern portfolio theory: that is, it has the lowest instantaneous volatility for a given rate of return.

Classical portfolio theory has been extended to include liabilities by Wise (1984), Wilkie (1985), Sharpe and Tint (1990) and Keel and Müller (1995). Working in discrete time and using a quadratic loss function Keel and Müller (1995) find that the composition of efficient portfolios can be altered by the inclusion of liabilities: in particular, where liabilities are random and not independent of the asset returns.

The precise form for $V(t,x)$ is, of course, still not yet known: we only have expressions for c^* and p^* involving $V(t,x)$.

It is necessary that the loss function is a strictly convex function of c. This ensures that the inverse of L_c exists. This requirement excludes, for example, downside loss functions which are convex but not strictly convex.

Here we restrict ourselves to the following quadratic loss function:

$$L(t,c,x) = (c-c_m)^2 + 2\rho(c-c_m)(x-x_p) + (k+\rho^2)(x-x_p)^2 \tag{11}$$

where $k \geq 0$.

Thus
$$L_c^{-1}\left(-e^{\beta t}V_x\right) = c_m - \rho(x-x_p) - \frac{1}{2}e^{\beta t}V_x \tag{12}$$

(that is, if $c = c_m - \rho(x-x_p) - \exp(\beta t)V_x/2$ we have $L_c(t,c,x) = -\exp(\beta t)V_x$). A special case of this loss function is the one suggested by Haberman and Sung (1994) (in a discrete-time framework).

We apply this to the Hamilton-Jacobi-Bellman equation to give:

$$\begin{aligned} 0 = e^{-\beta t}&\left[\left(-\frac{1}{2}e^{\beta t}V_x\right)^2 + k(x-x_p)^2\right] + V_t + (\delta_0 x - B)V_x - \lambda^T D^{-1}\frac{V_x}{xV_{xx}}\lambda x V_x \\ &+ \left(c_m - \rho(x-x_p) - \frac{1}{2}e^{\beta t}V_x\right) + \frac{1}{2}V_{xx}\left[x^2\left(\frac{V_x}{xV_{xx}}\right)^2\lambda^T D^{-1}DD^{-1}\lambda + \sigma_b^2\right] \end{aligned} \tag{13}$$

Given the form of the objective function (Markov and time-homogeneous) it is clear that the optimal strategies c^* and p^* depend only upon x and not upon t. Thus $V(t,x)$ will be of the form $e^{-\beta t}F(x)$ and therefore:

$$\begin{aligned} 0 = &\frac{1}{4}F_x^2 + k(x-x_p)^2 - \beta F + (\delta_0 x - B)F_x - \lambda^T D^{-1}\lambda\frac{F_x^2}{F_{xx}} \\ &+ \left(c_m - \rho(x-x_p) - \frac{1}{2}F_x\right)F_x + \frac{1}{2}F_{xx}\left[\frac{F_x^2}{F_{xx}^2}\lambda^T D^{-1}\lambda + \sigma_b^2\right] \end{aligned} \tag{14}$$

Try $F(x) = Px^2 + Qx + R$, and write $\varepsilon = \lambda^T D^{-1}\lambda$. Then:

$$0 = -\frac{1}{4}(2Px+Q)^2 + k(x-x_p)^2 - \beta(Px^2+Qx+R) + (\delta_0 x - B)(2Px+Q)$$
$$-\frac{1}{2}\varepsilon\frac{(2Px+Q)^2}{2P} + \left(c_m - \rho(x-x_p)\right)(2Px+Q) + P\sigma_b^2 \tag{15}$$
$$\Rightarrow 0 = x^2\left[-P^2 + k - \beta P + 2P\delta_0 - P\varepsilon - 2\rho P\right]$$
$$+ x\left[-PQ - 2kx_p - \beta Q - 2PB + Q\delta_0 - Q\varepsilon - \rho Q + 2P(c_m + \rho x_p)\right]$$
$$+ \left[-\frac{1}{4}Q^2 + kx_p^2 - \beta R - BQ - \frac{Q^2\varepsilon}{4P} + (c_m + \rho x_p)Q + P\sigma_b^2\right] \tag{16}$$

Define $\hat{P} = 2\delta_0 - \beta - \varepsilon - 2\rho$. Then we find that:

$$P(k) = \frac{\hat{P} + \sqrt{\hat{P}^2 + 4k}}{2} \tag{17}$$
$$Q(k) = \frac{2\left[P(k)(B - c_m - \rho x_p) + kx_p\right]}{-P(k) + \delta_0 - \beta - \varepsilon - \rho} \tag{18}$$
$$R(k) = \frac{1}{\beta}\left[-\frac{1}{4}Q(k)^2 + kx_p^2 - BQ(k) - \frac{Q(k)^2\varepsilon}{4P} + (c_m + \rho x_p)Q(k) + P(k)\sigma_b^2\right] \tag{19}$$

This is an admissible solution provided $\hat{P} > 0$.
We find then that:

$$c^*(x) = c_m - \rho(x - x_p) - \frac{1}{2}(2P(k)x + Q(k)) \tag{20}$$
$$\text{or } c^*(x) = c_0^* - c_1^* x \tag{21}$$
$$p^*(x) = -\frac{(2P(k)x + Q(k))}{2P(k)x}D^{-1}\lambda \tag{22}$$
$$\text{or } p^*(x) = \frac{p_0^* + p_1^* x}{x} \tag{23}$$

where p_0^* and p_1^* are both $n \times 1$ vectors which are proportional to $D^{-1}\lambda$.

Note that when $x = -Q(k)/2P(k)$, $p^*(x) = 0$: that is, we are invested entirely in the risk-free asset. Furthermore if a portfolio, A, is synthesised from the n risky assets in the proportions $D^{-1}\lambda$ as described earlier, then, given a funding level of x, we should hold a proportion of the fund:

$$\tilde{p}(x) = e^T p^*(x) = -\frac{2P(k)x + Q(k)}{2P(k)x}e^T D^{-1}\lambda \tag{24}$$

in portfolio A and $1 - \tilde{p}(x)$ in cash. ($e^T = (1, \ldots, 1)$ is the unit vector.)

We can also note that $F(x)$ is minimised at $x = -Q(k)/2P(k)$, which we will denote by x_{min} say. As discussed in Section 3.6 this presents, to a certain extent, a barrier through which it is difficult for the funding level, $X(t)$, to pass. Depending upon the relationship between c_m, k and x_p this could take the form of a ceiling or a floor.

It is important to note that $P(k)$ and $Q(k)$ do not depend upon σ_b. It follows, therefore, that the optimal control strategy (both contributions and investments) do not depend upon σ_b. Thus, demographic variability is a factor which affects the value function $V(t,x)$ only and we should treat small funds in the same way as large funds.

It is also important to note that the precise proportions of each asset held in portfolio A do not depend upon the form of the loss function, nor does it depend upon σ_b.

Remark

The non-linear ordinary differential equation (14) is subject to the boundary condition $0 \leq F(x)$ for all x. We have two degrees of freedom in how we solve this equation. Numerical work suggests that there are also solutions to (14) which either have singularities (which we regard as an inadmissible solution) or which are asymptotically linear as $x \rightarrow \pm\infty$. Now if $F(x) \sim a + bx$ as $x \rightarrow +\infty$, $c^*(x) \sim c_m - b$ as $x \rightarrow +\infty$. With such a solution we may find that $X(t)$ will drift off to infinity. This drift, however, is countered by the asset-allocation strategy which is quite extreme:

- As $X(t)$ gets very large the fund goes very long in cash and very short in risky assets. This ensures that there is a very inefficient strategy which more-or-less throws away money in order to get back to the target funding level x_p.
- As $X(t)$ gets very small the fund goes very long in risky assets and very short in cash to get a high expected return to help us get back to a better funded position as quickly as possible.

In the quadratic-$F(x)$ case these problems with the asset-allocation strategy also apply but they are much less extreme. Furthermore, the optimal contribution rate is a linear function of $X(t)$. It is a necessary condition for stationarity that the contribution rate is at least linear. (Note, however, that linearity is sufficient only when the slope c_1^* is greater than a certain minimum level described later in this paper.)

Thus we can reasonably put in the further boundary condition that $F(x)/x^2 \rightarrow$ constant as $x \rightarrow \pm\infty$.

3.2. Constraints on cash

We have up until now assumed that the amount of money invested in cash could vary without bound. Here we go to the other extreme and assume that we invest a proportion p_m of the fund in risky assets and $1 - p_m$, in cash, where p_m is fixed. It is reasonable that $p_m < 1$ allowing for a small but fixed amount in cash to provide short-term liquidity for the fund to cover immediate benefit payments. (A typical figure for a UK pension fund in the UK is 5% cash and $p_m = 95\%$ risky assets.) Subject to this constraint, there is total freedom in the proportions invested in the n risky assets.

Recall the Hamilton-Jacobi-Bellman equation:

$$0 = \inf_{\substack{c,p \\ e^T p = p_m}} \left(e^{-\beta t} L(t,c,x) + V_t + \left[(\delta_0 + p^T \lambda)x + c - B\right] V_x + \frac{1}{2} V_{xx}(x^2 p^T D p + \sigma_b^2) \right) \tag{25}$$

where $e = (1, \ldots, 1)^T$.

We differentiate the expression in brackets with respect to c as before to get:

$$\frac{\partial}{\partial c}(\cdot) = e^{-\beta t} L_c + V_x = 0, \qquad \text{where } L_c = \partial L / \partial c \tag{26}$$

$$\Rightarrow c^*(t,x) = L_c^{-1}\left(e^{-\beta t} V_x\right) \tag{27}$$

To minimise over p subject to the constraint we use the method of Lagrangians. Thus we minimise the function:

$$G(p,\gamma) = x V_x \lambda^T p + \frac{1}{2} x^2 V_{xx} p^T D p + \gamma(e^T p - p_m) \tag{28}$$

over p and γ.

$$\frac{\partial G}{\partial p} = x V_x \lambda + x^2 V_{xx} D p + \gamma e = 0 \tag{29}$$

$$\frac{\partial G}{\partial \gamma} = e^T p - p_m = 0 \tag{30}$$

for which the solution is:

$$p = p(x) = \left(p_m + \frac{V_x}{x V_{xx}} e^T D^{-1} \lambda \right) \frac{1}{e^T D^{-1} e} D^{-1} e - \frac{V_x}{x V_{xx}} D^{-1} \lambda \tag{31}$$

$$= \left(d_0 + d_1 \frac{V_x}{x V_{xx}} \right) D^{-1} e - \frac{V_x}{x V_{xx}} D^{-1} \lambda \tag{32}$$

where $d_0 = \dfrac{p_m}{e^T D^{-1} e}$

and $d_1 = \dfrac{e^T D^{-1} \lambda}{e^T D^{-1} e}$

We note, as in the previous section, the connection with modern portfolio theory. We have already discussed the relevance of $D^{-1}\lambda$. Here we note that portfolios which invest in the same proportion as $D^{-1}e$ have the minimum variance given that there are to be no investments in cash. Furthermore, all efficient portfolios are linear combinations of $D^{-1}\lambda$ and $D^{-1}e$.

Again because of the form of the value function we substitute $V(t,x) = e^{-\beta t}F(x)$.

Now $L(t,c,x) = (c - c_m)^2 + 2\rho(c - c_m)(x - x_p) + (k + \rho^2)(x - x_p)^2$ (33)

$$\Rightarrow L_c^{-1}\left(-e^{\beta t}V_x\right) = c_m - \rho(x - x_p) - \frac{1}{2}F_x \tag{34}$$

We apply this to the Hamilton-Jacobi-Bellman equation to give:

$$\begin{aligned}0 =& \frac{1}{4}F_x^2 + k(x - x_p)^2 - \beta F \\ &+ \left[\left(\delta_0 + \lambda^T D^{-1}\left(d_0 e + (d_1 e - \lambda)\frac{F_x}{xF_{xx}}\right)\right)x + c_m - \rho(x - x_p) - \frac{1}{2}F_x - B\right]F_x \\ &+ \frac{1}{2}F_{xx}\left[x^2\left(d_0 e + (d_1 e - \lambda)\frac{F_x}{xF_{xx}}\right)^T D^{-1}\left(d_0 e + (d_1 e - \lambda)\frac{F_x}{xF_{xx}}\right) + \sigma_b^2\right]\end{aligned} \tag{35}$$

As in the unconstrained case this has a quadratic solution $F(x) = P(k)x^2 + Q(k)x + R(k)$. The form of $p(x)$ indicates that we require two portfolios A and B. Portfolio A is made up of fixed proportions of assets 1 to n in proportion to the vector $D^{-1}\lambda$, while portfolio B is synthesised similarly but in proportion to the vector $D^{-1}e$.

As in Section 3.1 portfolios A and B are independent of the form of the loss function.

3.3. Further discussion of the general model

We now consider the optimal asset-allocation strategy in more detail. In particular, consider the instantaneous rate of return on the investments: that is, $\delta_0 + \lambda^T p^*(x)$. Consider the unconstrained case first:

$$\delta_0 + \lambda^T p^*(x) = \delta_0 - \frac{2P(k)x + Q(k)}{2P(k)x}\lambda^T D^{-1}\lambda \tag{36}$$

$$= \delta_0 - \lambda^T D^{-1}\lambda - \frac{Q(k)}{2P(k)x}\lambda^T D^{-1}\lambda \tag{37}$$

Now D is positive definite so that $\lambda^T D^{-1}\lambda$ is positive. Furthermore, $P(k)$ is positive and $Q(k)$ is normally negative. Hence $\delta_0 + \lambda^T p^*(x)$ is normally a decreasing function of x.

Similarly consider the constrained case.

$$\lambda^T p^*(x) = \lambda^T\left[d_0 D^{-1}e + \frac{2P(k)x + Q(k)}{2P(k)x}(d_1 D^{-1}e - \lambda)\right] \tag{38}$$

$$= d_0\lambda^T D^{-1}e + d_1\lambda^T D^{-1}e - \lambda^T D^{-1}\lambda + \frac{Q(k)}{2P(k)x}\lambda^T D^{-1}(d_1 e - \lambda) \tag{39}$$

Note that $e^T D^{-1}(d_1 e - \lambda) = 0$. Hence $\lambda^T D^{-1}(d_1 e - \lambda) = -(d_1 e - \lambda)^T D^{-1}(d_1 e - \lambda) < 0$ since D is positive definite. Again $P(k) > 0$ and normally $Q(k) < 0$ so that $\lambda^T p^*(x)$ is a decreasing function of x.

Furthermore an analysis of the instantaneous variance of the investments confirms that as the instantaneous rate of return decreases, the instantaneous variance decreases also and then starts to increase as we go long (effectively) in low-risk assets and short in high-risk assets.

Thus we find that when the funding level is low we invest more in high-risk assets and as the funding level rises we shift from high-risk into low-risk assets. This is a rather counterintuitive investment strategy. We would expect that as the funding level falls that we might shift into lower-risk assets to protect our position. The strategy we have found here does the opposite. The reason for this is because of the quadratic form of the objective function. This, in a sense, defines an ideal funding level x_p and an ideal contribution rate c_m. If the funding level is below this then we invest in high-return, high-risk assets to increase the chance of getting quickly back to the ideal level. Conversely if the funding level is too high then we are prepared to invest in what is effectively an inefficient, high-risk, low-return investment strategy in order to get back to the ideal level. Indeed the fund will go long in cash and short in equities. In effect the scheme would be throwing money away since, for the same level of risk (that is, volatility of asset returns) it could have a higher expected return. The inefficiency here turns out, with hindsight, to be a result of the quadratic loss function. This actually prefers the positive target contribution rate, c_m, to refunds. In other words, it is better to throw money away than to take a refund. (There is nothing new in this observation. Related problems in other branches of financial economics come to the same counter-intuitive conclusions where, for example, quadratic utility functions are employed.)

Now consider the optimal contribution rate. Sometimes this is written in the form $(c_0 - c_1 x_p) - c_1(x - x_p)$ where $x - x_p$ is the surplus relative to the target fund size x_p. c_1 is the rate at which we try to remove surplus or *amortize* this surplus. It can be noted that the optimal amortization rate, $c_1^* = \hat{P}(k)$, depends on k, δ_0, λ and D but not on c_m, x_p or σ_b^2.

On the other hand, c_0^* also depends on x_p and c_m but again not on σ_b^2.

Similarly it can be seen that $p^*(x)$ does not depend upon σ_b^2. Thus, it has been demonstrated that for such a quadratic loss function $L(\cdot)$ the optimal contribution and asset-allocation strategies do not depend in any way upon the randomness in the level of benefit outgo (at least where this uncertainty is uncorrelated with investment returns).

Later in this paper we will return to the dynamic optimisation problem where we have a different objective function and where there are constraints on the investment strategy and on the funding level.

3.4. Optimal strategy when p is fixed

Suppose instead that the asset-allocation strategy is static: that is, $p(t,x) = p$ for all t, x, for some p. We can still apply the Bellman equation but minimise over $c(t,x)$ only. Thus we find that

$$0=\left[\frac{1}{4}F_x^2+k(x-x_p)^2\right]-\beta F$$
$$+\left[\delta_0 x-B-p^T\lambda x+\left(c_m-\rho(x-x_p)-\frac{1}{2}F_x\right)\right]F_x+\frac{1}{2}F_{xx}(p^T Dpx^2+\sigma_b^2) \quad (40)$$

Again we try to find a solution of the form $F(x) = Px^2 + Qx + R$ and we find that

$$P = P(k) = \frac{\hat{P} + \sqrt{\hat{P}^2 + 4k}}{2} \quad (41)$$

where $\hat{P} = 2\delta_0 - \beta - 2p^T\lambda - 2\rho + p^T Dp \quad (42)$

$$Q = Q(k) = -\frac{2[kx_p + P(k)(B - c_m - \rho x_p)]}{P(k) + \beta - \delta_0 + p^T\lambda + \rho} \quad (43)$$

$$R = R(k) = \frac{1}{\beta}\left[-\frac{1}{4}Q(k)^2 + kx_p^2 - BQ(k) + (c_m + \rho x_p)Q(k) + P(k)\sigma_b^2\right] \quad (44)$$

The question now arises: how do we choose the optimal static p?

We will consider one option here: minimise $P(k)$ over p. This means that the optimal curve $F(x)$ will be as close as possible in the limit as x tends to $\pm\infty$ to the superior solution derived in Section 3.1. Clearly the solution derived in Section 3.1 will be lower for all x regardless of the value of p. (Other possibilities include minimising $F(x)$ over p for a specific value of x, or minimising the minimum of $F(x)$ over p.)

To minimise $P(k)$ over p we differentiate:

$$\frac{dP}{dp} = \frac{dP}{d\hat{P}}\frac{d\hat{P}}{dp} \quad (45)$$

$$= \left(\frac{1}{2} + \frac{1}{2}\hat{P}(\hat{P}^2 + 4k)^{-\frac{1}{2}}\right)(-2\lambda + 2Dp) \quad (46)$$

$$\Rightarrow \hat{p} = D^{-1}\lambda \quad (47)$$

A consideration of the form of $\hat{P}$ as a function of p shows that this is a minimum at $\hat{p}$.

We note that the proportions in the risky assets as given in $\hat{p}$ are the same as those derived in Section 3.1. Furthermore, we find that, given

$p = \hat{p}$, $\hat{P} = 2\delta_0 - \beta - 2\rho - \lambda^T D^{-1}\lambda$ (again the same as in Section 3.1). This means that:

- for large or small values of x the loss of optimality as a result of fixing p does not become too great;
- if we write $c^*(t,x) = c_0^* - c_1^* x$ then c_1^* is not affected by the restriction on the investment strategy (that is, the rate of amortization of surplus or deficit is not affected).

3.5. Comparison of the strategies

Let us consider a specific example to compare the effectiveness of the optimal strategies derived in Sections 3.3 and 3.5 compared to that in Section 3.1. The fixed parameters are as follows:

$$\delta_0 = 0.03, \quad \delta = \begin{pmatrix} 0.04 \\ 0.06 \end{pmatrix}, \quad S = \begin{pmatrix} 0.05 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}, \quad B = 1, \quad \sigma_b = 0.1 \quad (48)$$

The control parameters are:

$$c_m = 0.6, \quad k = 0.001, \quad x_p = 10, \quad \beta = 0.03, \quad \rho = 0 \qquad (49)$$

In this and in subsequent sections we define the funding level, $X(t)$, as the value of the assets divided by the expected rate of benefit outgo. Alternatively, if expected benefit outgo is defined as it is here as $B = 1$ then $X(t)$ is also the fund size.

The optimal value functions $F(x)$ are plotted in Figure 1 and their stationary distributions (as derived later on in Section 4) are plotted in Figure 2.

Selected statistics are given in Table 1. From Table 1 and Figure 1 we can see that the unconstrained solution is significantly better that the other two. The unconstrained and static cases are quite similar in some ways (shape and contribution strategy) but the lack of flexibility in the investment strategy adds on a fixed and substantial penalty. The constrained (no cash) case looks much more different. By reference to Figure 1 it performs well in the middle of the range and, indeed, attempts to stay there by applying a more aggressive amortization strategy. For more extreme values of x this strategy is much poorer than the static case. However, by looking also at the stationary densities of the funding level under the three strategies (Figure 2) we can see that such extreme values will occur very rarely indeed.

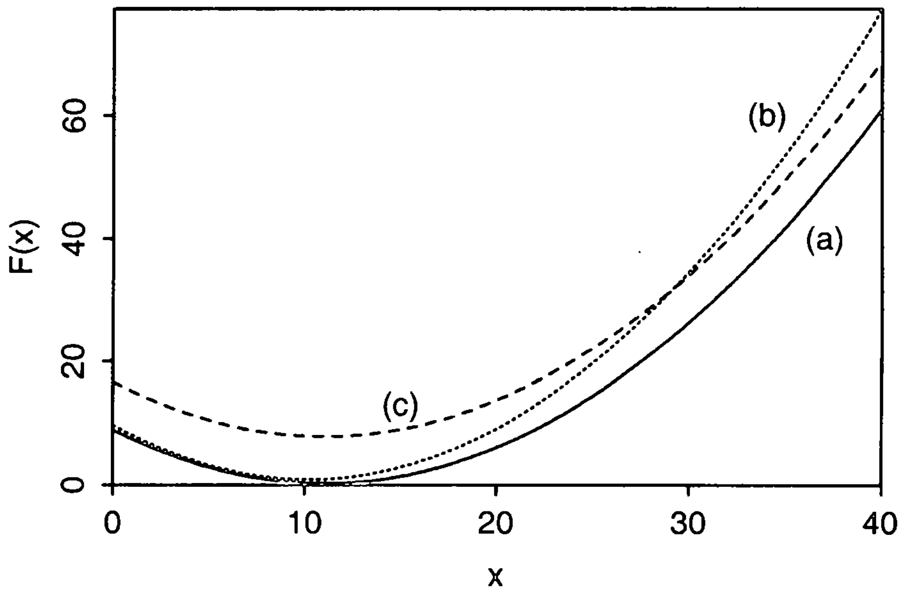


FIGURE 1: Comparison of value functions for three investment/contribution strategies. (a) (solid line) unconstrained optimum. (b) (dotted line) optimum under the constraint of no cash ($p_m = 1$). (c) (dashed line) optimum under a static investment strategy.

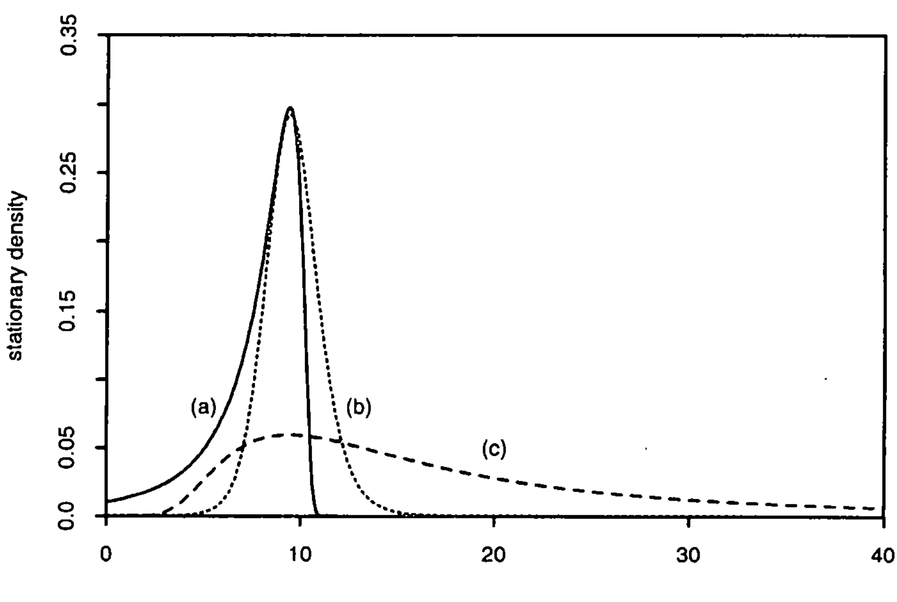


FIGURE 2: Comparison of stationary densities for three investment/contribution strategies. (a) (solid line) unconstrained optimum. (b) (dotted line) optimum under the constraint of no cash ($p_m = 1$). (c) (dashed line) optimum under a static investment strategy. (The funding level is defined here as fund size divided by the expected rate of benefit outgo.)

TABLE I

COMPARISON OF OPTIMAL STRATEGIES WITH AND WITHOUT CONSTRAINTS

	P	Q	R	Minimum $F(x)$	c_0^*	c_1^*
Unconstrained	0.073	−1.60	9.00	0.19	1.40	0.073
Constrained ($p_m = 1$)	0.086	−1.75	9.79	0.85	1.48	0.086
Static	0.073	−1.60	16.70	7.89	1.40	0.073

As mentioned in Section 3.1 we can look at Figures 1 and 2 and see that under the unconstrained investment strategy the value function is minimised at $x_{min} = -Q/2P = 11.01$ and that this, in effect, turns out to be a ceiling (although $X(t)$ can have brief excursions above this value because of volatility in the benefit outgo). Under such circumstances (that is the existence of, effectively, a ceiling) some of the criticisms of the approach with the quadratic loss function become somewhat irrelevant since we are practically never at a funding level where we choose effectively to throw away money (in the sense described in Section 3.4).

Under other circumstances (for example, here if we took $k = 0.001$ and $x_p = 10$ as before but changed c_m from 0.6 to 0.8) the ceiling would turn into a floor at 8.99 and the funding level would spend most of the time above this floor. While this appears to be an appealing strategy the reservations about the investment strategy discussed in Section 3.4 are well founded here.

Under the constrained strategy the value of x_{min} is 10.19 but we can see that the funding level can frequently go above this level. At x_{min} note that the fund here is invested in proportion to the minimum variance portfolio $D^{-1}e$.

Finally we can see from Figure 2 that the static investment strategy leads to much wider fluctuations in the funding level which could only be reduced by increasing the value of k in the loss function.

We will return to this example in Section 5.

3.6. Power and exponential loss functions

3.6.1. Power loss function

Let us complete this section now with a short analysis of the special case where $\sigma_b = 0$ and

$$L(t,c,x) = \begin{cases} -\frac{1}{\gamma}(c_m - c)^\gamma & \text{for } c \le c_m \\ +\infty & \text{for } c > c_m \end{cases} \quad \text{for } 0 < \gamma < 1. \tag{50}$$

Again we assume that the optimal value function takes the form $V(t,x) = \exp(-\beta t)F(x)$. Then the Hamilton-Jacobi-Bellman equation takes the form:

$$\inf_{c,p}\left\{-\frac{1}{\gamma}(c_m - c)^\gamma - \beta F + \left[(\delta_0 + p^T\lambda)x + c - B\right]F_x + \frac{1}{2}F_{xx}x^2p^TDp\right\} = 0 \quad (51)$$

(We restrict optimisation to strategies which keep the fund size positive. Without this condition it is clearly optimal to take contribution refunds of infinite size.)

Now

$$\frac{\partial}{\partial c}(\cdot) = 0 \quad (52)$$

$$\Rightarrow (c_m - c)^{\gamma-1} + F_x = 0 \quad (53)$$

$$\Rightarrow c^*(x) = c_m - (-F_x)^{1/(\gamma-1)} \quad (54)$$

$$\frac{\partial}{\partial p}(\cdot) = 0 \quad (55)$$

$$\Rightarrow p^*(x) = -D^{-1}\lambda\frac{F_x}{xF_{xx}} \quad (56)$$

Inserting $c^*(x)$ and $p^*(x)$ into equation (51) we get

$$(-F_x)^{\gamma/(\gamma-1)}\left(\frac{\gamma-1}{\gamma}\right) - \beta F + (\delta_0 x + c_m - B)F_x - \frac{1}{2}\frac{F_x^2}{F_{xx}}\lambda^TD^{-1}\lambda = 0 \quad (57)$$

We try for a solution of the form $F(x) = -k(x - x_m)^\alpha$.
Inserting this into equation (57) we get, for all x:

$$\begin{aligned}&(k\alpha)^{\gamma/(\gamma-1)}(x - x_m)^{\gamma(\alpha-1)/(\gamma-1)}\left(\frac{\gamma-1}{\gamma}\right)\\ &+\beta k(x - x_m)^\alpha - \delta_0\left(x - \frac{B - c_m}{\delta_0}\right)k\alpha(x - x_m)^{\alpha-1}\\ &+\frac{1}{2}\frac{k^2\alpha^2(x - x_m)^{2\alpha-2}}{k\alpha(\alpha-1)(x - x_m)^{\alpha-2}}\lambda^TD^{-1}\lambda = 0 \end{aligned} \quad (58)$$

$$\Rightarrow x_m = \frac{B - c_m}{\delta_0} \quad (59)$$

$$\alpha = \gamma \quad (60)$$

$$\text{and } k = \frac{1}{\gamma}c_1^{\gamma-1} \quad (61)$$

$$\text{where } c_1 = \left(\frac{\beta - \delta_0\gamma + \frac{1}{2}\frac{\gamma}{\gamma-1}\lambda^TD^{-1}\lambda}{1-\gamma}\right) \quad (62)$$

Hence

$$c^*(x) = c_m - c_1(x - x_m) \tag{63}$$

$$p^*(x) = D^{-1}\lambda\frac{(x - x_m)}{(1 - \gamma)x} \tag{64}$$

We note the similarity of the problems and solutions here with a well-known optimal-consumption problem described by Merton (1971, 1990). Equivalence is achieved by equating the controllable level of consumption with $B - c^*(t)$ in the current model. This enables us to speculate that recent extensions of this work to include the effects of transactions costs can be applied to the present problem. For example, the problem of proportional transactions costs has been considered by, amongst others, Davis & Norman (1990) and Shreve & Soner (1994).

As with the quadratic loss function, contributions decrease linearly with x with the amortisation rate c_1 being detennined by the discount rate β and the risk-aversion parameter γ (but not the maximum acceptable contribution rate, c_m).

Investment in risky assets, $p^*(x) \cdot x$ increases linearly in x above the minimum x_m, and therefore appears to conform better with conventional wisdom. However, it turns out that this solution gives rise to one of two trivial stationary solutions for $X(t)$: that is, $X(t) \to x_m$ or $+\infty$ depending upon the value of β.

Returning to the dynamics of the funding level $X(t)$ we find that $c^*(x)$ and $p^*(x)$ give rise to

$$dX(t) = (X(t) - x_m)\left[\left(\delta_0 + \frac{1}{1-\gamma}\lambda^T D^{-1}\lambda - c_1\right)dt + \frac{1}{1-\gamma}\lambda^T S^{-1} dZ(t)\right] \tag{65}$$

$$\stackrel{\mathcal{D}}{=} (X(t) - x_m)\left[\left(\delta_0 + \frac{1}{1-\gamma}\lambda^T D^{-1}\lambda - c_1\right)dt + \frac{1}{1-\gamma}\sqrt{\lambda^T D^{-1}\lambda}\, d\tilde{Z}(t)\right] \tag{66}$$

where $\tilde{Z}(t)$ is another Brownian motion. It follows that (and inserting the known form of c_1):

$$X(t) - x_m = (X(0) - x_m)\exp\left[\left(\delta_0 + \frac{1}{2(1-\gamma)}\lambda^T D^{-1}\lambda - \beta\right)t + \frac{1}{1-\gamma}\sqrt{\lambda^T D^{-1}\lambda}\,\tilde{Z}(t)\right] \tag{67}$$

That is, $X(t) - x_m$ is a geometric Brownian motion which tends to zero if $\beta > \delta_0 + \lambda^T D^{-1}\lambda/2(1 - \gamma)$ and to $+\infty$ if $\beta < \delta_0 + \lambda^T D^{-1}\lambda/2(1 - \gamma)$.

There are some similarities between this solution and that of Boulier *et al.* (1995) under which $x_m - X(t)$ is also a geometric Brownian motion.

With either the introduction of volatility in benefit outgo ($\sigma_b > 0$) or with restrictions on the amount of cash we cannot have both a lower bound on the funding level and an upper bound on the contribution rate.

The loss function $L(c) = c^\gamma/\gamma$ for $c > 0$ and $\gamma > 1$ has been considered by Siegmann & Lucas (1999). They obtain similar results to those described above, except that x_m becomes a maximum, and contributions are bounded below by 0 rather than above by c_m.

3.6.2. *Exponential loss function*

Similarly, we can consider the exponential loss function (for example, see Siegmann & Lucas, 1999):

$$L(t, c, x) = \exp(\gamma c - \theta x) \tag{68}$$

where $\gamma > 0$ and $\theta > 0$. Here the relationship between γ and θ determines the relative emphasis on the employer and the members.

This gives us the solution:

$$F(x) = \exp(a - bx) \tag{69}$$

where $b = \gamma\delta_0 + \theta$

$$\text{and } a = \log\frac{\gamma}{\gamma\delta_0 + \theta} + \frac{-\beta + (\gamma\delta_0 + \theta)B + (\gamma\delta_0 + \theta)/\gamma - \frac{1}{2}\lambda^T D^{-1}\lambda}{(\gamma\delta_0 + \theta)/\gamma}$$

$$\Rightarrow c^*(t) = c_0 - c_1 x \tag{70}$$

$$p^*(t) = \frac{1}{x}p_0 \tag{71}$$

$$\text{where } c_0 = \frac{-\beta + (\gamma\delta_0 + \theta)B + (\gamma\delta_0 + \theta)/\gamma - \frac{1}{2}\lambda^T D^{-1}\lambda}{\gamma\delta_0 + \theta}$$

$$c_1 = \delta_0$$

$$\text{and } p_0 = \frac{D^{-1}\lambda}{\gamma\delta_0 + \theta}$$

This solution is more like the quadratic loss function considered in earlier sections: that is, the proportion of the fund invested in risky assets decreases as x increases. If we increase θ then p_0 decreases. This reflects the fact that there is a greater degree of risk aversion when we consider the interests of the members, so we invest less in risky assets.

With a little algebra we can see that $X(t)$ follows a Brownian motion with drift $\mu = \left(-\beta + (\gamma\delta_0 + \theta)/\gamma + \frac{1}{2}\lambda^T D^{-1}\lambda\right)/(\gamma\delta_0 + \theta)$ and volatility $\sqrt{\alpha} = \sqrt{\lambda^T D^{-1}\lambda}/(\gamma\delta_0 + \theta)$.This means that the solution is unsatisfactory because it is both non-stationary and because it gives rise to a 'counter-intuitive' investment strategy.

4. THE STATIONARY DISTRIBUTION OF $X(t)$

4.1. General model

Assume now that

$$c(t,x) = c(x) = c_0 - c_1 x \tag{72}$$

$$p(t,x) = p(x) = \frac{p_0 + p_1 x}{x} \tag{73}$$

where p_0 and p_1 are $n \times 1$ vectors.

The reason for assuming a linear form for $c(x)$ and $xp(x)$ is simple. They are consistent with the optimal dynamic controls derived in Section 3 when we use a quadratic loss function. Furthermore let us recall the value function

$$W(t,x)(c,p) = E\left[\int_t^\infty \exp(-\beta s)L(s,c(s,X(s)),X(s))ds | X(t) = x\right] \tag{74}$$

As $\beta \to 0$, $\beta W(t,x) \to E[L(s,c,X)]$: that is, the limiting optimal dynamic controls are also optimal in the static case if we use the same quadratic loss function.

The dynamics of the fund size, $X(t)$, are then

$$dX = X\left[\left(\delta_0 + p(x)^T\lambda\right)dt + p(x)^T S d\tilde{Z}\right] + (c_0 - c_1 X - B)dt + \sigma_b dZ_b \tag{75}$$

$$= \left[\left(\delta_0 X + (p_0 + p_1 X)^T\lambda\right)dt + (p_0 + p_1 X)^T S d\tilde{Z}\right] + (c_0 - c_1 X - B)dt + \sigma_b dZ_b \tag{76}$$

$$\stackrel{\mathcal{D}}{=} \mu \cdot dt - \nu X \cdot dt + (\alpha + \beta X + \gamma X^2)^{1/2} dZ \tag{77}$$

where $\tilde{Z}(t)$ is a standard n-dimensional Brownian Motion, and $Z(t)$ is a standard Brownian motion which depends upon $\tilde{Z}(t)$ and $Z_b(t)$,

$$\mu = c_0 - B + p_0^T\lambda \tag{78}$$

$$\nu = c_1 - \delta_0 - p_1^T\lambda \tag{79}$$

$$\alpha = p_0^T D p_0 + \sigma_b^2 \tag{80}$$

$$\beta = 2 p_0^T D p_1 \tag{81}$$

$$\gamma = p_1^T D p_1 \tag{82}$$

In order to discuss the properties of this model we state the following theorem:

Theorem 4.1.1
Let the continuous-time stochastic process X, satisfy the stochastic differential equation

$$dX_t = (\alpha + \beta X_t + \gamma X_t^2)^{1/2} dZ + \mu dt - \nu X_t dt \tag{83}$$

subject to the constraints on the parameters $\alpha > 0$, $\gamma > 0$, $\beta^2 - 4\alpha\gamma \leq 0$, $\mu > 0$ *and* $\nu > 0$.

(a) If $\beta^2 - 4\alpha\gamma < 0$, *the stationary density function of* X_t *is*

$$f_X(x) = K \exp\left[2a \tan^{-1}\frac{x+b}{c}\right](\alpha + \beta x + \gamma x^2)^{-1-\nu/\gamma} \qquad (84)$$

$$\textit{for} \quad -\infty < x < \infty$$

$$\textit{where } a = \frac{1}{\sqrt{4\alpha\gamma - \beta^2}}\left(\frac{\nu\beta}{\gamma} + 2\mu\right)$$

$$b = \frac{\beta}{2\gamma}$$

$$c = \frac{\sqrt{4\alpha\gamma - \beta^2}}{2\gamma}$$

(b) If $\beta^2 - 4\alpha\gamma = 0$ *and* $X_0 > -b$, *the stationary density function of* X_t *is*

$$f_X(x) = \begin{cases} K(x+b)^{-\theta}\exp[-\phi/(x+b)] & \begin{cases} \textit{for} \quad x > -b \quad \textit{if} \quad \theta > 0 \\ \textit{for} \quad x < -b \quad \textit{if} \quad \theta < 0 \end{cases} \\ 0 & \textit{otherwise} \end{cases} \qquad (85)$$

$$\textit{where } b = \frac{\beta}{2\gamma}$$

$$\theta = 2\left(1 + \frac{\nu}{\gamma}\right)$$

$$\phi = \frac{\nu\beta + 2\mu\gamma}{\gamma^2}$$

that is, the Translated-Inverse-Gamma distribution with parameters $-b$, $\theta - 1 > 0$ *and* $\phi > 0$ $(TIG(-b,\ \theta - 1,\ \phi))$. *(If* $X \sim TIG(k, \alpha, \beta)$ *then* $(X-k)^{-1} \sim Gamma(\alpha, \beta)$.*)*

In each case K is a normalizing constant.

Proof: Proofs of these two results have been provided before by a number of authors. Distribution (b) was first derived in the context of pension funding by Dufresne (1990) in the case where there is one risky asset, no cash and no demographic volatility ($\sigma_b = 0$). Dufresne also noted that the stationary distribution of the funding level was the same as the distribution of a perpetuity. An alternative proof for the distribution of the present value of a perpetuity was also shown to have distribution (b) by Yor (1992) and by De Schepper *et al.* (1994). Föllmer and Schweizer (1993, Theorem 5.1 and erratum) considered the diffusion process defined above as underlying a model for stock prices. They derived both of the limiting distributions given in (a) and (b).

The two distributions above are also known as Pearson type IV and type V distributions respectively (for example, see Johnson *et al.*, 1994).

Let us consider the Pearson Type IV distribution. This distribution has four degrees of freedom. The fifth degree of freedom used in the dynamics of the fund size determines the speed of the process.

Following the notation of Johnson *et al.* (1994) we define $\mu'_r = E[X^r]$. We define $\mu'_{-1} = 0$ and have $\mu'_0 = 1$. It is easy to show that the μ'_r satisfy the following recursive relationship:

$$-k_0 r\mu'_{r-1} + (k_3 - (r+1)k_1)\mu'_r + (1-(r+2)k_2)\mu'_{r+1} = 0 \tag{86}$$

$$\text{where } k_0 = \frac{\alpha}{2(\gamma+\nu)}, \quad k_1 = \frac{\beta}{2(\gamma+\nu)}, \quad k_2 = \frac{\gamma}{2(\gamma+\nu)}, \quad k_3 = \frac{\beta-2\mu}{2(\gamma+\nu)} \tag{87}$$

Hence we have

$$E[X] = \mu'_1 = \frac{k_1 - k_3}{1-2k_2} = \frac{\mu}{\nu} \tag{88}$$

$$E\left[X^2\right] = \mu'_2 = \frac{k_0 + (2k_1 - k_3)\mu/\nu}{1-3k_2} = \frac{\alpha\nu + \beta\mu + 2\mu^2}{\nu(2\nu-\gamma)} \tag{89}$$

$$\Rightarrow Var[X] = \frac{\alpha\nu+\beta\mu+2\mu^2}{\nu(2\nu-\gamma)} - \frac{\mu^2}{\nu^2} = \frac{\nu^2\left(\alpha+\beta\frac{\mu}{\nu}+\gamma\frac{\mu^2}{\nu^2}\right)}{\nu^2(2\nu-\gamma)} = \frac{\alpha+\beta E[X]+\gamma E[X]^2}{2\nu-\gamma} \tag{90}$$

$$E[c(X)] = c_0 - c_1 E[X] = c_0 - c_1\frac{\mu}{\nu} \tag{91}$$

$$Var[c(X)] = c_1^2 Var[X] = c_1^2\frac{\alpha+\beta E[X]+\gamma E[X]^2}{2\nu-\gamma} \tag{92}$$

The stationary distribution exists if and only if $2(1+\nu/\gamma) > 1$: that is $c_1 > \delta_0 + p_1^T\lambda - \frac{1}{2}\gamma$ (in fact, $\gamma^{1/2}$ is the asymptotic volatility as $|x| \to \infty$). Similarly $E[X]$ exists if and only if $c_1 > \delta_0 + p_1^T\lambda$ and $Var[X]$ exists if and only if $c_1 > \delta_0 + p_1^T\lambda + \frac{1}{2}\gamma$.

Coming back to the optimal solution for the dynamic problem we found that $c_1^* = P(k) = P(k,\beta)$. Thus the condition for stationarity is

$$P(k) > \delta_0 + p_1^{*T}\lambda - \frac{1}{2}p_1^{*T}Dp_1^* \tag{93}$$

$$\text{where } p_1^* = -D^{-1}\lambda \quad \text{(unconstrained case).}$$

$$\Rightarrow P(k) > \delta_0 + \frac{3}{2}\lambda^T D^{-1}\lambda \tag{94}$$

Now $P(k, \beta)$ is a decreasing function of β and an increasing function of k, so the condition above is less likely to be satisfied if β is large or k is small. Under such circumstances the funding level will diverge as t tends to infinity. The situation, therefore, is that with a relatively large value of β we pay more attention to control of short-term variability in the contribution rate at the expense of larger fluctuations in the long term. Likewise, if the value of k is too small then we also pay too much attention to short-term contribution-rate stability.

4.2. Continuous proportion portfolio insurance

The idea of continuous proportion portfolio insurance (CPPI) was introduced by Black and Jones (1988) and Black and Perold (1992).

The previous sections in this paper have concentrated upon quadratic loss functions. The motivation behind CPPI is that in certain countries there exist minimum funding constraints: that is, there exists a floor below which the funding level must not fall. CPPI was proposed as a means of reducing the risk that the fund falls below this floor.

Under CPPI if the funding level is low then the fund will be invested more in low-risk assets (in particular, those which will best match variations in the floor). As the funding level improves the fund can be shifted more into risky assets which provide the fund with higher upside potential.

Suppose that the minimum funding level (or floor) is M. We have a low-risk portfolio A with a proportion π_{Ai} of the fund invested in asset i ($i = 1, 2, \ldots, n$). We also have a higher-risk portfolio B which invests in proportion to the vector π_B. At funding level x a proportion $p_A(x)$ of the fund is invested in portfolio A and $p_B(x) = 1 - p_A(x)$ in portfolio B. Since A is less risky we have (normally):

$$\pi_A^T\lambda < \pi_B^T\lambda \quad \text{(that is, A has a lower expected return)} \tag{95}$$

$$\pi_A^T D\pi_A < \pi_B^T D\pi_B \quad \text{(that is, A is lower risk)} \tag{96}$$

We define $p_B(x)$ in one of the following ways:

$$p_B(x) = \frac{x - M}{x} \tag{97}$$

$$\text{or } \bar{p}_B(x) = \max\left\{\frac{x - M}{x}, 0\right\} \tag{98}$$

We will concentrate here on $p_B(x)$ for the sake of mathematical convenience since it is normally the case that the probability that $X(t)$ falls below M under this strategy is very small if A is very low risk.

The vector of proportions invested in each asset under CPPI is thus:

$$p_c(x) = \frac{M}{x}\pi_A + \frac{x-M}{x}\pi_B \tag{99}$$

$$= \frac{p_{c0} + p_{c1}x}{x} \tag{100}$$

$$\text{where } p_{c0} = M(\pi_A - \pi_B)$$

$$p_{c1} = \pi_B$$

We can, therefore, apply all of the results discussed in Section 4.1 to CPPI. For example, it is of interest to compare the effectiveness of CPPI relative to a static investment strategy. Let us look first at the stationary mean and variance of the funding level. In the equations below we use a subscript c for calculations under CPPI and s where we are considering the static strategy. Thus:

$$E[X_c] = \frac{\mu_c}{\nu_c} = \frac{c_0 - B + p_{c0}^T\lambda}{c_1 - \delta_0 - p_{c1}^T\lambda} = m_c \tag{101}$$

$$\text{and } Var[X_c] = \frac{\alpha_c + \beta_c m_c + \gamma_c m_c^2}{2\nu_c - \gamma_c} = s_c^2 \quad \text{say.} \tag{102}$$

We will assume that the floor, M, is sufficiently small and that portfolios A and B, and the contribution strategy have been chosen in such a way that $X(t)$ is stationary with $M < m_c < \infty$.

Now suppose that we will employ a static investment strategy under which we hold assets in proportion to the vector $p_s\pi_A + (1 - p_s)\pi_B$ for all x where p_s is some scalar quantity. Then we have:

$$E[X_s] = \frac{c_0 - B}{c_1 - \delta_0 - (p_s\pi_A + (1-p_s)\pi_B)^T\lambda} = m_s. \tag{103}$$

Now choose p_s in such a way that $m_s = m_c$: that is,

$$p_s = \frac{m_c(c_1 - \delta_0 - \pi_B^T\lambda) - (c_0 - B)}{m_c(\pi_A - \pi_B)^T\lambda} = \frac{M(\pi_A - \pi_B)^T\lambda}{m_c(\pi_A - \pi_B)^T\lambda} = \frac{M}{m_c} \tag{104}$$

Note that $0 < p_s < 1$.

We now claim that $Var[X_s] < Var[X_c]$.

$$\alpha_c + \beta_c m_c + \gamma_c m_c^2 \tag{105}$$

$$= \sigma_b^2 + M^2(\pi_A - \pi_B)^T D(\pi_A - \pi_B) + 2M(\pi_A - \pi_B)^T D\pi_B m_c + \pi_B^T D\pi_B m_c^2 \tag{106}$$

$$= \sigma_b^2 + (M(\pi_A - \pi_B) + m_c\pi_B)^T D(M(\pi_A - \pi_B) + m_c\pi_B) \tag{107}$$

$$= \sigma_b^2 + (p_c\pi_A + (1-p_c)\pi_B)^T D(p_c\pi_A + (1-p_c)\pi_B)m_c^2 \tag{108}$$

where $p_c = M/m_c$.

We also have:

$$\alpha_s + \beta_s m_s + \gamma_s m_s^2 = \sigma_b^2 + (p_s\pi_A + (1-p_s)\pi_B)^T D(p_s\pi_A + (1-p_s)\pi_B)m_s^2 \quad (109)$$

But $m_c = m_s$ and $p_c = p_s$ so that $\alpha_c + \beta_c m_c + \gamma_c m_c^2 = \alpha_s + \beta_s m_s + \gamma_s m_s^2$.

Next consider:

$$\begin{aligned} 2\nu_s - \gamma_s &= 2\Big(c_1 - \delta_0 - \pi_B^T\lambda - p_s(\pi_A - \pi_B)^T\lambda\Big) && (110)\\ &\quad - (p_s\pi_A + (1-p_s)\pi_B)^T D(p_s\pi_A + (1-p_s)\pi_B) && (111)\\ &= 2(c_1 - \delta_0 - \pi_B^T\lambda) - \pi_B^T D\pi_B - 2p_s(\pi_A - \pi_B)^T\lambda && (112)\\ &\quad + \Big[\pi_B^T D\pi_B - (p_s\pi_A + (1-p_s)\pi_B)^T D(p_s\pi_A + (1-p_s)\pi_B)\Big] && (113) \end{aligned}$$

Now $0 < p_s < 1$, $(\pi_A - \pi_B)^T\lambda < 0$ and

$$\Big[\pi_B^T D\pi_B - (p_s\pi_A + (1-p_s)\pi_B)^T D(p_s\pi_A + (1-p_s)\pi_B)\Big] > 0 \quad (114)$$

(since the expression in square brackets is convex, quadratic in p_s and $\pi_A^T D\pi_A < \pi_B^T D\pi_B$).

Therefore:

$$2\nu_s - \gamma_s > 2\nu_c - \gamma_c. \quad (115)$$

Hence:

$$\frac{\alpha_s + \beta_s m_s + \gamma_s m_s^2}{2\nu_s - \gamma_s} < \frac{\alpha_c + \beta_c m_c + \gamma_c m_c^2}{2\nu_c - \gamma_c} \quad (116)$$

$$\Rightarrow Var[X_s] < Var[X_c] \quad (117)$$

This can be summarised in the following theorem:

Theorem 4.2.1
For any CPPI investment strategy let m_c and s_c^2 be the stationary mean and variance of the funding level X_t. There exists a static investment strategy under which the stationary mean funding level, m_s, is equal to m_c but the stationary variance of the funding level, s_s^2, is less than s_c^2.

Interpretation: In the variance sense, the static strategy is *more efficient* than CPPI: that is, given a CPPI strategy we can always find a static strategy which delivers the same mean funding level but a lower variance.

One example illustrating this result is plotted in Figure 3. Here we use the same fixed parameters as in Section 3.5. In addition we have $c_0 = 1.5$ and $c_1 = 0.07$ for both the static and CPPI strategies. Under CPPI we have a floor of $M = 10$ with $\pi_A^T = (0,0)$ and $\pi_B^T = (0.2, 0.8)$ (meaning that at the floor ($(X(t) = M)$ the fund is invested 100% in cash). This gives rise to a mean funding level (that is, assets divided by expected benefit outgo) of 17.1

while the variance of the funding level is, in fact, infinite. Under the matching static investment strategy the expected funding level is also 17.1 while the standard deviation of the funding level is 5.4. This marked difference in the variances is caused by the fatness of the tail of the CPPI distribution although this is not clear from Figure 3. What we can see in Figure 3 is that the two distributions are quite different.

One might ask why would we use CPPI when the static strategy has been shown to be more efficient. The answer to this is that it depends upon the objectives of the pension fund. If the objective is to minimise variance then clearly the static strategy is superior (although we have shown in Section 3 that a form of "inverse" CPPI is better still). On the other hand, if the objective is to minimise the probability that the funding level falls below the floor, M, then CPPI is clearly superior.

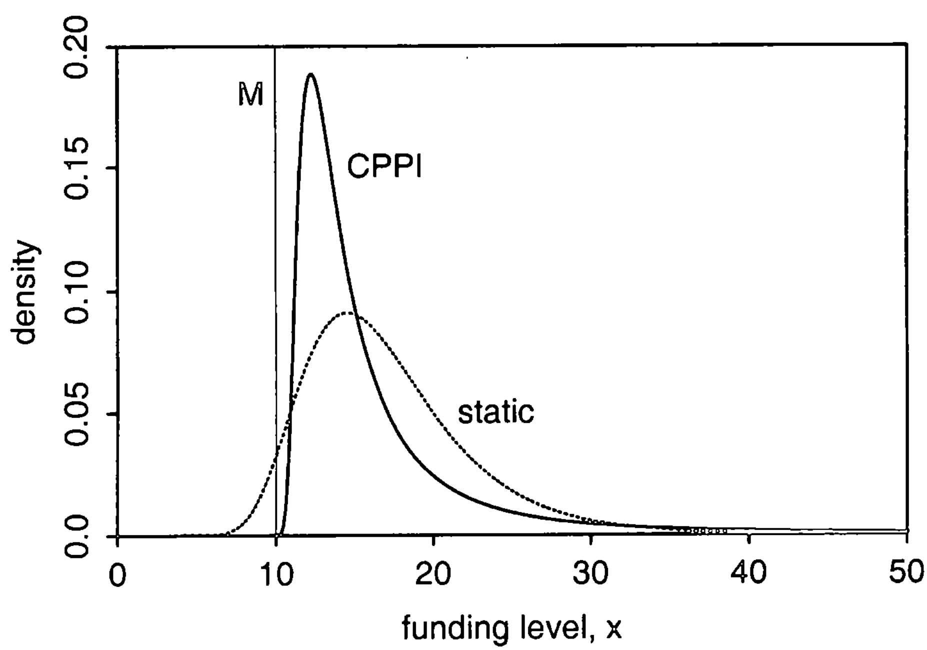


FIGURE 3: Comparison of stationary distributions for static and CPPI investment strategies. Both distributions have the same mean.

5. Numerical examples

We consider now an example in which the following parameters are fixed (as in Section 3.6):

$$\delta_0 = 0.03, \quad \delta = \begin{pmatrix} 0.04 \\ 0.06 \end{pmatrix}, \quad S = \begin{pmatrix} 0.05 & 0.05 \\ 0.05 & 0.2 \end{pmatrix}, \quad B = 1, \quad \sigma_b = 0.1 \tag{118}$$

Here we consider an analysis of the sensitivity of the optimal control strategies to variation of the input parameters in the value function and the loss function. The central parameter values which we will use are:

$$c_m = 0.6, \quad k = 0.001, \quad x_p = 10, \quad \beta = 0.03 \tag{119}$$

Furthermore, we assume that none of the fund can be invested in cash (as in Section 3.3 with $p_m = 1$). Throughout this analysis we keep the fifth input parameter ρ equal to 0.

In Tables 2 (Dynamic optimisation) and 3 (Stationary optimisation) below we give the values of the input parameters (c_m, k, x_p and β), the optimal values of p_0, p_1, c_0 and c_1, and the mean and standard deviation of the stationary fund size and the contribution rate.

The values given for p_{B0} and p_{B1} relate to the proportion of the fund invested in the more risky but efficient portfolio B: that is, the portfolio in which investments are in proportion to the vector $D^{-1}\lambda$. In particular, the proportion of the fund invested in portfolio B is $p_B(x) = (p_{B0} + p_{B1}x)/x$. Since $p_m = 1$ the remainder of the assets are invested in the minimum variance portfolio A: $D^{-1}e/\left(e^T D^{-1} e\right)$.

TABLE 2

Dynamic Optimisation

Ex.	c_m	k	x_p	β	σ_b	p_{B0}	p_{B1}	c_0	c_1	$E[X]$	$SD[X]$	$E[C]$	$SD[C]$
1	0.6	0	–	0.02	0.1	20.7	−1.97	1.00	0.038	8.9	3.8	0.660	0.146
2	0.6	0	–	0.03	0.1	20.7	−1.97	0.89	0.028	7.0	∞	0.699	∞
3(*)	0.6	0.005	10	0.03	0.1	20.1	−1.97	1.48	0.086	9.6	1.5	0.651	0.132
4	0.5	0.05	10	0.03	0.1	19.8	−1.97	3.00	0.238	9.9	0.8	0.647	0.194
5	0.6	0.005	10	0.03	0.1	22.1	−1.97	1.47	0.086	9.7	1.6	0.633	0.141
6	0.6	0.005	15	0.03	0.1	26.0	−1.97	1.74	0.086	14.2	2.3	0.512	0.194
7	0.6	0.005	10	0.03	0.2	20.1	−1.97	1.48	0.086	9.6	1.6	0.651	0.139

TABLE 3

STATIONARY OPTIMISATION

Ex.	c_m	k	x_p	β	σ_b	p_{B0}	p_{B1}	c_0	c_1	$E[X]$	$SD[X]$	$E[C]$	$SD[C]$
1	0.6	0	–	–	0.1	20.7	−1.97	1.21	0.058	9.7	2.2	0.644	0.127
2			–	–									
3(*)	0.6	0.005	10	–	0.1	20.1	−1.97	1.68	0.105	9.8	1.3	0.647	0.141
4	0.6	0.05	10	–	0.1	19.9	−1.97	3.16	0.254	9.9	0.8	0.647	0.200
5	0.5	0.005	10	–	0.1	22.4	−1.97	1.70	0.105	10.3	1.4	0.619	0.152
6	0.6	0.005	15	–	0.1	25.6	−1.97	1.97	0.105	13.7	1.9	0.526	0.197
7	0.6	0.005	10	–	0.2	20.1	−1.97	1.68	0.105	9.8	1.4	0.647	0.149

5.1. Notes on the numerical examples

- Examples 1 and 2 show the effect of changing the risk-discount rate β when $k = 0$. Note how the variances in the dynamic case become infinite as β increases. However, when $\beta = 0.03$ the dynamic optimum still has a stationary distribution, albeit with infinite variances.
  When $k = 0$ we can also see that the optimal asset-allocation strategies for the dynamic and static cases are the same and do not depend upon β. We also see that $c_1^D = c_1^S - \beta$ if $k = 0$.
  As β tends to 0 the optimal dynamic solutions converge to the same values as the optimal static solution. The effect of β is therefore to suppress variance in the short term through a lower value of c_1. A low value of c_1 may reduce variance in the short term but it increases it in the long run by allowing fluctuations in the fund size to persist.
  In Example 1 we also see that $E[C] > c_m$. This reflects that fact that the minimum variance of C falls as $E[C]$ increases (and $E[X]$ falls).
- In Example 1, the fund is invested 100% in portfolio B when X equals about 7.0. Below this the fund goes long in portfolio B and short in portfolio A. Conversely, when X reaches just above 10.5 the fund has 100% in portfolio A. When X goes above this there is a long position in portfolio A and a short position in portfolio B.
  Similar ranges apply for each of the other examples.
- Examples 2, 3 and 4 show the effect of increasing k. This shifts the emphasis onto reducing the variance of the fund size rather than of the contribution rate. The principle effect is that c_1 increases with k: that is, surplus or deficit is amortised more quickly. The changes in p_0 and c_0 are primarily a knock on effect.
- Examples 3 and 5 demonstrate the consequences of changing c_m. p_1 remains unchanged as it does throughout. The changes in the remaining control parameters have the effect of shifting the mean values principally but also affect the variances.

- Examples 3 and 6 consider the effect of changing the target fund size x_p. There is no change in c_1 or p_1. p_0 and c_0 change in order to shift the mean fund size. The variance rises because the target fund size is being moved away from the more natural mean observed in Example 3. This increases the tension on the mean contribution rate since a target fund size of 15 is not entirely consistent with a target contribution rate of 0.6.
- Examples 3 and 7 show the influence of the uncertainty in the level of benefit outgo. As was remarked in Section 3, σ_b has no effect on the optimal values of p_0, p_1, c_0 and c_1. Furthermore, the increases in the variances are small indicating that at this level ($\sigma_b = 0.1$ or 0.2) the main source of variability in the contribution rate is due to investment risk.
- The stationary distributions for the fund size for the dynamic and the stationary optima in Example 3 are plotted in Figure 4. It can be seen that the results are similar although the dynamic optimum gives rise to a stationary distribution which is less peaked and which has fatter tails. In other cases (for example, Example 6) if there is some tension between the target funding level, x_p, and the target contribution rate, c_m, there will be more of a difference between the two stationary distributions.

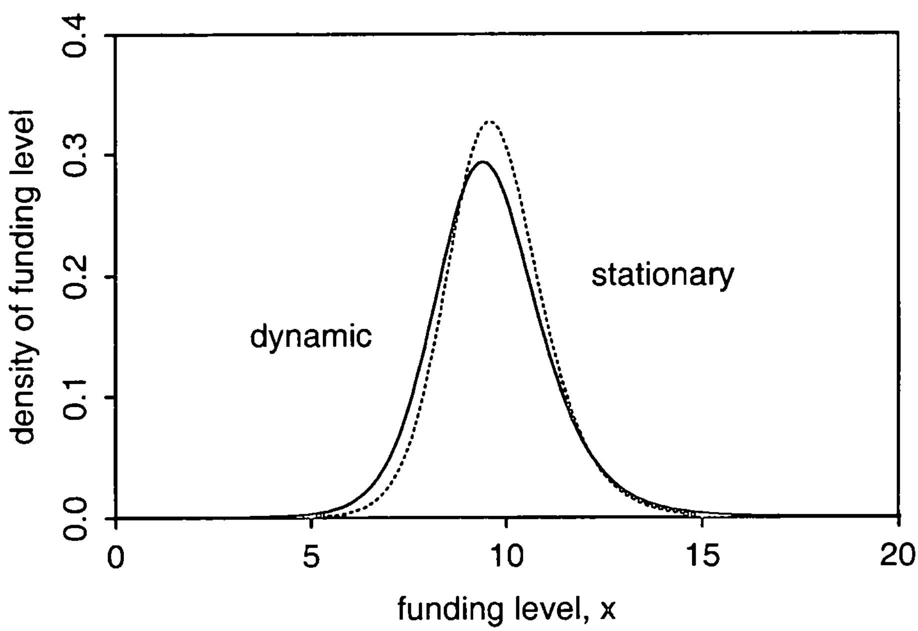


FIGURE 4: Example 3: Comparison of the stationary distribution of the funding levels for the dynamic and stationary optimal solutions.

6. Constraints and Discontinuities

A number of possible constraints can be put in place which complicate considerably the proceeding analyses. These are:

- upper and lower barriers for the funding level, $X(t)$. These might be legislative requirements or self-imposed by the fund sponsor and the trustees.
- an upper limit set by the fund sponsor on the contribution rate.
- restrictions on the short-selling of assets.

Further discontinuities might exist where the objective function has a non-standard form. For example, we may have

$$L(t,c,x) = (c - c_m)^2 + k\big(\max\{x_p - x, 0\}\big)^2 \tag{120}$$

where the second term only introduces a penalty when the funding level drops below x_p. Such a function can also be used as a means of investigating the effects of a barrier since as k gets larger and larger the optimal contribution rate below x_p will increase in an effort to raise the funding level above x_p as quickly as possible. For large k this will have the effect of looking like a reflection off the barrier

Analysis of many of these problems is under way but there are only a few interesting results to discuss at this stage.

6.1. Dynamics in the presence of a minimum barrier

A much simplified version of the minimum funding requirement in the UK is as follows. There is a floor M below which the funding level should not fall. If $X(t)$ does drop below M then it is immediately increased to M by a special contribution.

This problem can be approached by modifying the original setup described in Section 2 by adding an additional contribution rate $c^+ \cdot \max\{M - X(t), 0\}$: that is, when the funding level is below M. As c^+ tends to infinity the dynamics of the model approach that described above and the process reflects off the barrier M. In this limit the process can be written as follows:

$$dX = X\Big[\Big(\delta_0 + p(X)^T\lambda\Big)dt + p(X)^T S dZ\Big] + c(X)\cdot dt - B\cdot dt - \sigma_b\cdot dZ_b + dL_t^M. \tag{121}$$

The new term in this formula, dL_t^M, is called the *local time* of the process, $X(t)$, at M and is defined as

$$L_t^M = \int_0^t dL_s^M \tag{122}$$

$$= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_0^t I(M \le X(t) < M + \varepsilon)ds \tag{123}$$

where $I(\cdot)$ is equal to 1 when $X(t)$ lies between M and $M+\varepsilon$ and 0 otherwise. L_t^M is a measure of how much time the process spends in the vicinity of M.

The dL_t^M term represents the additional contributions required when the process hits the barrier to keep $X(t)$ above M. In a sense it gives the process a small upwards 'kick' every time it hits the barrier.

It is possible to analyse the stationary distribution of such a process when $c(x)$ and $p(x)\cdot x$ are linear in x away from the barrier: it has a truncated Pearson type IV distribution. However, $Var\left[c(X)dt + dL_t^M\right]/dt^2$ is infinite whereas the variance is finite when there is no barrier. This inhibits the optimisation of, for example, quadratic objective functions. For such problems it is easier to replace dL_t^M by $c^+\max\{M - X(t), 0\}dt$ and consider what happens as c^+ tends to infinity.

A more suitable loss function which accommodates local time as a result of the existence of upper and lower barriers is:

$$L(c) = l_0 c + \sqrt{1 + l_1^2(c - l_2)^2}. \tag{124}$$

Note that as $c \to +\infty$, $L(c) \sim (l_0 + l_1)c$, while as $c \to -\infty$, $L(c) \sim (l_0 - l_1)c$. This asymptotic linearity is required to ensure that the expected value of the loss function does not become infinite when a reflecting barrier and local time is introduced. If $l_0 < l_1$ then $L(c)$ is increasing and convex. In other words, the fund sponsor prefers to pay less rather than more and prefers stability to instability. Furthermore, the employer will be prepared to pay a higher average contribution rate in the long run in return for lower volatility in the contribution rate.

6.2. No short-selling of assets

Suppose that the holdings in each asset must be non-negative: that is, $0 \le p(t,x) \le 1$ for all t, x.

Let us consider the following piecewise linear model for the proportion of the fund invested in the more risky asset 2 in a 2-asset model:

$$c(x) = c_0 - c_1 x \text{ for all } x \tag{125}$$

$$p_2(x) = \begin{cases} 1 & \text{if } x < x_0 \\ \dfrac{x_0}{(x_1 - x_0)}\dfrac{(x_1 - x)}{x} & \text{if } x_0 \le x < x_1 \\ 0 & \text{if } x_1 \le x \end{cases} \tag{126}$$

In the unconstrained case:

$$p_2(x) = \frac{x_0}{(x_1 - x_0)}\frac{(x_1 - x)}{x} \tag{127}$$

for all x. The use of x_0 and x_1 here makes it easier to see where the constraints lock in. For asset 1 and cash we have:

$$p_1(x) = 1 - p_2(x) \tag{128}$$

$$p_0(x) = 0 \tag{129}$$

Over each interval $[0, x_0)$, $[x_0, x_1)$ and $[x_1, \infty)$ the stationary distribution function is a scaled Pearson type IV with different parameters over each interval. Since $p(x)$ is continuous the stationary density function is continuous. This allows numerical evaluation of an objective function and hence optimisation over c_0, c_1, x_0 and x_1.

Let us consider a numerical example. We use the same model parameter values and objective function as in Section 5. In the unconstrained problem the optimal solution is linear in x as usual. In the constrained problem the optimal solution will not be linear or piecewise linear in x, but here we optimise only over piecewise linear strategies.

It can be seen by referring to Table 4 that the effects of the constraints in this example are fairly small but, nevertheless, significant. The size of the effect of the constraint depends upon to what extent the interval $[x_0, x_1)$ comes into play in the unconstrained case. If $X(t)$ falls into $[x_0, x_1)$ most of the time then the effect of the constraint will be small. Here, in the unconstrained case, most of the time the fund is invested long in the low-risk asset 1 and short in the high-risk asset 2.

TABLE 4

STATIONARY OPTIMISATION UNDER CONSTRAINTS

Case	x_0	x_1	c_0	c_1	$E[X]$	$SD[X]$	$E[C]$	$SD[C]$	$E[L(C, X)]$
Constrained	3.87	7.40	1.77	0.123	9.29	1.50	0.628	0.184	0.0485
Unconstrained	3.96	7.89	1.68	0.105	9.85	1.35	0.645	0.142	0.0315

6.3. Upper limit on the contribution rate

Boulier *et al.* (1996) considered the effect of an upper bound on the optimal contribution rate. This resulted in a nearly linear form for $c(x)$ and a bell-shaped curve for $p(x)$. Their solution required the existence of a risk-free asset for the fund and zero volatility in the benefit outgo: otherwise the dynamics of the model would be non-stationary since a sufficiently large deficit will eventually build up which cannot be eradicated.

7. Conclusions

This paper has considered the optimal control of a pension fund using the asset-allocation strategy and the contribution strategy.

Optimal solutions have been derived for power and exponential loss functions (with no demographic risk $-\sigma_b = 0$) and, in more detail, for a quadratic loss function. In most cases the contribution strategy appears to be sensible and conforms with current practice. In each case aspects of the solution were not completely satisfactory. First power and exponential loss functions were found to give rise to non-stationary solutions. Second, when we considered the quadratic and exponential loss functions, the optimal asset-allocation strategy derived was rather counterintuitive: moving, say, out of equities into bonds when the level of surplus is growing.

This has one of two explanations. Funds may be operating in a very non-optimal way. Alternatively, they may be operating optimally but with different objectives. For example, in the UK, the government has recently introduced minimum funding legislation. This should lead to loss functions which heavily penalise events when the fund size falls below the legal minimum. Boulier *et al.* (1996) considered a related problem in which the contribution rate was subject to an upper constraint (say, twice the target rate). However, in the present framework (in which all assets are risky and where there is volatility in the benefit outgo) it is not possible to constrain the contribution rate in this way, for otherwise the fund size would ultimately drift off to minus infinity.

There is, however, some sense in a shift out of equities if the fund size is well above its target level. First if there is too much surplus then there will be pressure on the sponsoring employer to use this surplus to pay for discretionary pension increases which, perhaps, had not been promised. In any event the members would be benefitting from good investment returns while the employer has to pay when things go badly. Second if the employer is able to take a refund, the refund may be liable to tax (for example, in the UK this is 40% with the aim of inhibiting exploitation of the tax advantages enjoyed by a pension fund). Third, too much surplus may lead to the removal of part or all of the fund's special tax status (again this is the case in the UK). All of these reasons mean that it should be advantageous to put a bigger proportion of the fund into low-risk assets when the fund has a large surplus. The results described in this paper back up this viewpoint.

It is clear from the results contained in this paper that we must look for alternative loss functions. The target are ones which give rise to stationary solutions and sensible asset-allocation strategies.

The results presented in this paper and in that of Boulier *et al.* (1995) also draw attention to the following issues:

- what objective functions (if any) are used by pension funds? Are pension funds currently operating in a sub-optimal way or do they have different objective functions from the one considered here?

- what constraints (if any) on contributions and investments are appropriate? Can investment constraints be circumvented by prudent use of derivatives?
- is too much emphasis placed on the calculation of the so-called actuarial liability when this may have no relationship to the target funding level under the optimised objective function? Here the problems have been analysed solely with reference to the objective function. A framework which relies heavily on the actuarial liability might result in a solution which is sub-optimal with reference to this stated objective function.

It must be stressed again that we have assumed a stable membership structure in the pension plan. In many problems there may be a reason, for example, to incorporate changes in the membership as a sponsoring company evolves or restructures. Such situations would require an adaptation of, for example, the use of $X(t) = F(t)/W(t)$ as the key process and of the objective function. The findings in this paper suggest that the interests of the employer and fund members might be served better by a combination of dynamic control theory and more traditional actuarial valuation techniques. In this respect, theoretical solutions to simplified problems give us a basis for investigations of more complex situations.

ACKNOWLEDGEMENTS

I would like to thank the many people who have commented on earlier versions of this paper and at various talks. In particular, I have had useful conversations with or suggestions from Jean-Francois Boulier, Daniel Dufresne, Martin Schweizer, Wolfgang Rungaldier, Michael Taksar, Hanspeter Schmidli, Ravi Mazumdar, Paul Embrechts, Gary Parker, Martino Graselli, Mary Hardy, David Wilkie, Terence Chan, Iain Currie, Jack Carr, Ken Brown, and Andrew Lacey.

REFERENCES

BLACK, F. and JONES, R. (1988) Simplifying portfolio insurance for corporate pension plans. *Journal of Portfolio Management* **14(4)**, 33-37.

BLACK, F. and PEROLD, A. (1992) Theory of constant proportion portfolio insurance. *Journal of Economic Dynamics and Control* **16**, 403-426.

BOULIER, J-F., TRUSSANT, E. and FLORENS, D. (1995) A dynamic model for pension funds management. *Proceedings of the 5th AFIR International Colloquium* **1**, 361-384.

BOULIER, J-F., MICHEL, S. and WISNIA, V. (1996) Optimizing investment and contribution policies of a defined benefit pension fund. *Proceedings of the 6th AFIR International Colloquium* **1**, 593-607.

CAIRNS, A.J.G. (1995) Pension funding in a stochastic environment: the role of objectives in selecting an asset-allocation strategy. *Proceedings of the 5th AFIR International Colloquium* **1**, 429-453.

CAIRNS, A.J.G. (1996) Continuous-time stochastic pension fund modelling. *Proceedings of the 6th AFIR International Colloquium* **1**, 609-624.

CAIRNS, A.J.G., BLAKE, D. and DOWD, K. (2000). Optimal dynamic asset allocation for defined contribution pensionplans. To appear in Proceedings of the 10th AFIR International Colloquium.

CAIRNS, A.J.G. and PARKER, G. (1997) Stochastic pension fund modelling. *Insurance Mathematics and Economics* **21**, 43-79.

DAVIS, M.H.A. and NORMAN, A.R. (1990) Portfolio selection with transaction costs. *Mathematics of Operations Research* **15**, 676-713.

DEELSTRA, G., GRASSELLI, M. and KOEHL, P-F. (1999) Optimal investment strategies in a CIR framework. Preprint.

DE SCHEPPER, A., TEUNEN, M. and GOOVAERTS, M. (1994) An analytical inversion of a Laplace transform related to annuities certain. *Insurance: Mathematics and Economics* **14**, 33-37.

DUFRESNE, D. (1988) Moments of pension contributions and fund levels when rates of return are random. *Journal of the Institute of Actuaries* **115**, 535-544.

DUFRESNE, D. (1989) Stability of pension systems when rates of return are random. *Insurance: Mathematics and Economics* **8**, 71-76.

DUFRESNE, D. (1990) The distribution of a perpetuity, with applications to risk theory and pension funding. *Scandinavian Actuarial Journal* **1990**, 39-79.

FLEMING, W.H. and RISHEL, R.W. (1975) *Deterministic and stochastic optimal control.* Springer-Verlag, New York.

FÖLLMER, H. and SCHWEIZER, M. (1993) A microeconomic approach to diffusion models for stock prices. *Mathematical Finance* **3**, 1-23. (Erratum: **4**:285.)

GERBER, H.U. and SHIU, E.S. (2000) Investing for retirement: optimal capital growth and dynamic asset allocation. To appear in North American Actuarial Journal **4** (2).

HABERMAN, S. and SUNG, J-H. (1994) Dynamic approaches to pension funding. *Insurance: Mathematics and Economics* **15**, 151-162.

JOHNSON, N.L., KOTZ, S. and BALAKRISHNAN, N. (1994) *Continuous Univariate Distributions: Volume 1, 2nd Edition.* Wiley, New York.

KEEL, A. and MÜLLER, H.H. (1995) Efficient portfolios in the asset liability context. *ASTIN Bulletin* **25**, 33-48.

MERTON, R.C. (1971) Optimal consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory* **3**, 373-413.

MERTON, R.C. (1990) *Continuous-Time Finance.* Blackwell, Cambridge, Mass.

O'BRIEN, T. (1986) A stochastic-dynamic approach to pension funding. *Insurance: Mathematics and Economics* **5**, 141-146.

O'BRIEN, T. (1987) A two-parameter family of pension contribution functions and stochastie optimization. *Insurance: Mathematics and Economics* **6**, 129-134.

ØKSENDAL, B. (1998) *Stochastic Differential Equations, 5th Edition.* Springer-Verlag, Berlin.

SHARPE, W.F. and TINT, L.G. (1990) Liabilities – a new approach. *Journal of Portfolio Management* **Winter 1990**, 5-10.

SIEGMANN, A.H. and LUCAS, A. (1999) Continuous-time dynamic programming for ALM with risk-averse loss functions. *Proceedings of the 9th AFIR International Colloquium* **2**, 183-193.

SHREVE, S.E. and SONER, H.M. (1994) Optimal investment and consumption with transaction costs. *Annals of Applied Probability* **4**, 609-692.

WILKIE, A.D. (1985) Portfolio selection in the presence of fixed liabilities. *Journal of the Institute of Actuaries* **112**, 229-277.

WISE, A. (1984) The matching of assets to liabilities. *Journal of the Institute of Actuaries* **111**, 445-501.

YOR, M. (1992) Sur certaines foncionnelles exponentielles du mouvement brownien réel. *Journal of Applied Probability* **29**, 202-208.

ANDREW CAIRNS
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh EH14 4AS
Scotland

A MULTIVARIATE GENERALIZATION OF THE GENERALIZED POISSON DISTRIBUTION

BY

RALUCA VERNIC

"Ovidius" University of Constanta, Romania

ABSTRACT

This paper proposes a multivariate generalization of the generalized Poisson distribution. Its definition and main properties are given. The parameters are estimated by the method of moments.

KEYWORDS

Multivariate generalized Poisson distribution (MGP_m); generalized Poisson distribution (GPD); bivariate generalized Poisson distribution (BGPD).

1. INTRODUCTION

The univariate generalized Poisson distribution (GPD), introduced by CONSUL and JAIN (1973), is a well-studied alternative to the standard Poisson distribution. CONSUL (1989) provided a guide to the current state of modeling with the GPD at that time, and documented many real life examples. GPD has also been making appearances in the actuarial literature (see GERBER, 1990; GOOVAERTS and KAAS, 1991; KLING and GOOVAERTS, 1993; AMBAGASPITIYA and BALAKRISHNAN, 1994 etc.). A bivariate generalization was developed by VERNIC (1997) and was applied in the insurance field.

The multivariate generalization that we present in this paper is derived from the GPD in a similar way with the BGPD. In consequence, the BGPD can be obtained from the MGP_m for $m = 2$. In section 2 we present some properties of the MGP_m. The method of moments is used in section 3 for the estimation of the parameters. In section 4 the particular case of the BGPD is considered together with its application in the insurance field, based on the paper of VERNIC (1997) and illustrated with a numerical example. Since the BGPD is well fitted to the aggregate amount of claims for a compound class of policies submitted to claims of two kinds whose yearly frequencies are a priori dependent, it is natural to consider that the MGP_m is a good candidate for the aggregate amount of claims for a class of policies submitted to claims of m kinds.

2. The multivariate generalized Poisson distribution

2.1. Development of the distribution

If $N \sim GPD(\lambda, \theta)$, then its probability function (p.f.) is given by (Consul and Shoukri, 1985)

$$f(n) = P(N = n) = \begin{cases} \frac{1}{n!}\lambda(\lambda + n\theta)^{n-1}\exp\{-\lambda - n\theta\}, & n = 0,\ 1,\ \ldots \\ 0, & \text{for } n > q \text{ when } \theta < 0 \end{cases}, \quad (2.1)$$

and zero otherwise, where $\lambda > 0$, $\max(-1, -\lambda/q) \leq \theta < 1$ and $q \geq 4$ is the largest positive integer for which $\lambda + \theta q > 0$ when $\theta < 0$.

Vernic (1997) used the trivariate reduction method to construct the BGPD in the following way: let N_i, $i = 1,\ 2,\ 3$, be independent generalized Poisson random variables (r.v.), $N_i \sim GPD(\lambda_i, \theta_i)$, $i = 1,\ 2,\ 3$, and let $X = N_1 + N_3$ and $Y = N_2 + N_3$. Then $(X, Y) \sim BGPD(\lambda_i, \theta_i;\ i = 1,\ 2,\ 3)$.

Similarly, we obtain the m-dimensional generalized Poisson distribution by taking $(m + 1)$ independent generalized Poisson random variables, $N_i \sim GPD(\lambda_i, \theta_i)$, $i = 0,\ \ldots,\ m$, and considering $X_1 = N_1 + N_0,\ \ldots,\ X_m = N_m + N_0$. Then $(X_1,\ \ldots,\ X_m) \sim MGP_m(\Lambda, \Theta)$, where $\Lambda = (\lambda_0,\ \lambda_1,\ \ldots,\ \lambda_m)$ and $\Theta = (\theta_0,\ \theta_1,\ \ldots,\ \theta_m)$. This method can be called **the multivariate reduction method**, as an extension of the trivariate reduction method.

It is easy to see that the joint p.f. of $(X_1,\ \ldots,\ X_m)$ reads

$$p(x_1,\ \ldots,\ x_m) = P(X_1 = x_1,\ \ldots,\ X_m = x_m) =$$
$$= \sum_{k=0}^{\min\{x_1,\ldots,x_m\}} f_1(x_1 - k) \cdot \ldots \cdot f_m(x_m - k) f_0(k), \quad (2.2)$$

where f_i is the p.f. of the r.v. N_i.

Using (2.1) in (2.2) we get

$$p(x_1,\ \ldots,\ x_m) = \left(\prod_{j=0}^{m}\lambda_j\right)\exp\left\{-\lambda - \sum_{j=1}^{m} x_j\theta_j\right\}\cdot$$
$$\cdot\sum_{k=0}^{\min\{x_1,\ldots,x_m\}}\left(\prod_{j=1}^{m}\frac{[\lambda_j + (x_j - k)\theta_j]^{x_j-k-1}}{(x_j - k)!}\right)\cdot$$
$$\cdot\frac{(\lambda_0 + k\theta_0)^{k-1}}{k!}\exp\left\{k\left(\sum_{j=1}^{m}\theta_j - \theta_0\right)\right\}, \quad (2.3)$$
$$x_1,\ \ldots,\ x_m = 0,\ 1,\ 2,\ \ldots,$$

where $\lambda = \sum_{j=0}^{m}\lambda_j$ and $0! = 1$.

2.2. Properties of the distribution

We will first make some remarks on the GPD.

The GPD reduces to the Poisson distribution when $\theta = 0$ and it possesses the twin properties of over-dispersion and under-dispersion according as $\theta > 0$ or $\theta < 0$. When θ is negative, the GPD model includes a truncation due to the fact that $f(n) = 0$ for all $n > q$ (see 2.1). In the following, the moments expressions and the other formulas for the GPD are valid only for the case $\lambda > 0, 0 \leq \theta < 1$ and $q = \infty$, as discussed in SCOLLNIK (1998). This is a point frequently misrepresented in the literature.

In conclusion, we will assume for simplicity that $\theta > 0$. From AMBGASPITIYA and BALAKRISHNAN (1994) we have the following formulas for $N \sim GPD(\lambda, \theta)$:

- the probability generating function (p.g.f.)

$$\Pi_N(t) = \exp\left\{-\frac{\lambda}{\theta}[W(-\theta t \exp\{-\theta\}) + \theta]\right\}. \tag{2.4}$$

- the moment generating function (m.g.f.)

$$M_N(t) = \exp\left\{-\frac{\lambda}{\theta}[W(-\theta \exp\{-\theta + t\}) + \theta]\right\}, \tag{2.5}$$

where the Lambert W function is defined as $W(x)\exp\{W(x)\} = x$. For more details about this function see CORLESS et al. (1996).

- the first four central moments

$$\left\{\begin{array}{ll} E(N) = \mu_1 = \lambda M\,; & Var(N) = \mu_2 = \lambda M^3 \\ \mu_3 = \lambda(3M-2)M^4\,; & \mu_4 = 3\lambda^2 M^6 + \lambda(15M^2 - 20M + 6)M^5 \end{array}\right\}, \tag{2.6}$$

where $M = (1-\theta)^{-1}$.

The probability generating function of the MGP$_m$

Let now $\Pi_i(t)$ denote the p.g.f. of the r.v. N_i, $i = 0, \ldots, m$. Then the joint p.g.f. of $(X_1, \ldots, X_m)$ is

$$\begin{aligned}\Pi(t_1, \ldots, t_m) &= E\left(t_1^{X_1} \cdot \ldots \cdot t_m^{X_m}\right) = E\left(t_1^{N_1} \cdot \ldots \cdot t_m^{N_m}(t_1 \cdot \ldots \cdot t_m)^{N_0}\right) = \\ &= \Pi_1(t_1) \cdot \ldots \cdot \Pi_m(t_m)\Pi_0(t_1 \cdot \ldots \cdot t_m).\end{aligned} \tag{2.7}$$

Using (2.4) in (2.7) and assuming that $\theta_i > 0$, $i = 0, \ldots, m$, we have

$$\Pi(t_1, \ldots, t_m) = \exp\left\{-\sum_{i=1}^{m}\frac{\lambda_i}{\theta_i}W(-\theta_i t_i e^{-\theta_i}) - \frac{\lambda_0}{\theta_0}W\left(-\theta_0 e^{-\theta_0}\prod_{i=1}^{m} t_i\right) - \lambda\right\}.$$

The moment generating function of the MGP_m

If the m.g.f. of N_i is $M_i(t)$, $i = 0, \ldots, m$, then the m.g.f. of $(X_1, \ldots, X_m)$ is

$$M(t_1, \ldots, t_m) = E(\exp\{t_1 X_1 + \ldots + t_m X_m\}) = E\left(e^{t_1 N_1} \cdot \ldots \cdot e^{t_m N_m} e^{(t_1 + \ldots + t_m) N_0}\right)$$
$$= M_1(t_1) \cdot \ldots \cdot M_m(t_m) M_0(t_1 + \ldots + t_m). \tag{2.8}$$

Using (2.5) in (2.8), the joint m.g.f. is given for $\theta_i > 0$, $i = 0, \ldots, m$, by

$$M(t_1, \ldots, t_m) =$$
$$\exp\left\{ -\sum_{i=1}^{m} \frac{\lambda_i}{\theta_i} W(-\theta_i \exp\{-\theta_i + t_i\}) - \frac{\lambda_0}{\theta_0} W\left(-\theta_0 \exp\left\{-\theta_0 + \sum_{i=1}^{m} t_i\right\}\right) - \lambda \right\}.$$

Moments

Let $\mu_{r_1, \ldots, r_m} = E\left(\prod_{j=1}^{m} (X_j - EX_j)^{r_j}\right)$ be the $(r_1, \ldots, r_m)^{th}$ central moment of $(X_1, \ldots, X_m)$. The equation for $\mu_{r_1, \ldots, r_m}$ given $\mu_k^{(j)}$ the k^{th} central moment of N_j, $j = 0, \ldots, m$, results as follows

$$\mu_{r_1, \ldots, r_m} = E\left[\prod_{j=1}^{m} (N_j - EN_j + N_0 - EN_0)^{r_j}\right] =$$
$$= E\left[\prod_{j=1}^{m} \sum_{i_j=0}^{r_j} \binom{r_j}{i_j} (N_j - EN_j)^{i_j} (N_0 - EN_0)^{r_j - i_j}\right] =$$
$$= \sum_{(i_1, \ldots, i_m) = (0, \ldots, 0)}^{(r_1, \ldots, r_m)} \left(\prod_{j=1}^{m} \binom{r_j}{i_j} \mu_{i_j}^{(j)}\right) \mu^{(0)}_{\sum_{j=1}^{m} (r_j - i_j)}. \tag{2.9}$$

From (2.6) and the independence of N_j, $j = 0, \ldots, m$, we also have for $\theta_i > 0$, $i = 0, \ldots, m$,

$$\begin{cases} EX_i = \lambda_i M_i + \lambda_0 M_0 \\ Var(X_i) = \lambda_i M_i^3 + \lambda_0 M_0^3 \end{cases}, \quad i = 1, \ldots, m, \tag{2.10}$$

and from (2.9) we have, for example

$$\begin{cases} \mu_{110\ldots0} = \mu_{0\ldots010\ldots010\ldots0} = \mu_2^{(0)} = \lambda_0 M_0^3 \\ \mu_{1110\ldots0} = \mu_{0\ldots010\ldots010\ldots010\ldots0} = \mu_3^{(0)} = \lambda_0 (3M_0 - 2) M_0^4 \\ \cdot \\ \cdot \\ \cdot \\ \mu_{11\ldots1} = \mu_m^{(0)} \end{cases}. \tag{2.11}$$

Marginal distributions

The marginal distributions are

$$P(X_i = r) = P(N_i + N_0 = r) = \lambda_0\lambda_i \exp\{-(\lambda_0 + \lambda_i) - r\theta_0\}\cdot$$
$$\cdot \sum_{j=0}^{r} \frac{1}{j!(r-j)!}(\lambda_i + j\theta_i)^{j-1}(\lambda_0 + (r-j)\theta_0)^{r-j-1}\exp\{-j(\theta_i - \theta_0)\}\ , \quad i = 1, \ldots, m.$$

In particular, if $\theta_i = \theta_0 = \theta$, this reduces to $X_i \sim GPD(\lambda_i + \lambda_0, \theta)$. Elsewhere, X_i is not a GPD.

Remark. From the development of the MGP$_m$, it is easy to see that if $(X_1, \ldots, X_m) \sim MGP_m(\Lambda, \Theta)$, then for any $\{i_1, \ldots, i_k\} \subset \{1, \ldots, m\}$ with $2 \le k < m$, $(X_{i_1}, \ldots, X_{i_k}) \sim MGP_k(\Lambda', \Theta')$, where $\Lambda' = (\lambda_0, \lambda_{i_1}, \ldots, \lambda_{i_k})$ and $\Theta' = (\theta_0, \theta_{i_1}, \ldots, \theta_{i_k})$.

For $k = 1$ the remark is not always true. But if we consider the particular case $\theta_0 = \theta_1 = \ldots = \theta_m = \theta$, then from $(X_1, \ldots, X_m) \sim MGP_m(\Lambda, \Theta)$ it follows that $X_i \sim GPD(\lambda_i + \lambda_0, \theta)$, $i = 1, \ldots, m$.

Recurrence relations

The marginal p.f. can be computed using the univariate generalized Poisson distribution, as it is seen from

$$p(0, \ldots, 0) = \exp\{-\lambda\}$$

$$p(0, \ldots, 0, \underset{j}{x_j}, 0, \ldots, 0) = f_j(x_j)\left(\prod_{\substack{i=1\\ i\neq j}}^{m} f_i(0)\right) f_0(0) =$$
$$= f_j(x_j)\exp\{-(\lambda - \lambda_j)\},\ j = 1, \ldots, m, \quad x_j > 0.$$

Given these probabilities, for $x_j > 0$, $j = 1, \ldots, m$, we have the following recurrence relation

$$p(x_1, \ldots, x_m) = \lambda_0 \exp\{(m-1)\lambda\} \sum_{k=0}^{\min\{x_1, \ldots, x_m\}} \left(\prod_{j=1}^{m} p(0, \ldots, 0, x_j - k, 0, \ldots, 0)\right)\cdot$$
$$\cdot \frac{(\lambda_0 + k\theta_0)^{k-1}}{k!} \exp\{-k\theta_0\}.$$

3. ESTIMATION OF THE PARAMETERS: METHOD OF MOMENTS

Let $(x_{1i}, \ldots, x_{mi})$, $i = 1, \ldots, n$ be a random sample of size n from the population. We will assume that the frequency of the m-tuple $(s_1, \ldots, s_m)$ is $n_{s_1,\ldots,s_m}$ for $s_1, \ldots, s_m = 0, 1, \ldots$. We recall that $\sum_{s_1,\ldots,s_m} n_{s_1,\ldots,s_m} = n$. Also

$$\begin{cases} n_{+\ldots+s_j+\ldots+} = \displaystyle\sum_{\{s_k|k=1,\ldots,m,\ k\neq j\}} n_{s_1,\ldots,s_m} \\ n_{+\ldots+s_i+\ldots+s_j+\ldots+} = \displaystyle\sum_{\{s_k|k=1,\ldots,m,\ k\neq j,\ k\neq i\}} n_{s_1,\ldots,s_m}\ , \quad i<j \end{cases}. \tag{3.1}$$

We denote

$$\begin{cases} \bar{x}_j = \dfrac{1}{n}\displaystyle\sum_{s_j} s_j n_{+\ldots+s_j+\ldots+} \\ \hat{\sigma}_j^2 = \dfrac{1}{n}\displaystyle\sum_{s_j} (s_j - \bar{x}_j)^2 n_{+\ldots+s_j+\ldots+} \end{cases}, \quad j = 1, \ldots, m, \tag{3.2}$$

and, with the notations in (3.1)

$$\begin{cases} \overline{x_i x_j} = \dfrac{1}{n}\displaystyle\sum_{s_i,s_j} s_i s_j n_{+\ldots+s_i+\ldots+s_j+\ldots+}\ , \quad i<j \\ \overline{x_i x_j x_k} = \dfrac{1}{n}\displaystyle\sum_{s_i,s_j,s_k} s_i s_j s_k n_{+\ldots+s_i+\ldots+s_j+\ldots+s_k+\ldots+}\ , \quad i<j<k \end{cases}.$$

It is easy to see that

$$\begin{cases} \mu_{0\ldots\underset{i}{0}1\!\!\!\!\,0\ldots\underset{j}{0}1\!\!\!\!\,0\ldots0} = E(X_iX_j) - E(X_i)E(X_j), \quad i<j \\ \mu_{0\ldots\underset{i}{1}0\ldots\underset{j}{1}0\ldots\underset{k}{1}0\ldots0} = E(X_iX_jX_k) - E(X_iX_j)E(X_k) - E(X_jX_k)E(X_i) - \\ \qquad -E(X_iX_k)E(X_j) + 2E(X_i)E(X_j)E(X_k), \quad i<j<k \end{cases},$$

so we can use the sample moments

$$\begin{cases} \hat{\mu}_{0\ldots\underset{i}{1}0\ldots\underset{j}{1}0\ldots0} = \overline{x_i x_j} - \bar{x}_i\bar{x}_j\ , \quad i<j \\ \hat{\mu}_{0\ldots\underset{i}{1}0\ldots\underset{j}{1}0\ldots\underset{k}{1}0\ldots0} = \overline{x_i x_j x_k} - \overline{x_i x_j}\,\bar{x}_k - \overline{x_i x_k}\,\bar{x}_j - \overline{x_j x_k}\,\bar{x}_i + \\ \qquad +2\bar{x}_i\bar{x}_j\bar{x}_k\ , \quad i<j<k \end{cases} \tag{3.3}$$

The general method

The classical method of moments consists of equating the sample moments to their populations equivalents, expressed in terms of the parameters. The number of moments required is equal to the number of parameters which equals $2(m+1)$. For example, using (2.10), (2.11), (3.2) and (3.3), we can choose the following $2(m+1)$ equations

$$\begin{cases} \bar{x}_j = \lambda_j M_j + \lambda_0 M_0 \\ \hat{\sigma}_j^2 = \lambda_j M_j^3 + \lambda_0 M_0^3 \\ \hat{\mu}_{110\ldots0} = \lambda_0 M_0^3 \\ \hat{\mu}_{1110\ldots0} = \lambda_0(3M_0 - 2)M_0^4 \end{cases}, \quad j = 1, \ldots, m.$$

Denoting $a = \dfrac{\hat{\mu}_{1110\ldots0}}{\hat{\mu}_{110\ldots0}}$, the solution of the system is

$$\begin{cases} M_0 = \dfrac{1+\sqrt{1+3a}}{3} \\ \lambda_0 = \dfrac{\hat{\mu}_{110\ldots0}}{M_0^3} \\ M_j = \sqrt{\dfrac{\hat{\sigma}_j^2 - \hat{\mu}_{110\ldots0}}{\bar{x}_j - \lambda_0 M_0}} \\ \lambda_j = \dfrac{\bar{x}_j - \lambda_0 M_0}{M_j} \end{cases}, \quad j = 1, \ldots, m. \tag{3.4}$$

We used the fact that $\theta < 1$, so $M = \frac{1}{1-\theta} > 0$.

Particular case: $\theta_0 = \theta_1 = \ldots = \theta_m = \theta$, so $M_0 = M_1 = \ldots = M_m = M$.

Method I. The number of parameters is now $(m+2)$: $\lambda_0, \ldots, \lambda_m$ and M, so we can use the following equations:

$$\begin{cases} \bar{x}_j = (\lambda_j + \lambda_0)M \\ \hat{\mu}_{110\ldots0} = \lambda_0 M^3 \\ \hat{\mu}_{1110\ldots0} = \lambda_0(3M-2)M^4 \end{cases}, \text{with the solution} \begin{cases} M = \dfrac{1+\sqrt{1+3a}}{3} \\ \lambda_0 = \dfrac{\hat{\mu}_{110\ldots0}}{M^3} \\ \lambda_j = \dfrac{\bar{x}_j}{M} - \lambda_0 \end{cases}, \quad j = 1, \ldots, m.$$

Method II. Another possibility is to use the method of moments in combination with the zero cell frequency method. If we denote by $f_{0\ldots0} = \dfrac{n_{0\ldots0}}{n}$ the frequency of the cell $(0, \ldots, 0)$, we can consider the system

$$\begin{cases} I. \quad f_{0\ldots0} = \exp\{-(\lambda_0 + \ldots + \lambda_m)\} \\ II. \quad \bar{x}_j = (\lambda_j + \lambda_0)M \\ III. \quad \hat{\sigma}_j^2 = (\lambda_j + \lambda_0)M^3 \end{cases}, \quad j = 1, \ldots, m.$$

We have here $(2m+1)$ equations. By summing equations I and II separately, we get

$$\begin{cases} IV.\ \sum_{j=1}^{m} \bar{x}_j = \left(\sum_{j=1}^{m} \lambda_j + m\lambda_0\right) M \\ V.\ \sum_{j=1}^{m} \hat{\sigma}_j^2 = \left(\sum_{j=1}^{m} \lambda_j + m\lambda_0\right) M^3 \end{cases}, \quad j = 1, \ldots, m.$$

Dividing the two relations gives $M^2 = \left(\sum_{j=1}^{m} \hat{\sigma}_j^2\right)\left(\sum_{j=1}^{m} \bar{x}_j\right)^{-1}$, hence the solution

$$M = \sqrt{\left(\sum_{j=1}^{m} \hat{\sigma}_j^2\right)\left(\sum_{j=1}^{m} \bar{x}_j\right)^{-1}}. \tag{3.5}$$

From equation I we have

$$-\ln f_{0\ldots0} = \lambda_0 + \sum_{j=1}^{m} \lambda_j ,$$

and using equation IV we are lead to

$$-\ln f_{0\ldots0} = \lambda_0 + \frac{1}{M}\sum_{j=1}^{m} \bar{x}_j - m\lambda_0 ,$$

so that

$$\lambda_0 = \frac{1}{m-1}\left(\frac{1}{M}\sum_{j=1}^{m} \bar{x}_j + \ln f_{0\ldots0}\right). \tag{3.6}$$

Then, from equation II we have

$$\lambda_j = \frac{1}{M}\bar{x}_j - \lambda_0 , \ j = 1, \ldots, m. \tag{3.7}$$

Finally, the solution $(M, \lambda_0, \lambda_j, j = 1, \ldots, m)$ is given by (3.5), (3.6) and (3.7).

Remark. In method II, the estimation of M is based on the empirical moments from all m variables, while in method I only three variables are taken into consideration by $\hat{\mu}_{1110\ldots0}$.

4. Particular case: Bivariate generalized Poisson distribution (BGPD)

Considering $m = 2$, the multivariate generalized Poisson distribution reduces to the bivariate generalized Poisson distribution. The BGPD was introduced by Vernic (1997) and was applied in the insurance field. The distribution was fitted to the aggregate amount of claims for a compound class of policies submitted to claims of two kinds whose yearly frequencies are a priori dependent. A comparative study with the classical bivariate Poisson distribution and with two bivariate mixed Poisson distributions has been carried out, based on two sets of data concerning natural events insurance in the U.S.A. and third party liability automobile insurance in France. The conclusion, after applying the χ^2 goodness-of-fit test, is that the BGPD fits better to the data, so it can be considered as a valid alternative to the usual bivariate Poisson or mixed Poisson distributions. For more details see Vernic (1997).

In the following, we will consider another example, based on the accident data of Cresswell and Frogatt (1963), with X_1 as the accidents in the first period and X_2 as the accidents in the second period. The data are given in table 1, first row in each cell.

The summary statistics for these data are:

$$\bar{x}_1 = 1.0014, \quad \bar{x}_2 = 1.291, \quad \hat{\sigma}_1^2 = 1.1935, \quad \hat{\sigma}_2^2 = 1.5961,$$
$$\hat{\mu}_{11} = 0.3258, \quad \hat{\mu}_{21} = 0.365.$$

Under the hypothesis $(X_1, X_2) \sim BGPD(\lambda_0, \lambda_1, \lambda_2; \theta_0, \theta_1, \theta_2)$, we have from (3.4)

$$\left\{ \begin{array}{lll} \theta_0 = 0.0286, & \theta_1 = 0.1057, & \theta_2 = 0.1200 \\ \lambda_0 = 0.2987, & \lambda_1 = 0.6206, & \lambda_2 = 0.8653 \end{array} \right\}.$$

The theoretical frequencies in this case are given in table 1, second row in each cell. After grouping in 32 categories: $(i,j)_{i=0..4;\, j=0..5}$; (0..4, 6 and above); (5 and above, 0 and above), we obtain $\chi^2_{obs} = \sum (obs - th)^2/th = 25.935$ and a significance level (P-value) verifying $0.45 \le \hat{\alpha} \le 0.75$. So the distribution is adequate.

We will now consider the particular case $\theta_0 = \theta_1 = \theta_2 = \theta$, so that we have the hypothesis $(X_1, X_2) \sim BGPD(\lambda_0, \lambda_1, \lambda_2; \theta)$. From (3.5), (3.6) and (3.7) we have $\theta = 0.0935$, $\lambda_0 = 0.2778$, $\lambda_1 = 0.63$, $\lambda_2 = 0.8925$, and the theoretical frequencies are given in table 1, last row in each cell. For the same categories we have $\chi^2_{obs} = 23.6082$ and $0.7 \le \hat{\alpha} \le 0.85$, so this particular distribution fits even better than the general one.

Acknowledgement

The author is grateful to the referees for their helpful advice in revising the paper.

TABLE 1

COMPARISON OF OBSERVED AND THEORETICAL FREQUENCIES

X_2 / X_1	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	Σ
0	117	96	55	19	2	2	0	0	291
	118.843	91.204	44.710	17.959	6.460	2.171	0.697	0.217	282.261
	117	95.100	46.748	18.088	6.081	1.865	0.537	0.148	285.567
1	61	69	47	27	8	5	1	0	218
	66.356	85.419	51.437	23.005	8.820	3.087	1.019	0.324	239.467
	67.132	84.165	50.881	22.205	8.065	2.608	0.780	0.220	236.056
2	34	42	31	13	7	2	3	0	132
	24.834	38.319	30.090	15.577	6.505	2.402	0.822	0.267	118.816
	24.976	37.584	30.048	15.739	6.427	2.249	0.711	0.209	117.943
3	7	15	16	7	3	1	0	0	49
	7.871	13.249	12.124	7.260	3.386	1.341	0.480	0.161	45.872
	7.694	12.602	12.004	7.911	3.849	1.516	0.520	0.162	46.258
4	3	3	1	1	2	1	1	1	13
	2.287	4.040	3.860	2.610	1.676	0.616	0.226	0.079	15.394
	2.138	3.685	3.774	2.927	1.799	0.844	0.327	0.111	15.605
5	2	1	0	0	0	0	0	0	3
	0.632	1.149	1.142	0.816	0.464	0.220	0.090	0.033	4.546
	0.558	0.995	1.075	0.910	0.647	0.382	0.176	0.068	4.811
6	0	0	0	0	1	0	0	0	1
	0.169	0.313	0.319	0.236	0.140	0.071	0.031	0.012	1.291
	0.140	0.255	0.285	0.255	0.198	0.136	0.079	0.036	1.384
7	0	0	0	1	0	0	0	0	1
	0.044	0.083	0.086	0.065	0.040	0.021	0.010	0.004	0.353
	0.034	0.063	0.072	0.067	0.055	0.041	0.028	0.016	0.376
	224	226	150	68	23	11	5	1	
Σ	221.036	233.776	143.768	67.528	27.491	9.929	3.375	1.097	708
	219.672	234.449	144.887	68.102	27.121	9.641	3.158	0.970	

REFERENCES

AMBAGASPITIYA, R.S. and BALAKRISHNAN, N. (1994) On the compound generalized Poisson distributions. *ASTIN Bulletin* **24**, 255-263.

CONSUL, P.C. (1989) *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker Inc., New York/Basel.

CONSUL, P.C. and JAIN, G.C. (1973) A generalization of the Poisson distribution. *Technometrics* **15**, 791-799.

CONSUL, P.C. and SHOUKRI, M.M. (1985) The generalized Poisson distribution when the sample mean is larger than the sample variance. *Communications in Statistics - Simulation and Computation* **14**, 1533-1547.

CORLESS, R.M., GONNET, G.H., HARE, D.E.G., JEFFREY, D.J. and KNUTH, D.E. (1996) On the Lambert W function. *Advances in Computational Mathematics* **5**, 329-359.

CRESSWELL, W.L. and FROGGATT, P. (1963) *The causation of bus driver accidents. An epidemiological study*. Oxford University Press, London.

GERBER, H.U. (1990) When does the surplus reach a given target? *Insurance: Mathematics and Economics* **9**, 115-119.

GOOVAERTS, M.J. and KAAS, R. (1991) Evaluating compound generalized Poisson distributions recursively. *ASTIN Bulletin* **21**, 193-197.

KLING, B. and GOOVAERTS, M. (1993) A note on compound generalized distributions. *Scandinavian Actuarial Journal* **1**, 60-72.

SCOLLNIK, D.P.M. (1998) On the analysis of the truncated generalized Poisson distribution using a Bayesian method. *ASTIN Bulletin* **28**, 135-152.

VERNIC, R. (1997) On the bivariate generalized Poisson distribution. *ASTIN Bulletin* **27**, 23-31.

RALUCA VERNIC
Dept. of Mathematics and Informatics
"Ovidius" University of Constanta
124, Bd. Mamaia
8700 Constanta, Romania
E-mail: rvernic@univ-ovidius.ro

A MATHEMATICAL MODEL OF ALZHEIMER'S DISEASE AND THE APOE GENE

BY

ANGUS MACDONALD AND DELME PRITCHARD

Heriot-Watt University, Edinburgh

ABSTRACT

Alzheimer's disease (AD) accounts for a significant proportion of long-term care costs. The recent discovery that the $\varepsilon 4$ allele of the ApoE gene indicates a predisposition to earlier onset of AD raises questions about the potential for adverse selection in long-term care insurance, about long-term care costs in general, and about the potential effects on costs of gene therapy, or better targetted treatments for AD. This paper describes a simple Markov model for AD, and the estimation of the transition intensities from the medical and epidemiological literature.

KEYWORDS

Alzheimer's Disease, Apolipoprotein E Gene, Genetics, Markov Models.

1. INTRODUCTION

Molecular genetics has far-reaching implications for all aspects of health economics, including the effectiveness, or even practicability, of insurance-based funding of all forms of care. In particular, if genetic test results are known by the individual but not by the insurer, there may be a tendency for those most at risk to buy more insurance, known as adverse selection. This is a natural subject for quantitative modelling.

Such discussion as has taken place (in the United Kingdom) has tended to be confined to life insurance, in which context it has been suggested that the overall costs of adverse selection might be limited (Macdonald, 1997, 1999; Pritchard, 1997). However, these papers acknowledge:

(a) that different conclusions might hold in respect of other forms of insurance; and
(b) the lack of sound epidemiological data in respect of any but a few genetic conditions.

Thus, the model proposed by Macdonald (1997, 1999) was based on parameter estimates intended to be extreme, and to suggest an upper bound on the cost of adverse selection, rather than on data-based statistical estimates. Tan (1997) applied a similar model to annuity business, with results that suggested higher costs.

The U.K. Government has so far avoided legislating on the very sensitive issue of the use of genetic test information by insurers, but instead has set up the Genetics and Insurance Committee (GAIC), charged with the assessment of the likely relevance and reliability of genetic test information as it relates to different kinds of insurance. Such considerations require actuarial models of the insurance process, allowing for the effects of specific genes on mortality and morbidity. To date, the only such study is Lemaire *et al.* (1999) and Subramanian & Lemaire (1999), treating breast and ovarian cancer and the BRCA1 and BRCA2 genes.

In this paper we propose a simple Markov model for Alzheimer's disease (AD) and estimate its transition intensities from medical and epidemiological studies. Genetic variation arises because the $\varepsilon 4$ allele of the Apolipoprotein E (ApoE) gene is known to indicate a predisposition to earlier onset of AD.

In Section 2 we give a brief outline of genetics terminology, then describe AD and the evidence for a genetic component of AD. In Section 3 we specify the model, and in Sections 4 and 5 we estimate the transition intensities. We show some results in Section 6, and our conclusions are in Section 7.

Applications of the model include the study of long-term care insurance, (Macdonald & Pritchard, 1999) and the study of long-term care costs (Warren *et al.*, 1999). The processes leading to LTC insurance claims are complex, when compared with other forms of insurance, and there are no insurance data to speak of; therefore we (and insurance companies themselves) have to rely on data collected and published for a variety of reasons, mostly in the medical literature. As a result, our model is far from definitive, but we found the process of extracting information from the medical literature and putting it to actuarial use very instructive, and we suggest that any shortcomings of our model shed useful light on the problems that might be faced in the future.

2. The Genetics of Alzheimer's Disease

For a recent survey of the genetic epidemiology of AD, see Slooter & van Duijn (1997); Breteler *et al.* (1992), reviews the position before much was known about the genetic component of AD.

2.1. A Brief Primer on Human Genetics

The following discussion gives only the briefest definitions of genetic terms, and omits all complications. For a proper account, see a text such as Strachan & Read (1999).

(a) Each cell of the human body (except sperm and egg cells) has a nucleus containing 23 pairs of chromosomes. Each cell carries an identical set of chromosomes at birth, unless some have been damaged.
(b) A chromosome is a very long double strand (the famous double helix) or linear molecule, made up of a sequence of base pairs of DNA.
(c) A gene is a sequence of base pairs at a fixed location on a given chromosome that acts as a functional unit, coding for the production of a protein (usually). The term 'gene' strictly refers to the locus on the chromosome.
(d) The gene at a given locus might have several variants (rather as chemical elements might have several isotopes); these are known as alleles. Different alleles are simply different sequences of base pairs. Differences between alleles are the result of mutations, namely alterations to a gene caused by (for example) errors in replication of the DNA when cells divide. If the mutation occurs in an adult (somatic) cell, it can only spread through cell division (which is what happens in cancers). If the mutation occurs in a gene carried by a sperm or egg cell, it can be passed to offspring.
(e) Because the chromosomes are paired (except the X and Y chromosomes that determine sex) each cell has two alleles of every gene, one inherited from each parent. They may be the same (homozygosity) or different (heterozygosity).
(f) Mutant genes will encode variants of the protein that the gene produces. These may have a beneficial, neutral or adverse effect on the cell or organism. If the altered protein is radically different, the cell will probably die, or the organism will fail to develop, or it will not survive to reproductive age. Less radical variations may manifest themselves as susceptibility to disease, or may be harmless but non-functional.
(g) The great range of genetic diseases arises from the range of effects of the protein products of different alleles, and the simple combinatorics of inheritance.
    (1) If an allele encodes a harmless but non-functional protein product, the disease will appear only in homozygous individuals (autosomal recessive inheritance, such as cystic fibrosis).
    (2) Heterozygous cells will produce a mixture of variants of the protein product; if just one of these is lethal it will cause disease (autosomal dominant inheritance, such as Huntington's disease).
    (3) In between these extremes, alleles encode protein products that are more or less dangerous or fully-functional, the effect often depending on the environment. Then different levels of susceptibility to disease will appear, and homozygous individuals will often be more susceptible than heterozygous individuals.
(h) Lifestyle and environment can affect the potency of susceptibility genes. For example, the activity of the protein produced by a dangerous allele may be enhanced in certain environments, or a protective allele may be put at greater risk of being knocked out by a mutation.

(i) Although the outline above mentions only single genes, most processes in the human body are complicated and involve very many proteins, each encoded by its own gene with its various alleles. Most genetic disease will result from combinations of several alleles, lifestyle and environment; the term for this is multifactorial disorder.

2.2. Cognitive Impairment

The term 'cognitive impairment' covers AD, which accounts for by far the greater number of cases, and other forms of mental deterioration, chiefly vascular in origin (for example, arising from strokes). Assessment is liable to be imprecise, making it difficult to decide on an exact date of inception of cognitive impairment, if such a thing exists. Moreover, although AD is the commonest form of cognitive impairment, it is hard to diagnose with certainty except by post-mortem examination. These factors introduce considerable uncertainty into epidemiological studies of AD. Breteler *et al.* (1992) noted that:
(a) AD itself can have a significant vascular component;
(b) some of the neuropathological symptoms of AD (see Section 2.4) can also be symptoms of vascular dementia; and
(c) studies by Tierney *et al.* (1988) found that post-mortem examination confirmed only 64-86% of diagnoses of AD.

2.3. The Pathology of Alzheimer's Disease

The pathology of AD includes:
(a) senile plaques (deposits on the outside of neurones (brain cells), consisting largely of the protein β-amyloid);
(b) neurofibrillary tangles (connections between neurones);
(c) amyloid angiopathy (deposits of amyloid protein in the arteries of the brain);
(d) loss of neurones; and
(e) decreased activity of choline acetyltransferase (an enzyme).

Therefore, any gene whose expression leads to the production, or over-production, of substances associated with these changes is potentially a genetic marker for AD.

2.4. General Evidence of a Genetic Contribution to Alzheimer's Disease

AD is a disease of old age; it is rare below ages 60-70. These rare cases are called 'early-onset' AD, which should not be confused with early onset of AD within the usual age range. We are concerned only with the latter.

Families with a history of AD are sometimes observed, but AD also occurs sporadically (that is, in the absence of a family history of AD) and it is always possible that a case of AD in an affected family is, in fact, sporadic. The differences between familial and sporadic AD are not clear, although the former may be marked by earlier onset and more rapid progression.

A very few families have several cases of early-onset AD in several generations, consistent with autosomal dominant transmission (Levy-Lahad & Bird, 1996), and three genes have been found. First was the gene encoding for amyloid precursor protein (APP), involved in the production of β-amyloid. It resides on chromosome 21, which is the chromosome affected in cases of Down syndrome, sufferers of which often develop AD in middle age. Several mutations have been found, but they are rare. Later, mutations in two genes labelled presenilin-1 (PS-1) and presenilin-2 (PS-2) were identified, which appeared to be associated with AD, though the mechanisms remain unclear.

Familial AD is not restricted to early-onset cases and family history remains an important risk factor for late-onset AD (Jarvik *et al.*, 1996). Susceptibility genes have been identified, of which the most studied is that which codes for apolipoprotein E.

2.5. The Apolipoprotein E Gene

This summary is not comprehensive, but we hope that it is detailed enough to give an actuarial reader an impression of the progress made in understanding genetic factors of AD, as well as some of the problems and (perhaps most important) the great speed at which human genetics is advancing.

Apolipoprotein E (ApoE) is found in senile plaques and neurofibrillary tangles in AD patients. It has also been studied because of its rôle in lipid metabolism. The gene that encodes it is on chromosome 19, which was linked to families with late-onset AD by Pericak-Vance *et al.* (1991), making it a clear candidate gene for familial AD. Strittmatter *et al.* (1993) confirmed this hypothesis, which was rapidly supported by many other studies. The basic facts are as follows:

(a) The ApoE gene has three common alleles – $\varepsilon 2$, $\varepsilon 3$ and $\varepsilon 4$ – whose frequencies are roughly 0.09, 0.77 and 0.14 respectively.
(b) Since each offspring receives one allele from each parent, there are six possible genotypes ($\varepsilon 2/\varepsilon 2$, $\varepsilon 2/\varepsilon 3$, $\varepsilon 2/\varepsilon 4$, $\varepsilon 3/\varepsilon 3$, $\varepsilon 3/\varepsilon 4$ and $\varepsilon 4/\varepsilon 4$). Offspring with two copies of the same allele are called homozygotes, while those with two different alleles are called heterozygotes.
(c) The ApoE $\varepsilon 4$ allele increases the risk of AD in a dose related fashion, such that $\varepsilon 4$ homozygotes ($\varepsilon 4/\varepsilon 4$) are at greater risk than $\varepsilon 4$ heterozygotes ($\varepsilon 2/\varepsilon 4$, $\varepsilon 3/\varepsilon 4$), who in turn are at greater risk than those without the $\varepsilon 4$ allele (Bickeböller *et al.*, 1997; Corder *et al.*, 1994; van Duijn *et al.*, 1995; Farrer *et al.*, 1997; Jarvik *et al.*, 1996; Kuusisto *et al.*, 1994; Lehtovirta *et al.*, 1995; Mayeux *et al.*, 1993; Myers *et al.*, 1996; Poirier *et al.*, 1993; Tsai *et al.*, 1994). See Section 5.2

for risk estimates. The risk depends on age, being highest at ages 60-70, tapering off at older ages (Bickeböller *et al.*, 1997; Corder *et al.*, 1994; Farrer *et al.*, 1997).

(d) It is also possible that the $\varepsilon 4$ allele is associated with earlier onset of AD (not to be confused with early-onset AD). The effect may be dose dependent (Farrer *et al.*, 1997; Frisoni *et al.*, 1995; Gomez-Isla *et al.*, 1996); or not (Lehtovirta *et al.*, 1995; Stern *et al.*, 1997; Corder *et al.*, 1995); or it may not exist at all (Liddell *et al.*, 1994; Masullo *et al.*, 1998; Norrman *et al.*, 1995).

(e) Investigations into the rate of mental decline of AD patients by genotype found no evidence for any difference (Basun *et al.*, 1995; Gomez-Isla *et al.*, 1996; Masullo *et al.*, 1998; Norrman *et al.*, 1995). There is conflicting evidence about mortality. It is possible that younger age at onset should imply longer survival times, because of the usual age-related mortality differentials, and therefore that the $\varepsilon 4$ allele should be associated with longer life after onset of AD. While some studies support this (Corder *et al.*, 1995; Gomez-Isla *et al.*, 1996; Norrman *et al.*, 1995), others have found no difference (Basun *et al.*, 1995; Stern *et al.*, 1997). If $\varepsilon 4$ is associated with lighter mortality in AD patients then risk estimates from cross-sectional studies (the vast majority to date) should be interpreted with caution. An incidence study (Evans *et al.*, 1997) confirmed $\varepsilon 4$ to be a significant risk factor, but the estimated increased risk of onset was at the lower end of the reported range.

(f) In contrast, the $\varepsilon 2$ allele has been found to have a protective effect against late-onset AD (Corder *et al.*, 1994; Farrer *et al.*, 1997; Gomez-Isla *et al.*, 1996; Jarvik *et al.*, 1996; Lambert *et al.*, 1998; Masullo *et al.*, 1998). However, a study of early-onset AD patients (van Duijn *et al.*, 1995), found a higher frequency of the $\varepsilon 2$ allele, and an association of $\varepsilon 2$ with a more aggressive form of AD, suggesting different rôles of ApoE in early-onset and late-onset AD. Findings relating to the $\varepsilon 2$ allele are based on the $\varepsilon 2/\varepsilon 3$ genotype, as $\varepsilon 2$ homozygotes are rare. The risk attached to the $\varepsilon 2/\varepsilon 4$ genotype is not clear, possibly because $\varepsilon 2$ and $\varepsilon 4$ have opposite effects (Jarvik *et al.*, 1996; Levy-Lahad *et al.*, 1996).

ApoE $\varepsilon 4$ is the most important genetic risk factor for AD identified yet. Though it is neither necessary nor sufficient to cause AD it does increase susceptibility. Approximately 26% of Caucasians carry at least one $\varepsilon 4$ allele and it has been estimated that between 42% and 79% of AD cases are attributable to the associated excess risk (Nalbantoglu *et al.*, 1994).

2.6. Other Genetic Factors of Alzheimer's Disease

In 1997, a gene for the K-variant of butyrylcholinesterase (BCHE K), not a risk factor by itself, was found to act in synergy with ApoE $\varepsilon 4$, such that carriers of both (an estimated 6% of Caucasians) were at over 30 times the risk of AD as a person with neither (Lehmann *et al.*, 1997). Subsequent studies (Brindle *et al.*, 1998; Singleton *et al.*, 1998) failed to reproduce the result. Although some explanations have been advanced, caution is advisable in using BCHE K as a risk factor for AD.

Payami *et al.* (1997) reported an association between AD and the A2 allele of the human leukocyte antigen (HLA); the HLA-A2 phenotype and ApoE $\varepsilon4/\varepsilon4$ genotype had similar and additive effects on reducing age at onset of AD, at ages below 60 and above 75. Further studies would be needed to confirm these findings.

Poduslo *et al.* (1998) found the apolipoprotein CI (apo CI) gene to be a risk factor for early-onset and late-onset AD, whether sporadic or familial. Apo CI A homozygotes had 4 to 5 times the odds of developing AD, heterozygotes about twice the risk. This was not unexpected, since Apo CI is closely linked to ApoE and in linkage disequilibrium with ApoE and AD. Linkage disequilibrium is the non-random assortment, in a population, of two genes on the same chromosome (the strength of the linkage is inversely proportional to the distance between them). It was thought that the association of AD with ApoE may be more significant.

3. A Model for Alzheimer's Disease

3.1. Model Specification

We use a continuous-time multiple state model. For general comments on these models, see Macdonald (1996a) or Waters (1984). Lives with each ApoE genotype are assumed to form a homogeneous population, suffering the different risks of AD discussed in Sections 2.5 and 5.2.

An important reason for using these models is that they allow the most complete representation of the underlying process. It is then necessary to estimate a large number of transition intensities, for which adequate data do not always exist, but it is preferable to obtain a clear picture of the data needed than to sweep the issue under the carpet by working with some less adequate model in the first place. In particular:

(a) if some simpler model is eventually recommended for use, because of missing data or for computational convenience, it is important to be able to assess its soundness in practice; and
(b) if missing data become available later, for example, as the insured lives experience develops, it is a hindrance if too much has been invested in a model that cannot incorporate it.

Modern computing power is such that the computational demands of multiple state models (numerical integration of differential equations) can quite reasonably be met, for arbitrarily complex Markov models (Norberg, 1995) and for many semi-Markov models (Waters & Wilkie, 1987; Waters, 1991). The techniques can all be found in standard texts on numerical analysis, and no actuary should be prevented from choosing an adequate model by the need to use them.

Figure 1 shows a simple model of AD. Each genotype is represented by such a model; the transition intensities in each model will differ, representing the different genetic risks. x denotes the age at outset (for example, when insurance is purchased) and t the elapsed duration. The choice of states is dictated entirely by the events that have been studied in the medical and epidemiological literature. For certain purposes, it would be desirable to model other events, such as the start of a long-term insurance claim. No data about that event are available; however a major event that has been studied is institutionalisation. Although becoming institutionalised need not coincide with the start of an insurance claim, it is the best available proxy.

Macdonald (1999) considered frailty models as an alternative to Markov models, for genetics and insurance applications. They offer the advantage of a simple model of the genetic variability, if that is justified by the circumstances. They may be especially useful for modelling multifactorial disorders, or genes with very many alleles or mutations, but for a single gene with just a few alleles it seems reasonable to model each separately. Other possible models (such as Cox-type models) might be useful for modelling individual transitions but do not lend themselves to the inclusion of payments contingent upon complicated life histories.

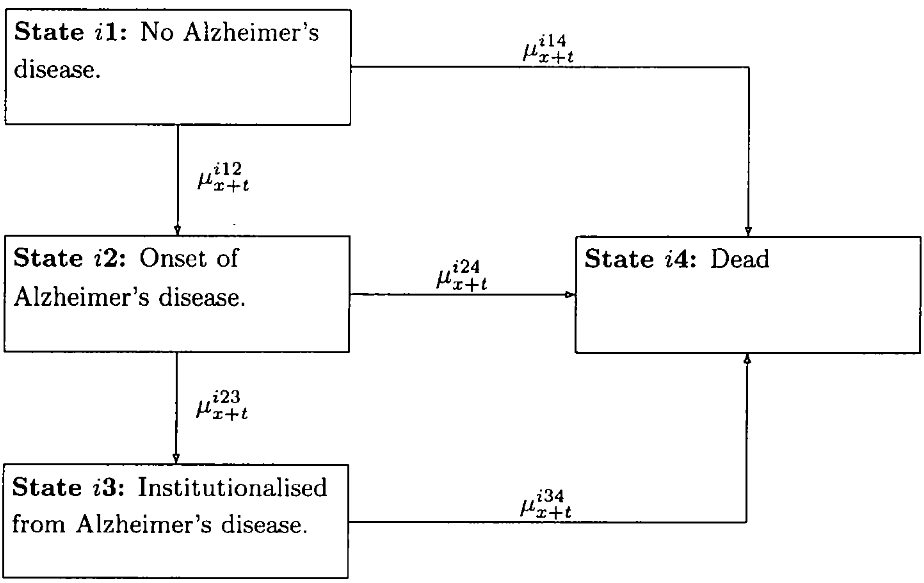


FIGURE 1: A simple model of Alzheimer's disease in the ith of M subgroups, each representing a different ApoE genotype. x is the age at outset, and t the elapsed duration.

3.2. An Expanded Long-Term Care Model

AD alone does not account for all long-term care costs. Broadly speaking, the need for care arises because of cognitive disorder (including AD) or loss of ability to perform Activities of Daily Living (ADLs) such as dressing, washing and feeding. A comprehensive model of long-term care costs can be specified in terms of these causes, with AD included as a component, and the impact of the ApoE gene on overall care costs can thereby be studied. However, incorporating AD explicitly in an expanded model will require data that describe, at the level of individual lives, the progress of AD and the loss of ADLs. Until such data are available, we can only estimate overall care costs by crude approximations: see Macdonald & Pritchard (1999).

4. Estimation of Transition Intensities Not Depending on ApoE Genotype

In this section we estimate the transition intensities for the events: onset of AD; institutionalisation; and death. All of these must be 'estimated' from results reported in the medical and epidemiological literature. Ideally, we would work with the original data, but these are almost never available. Reported results are not always ideal for the extraction of parameters for an actuarial model; often the age groups used are very wide, and not the same in different surveys; sometimes only graphs (such as Kaplan-Meier survival curves) are given.

4.1. Prospective and Case-Based Studies

A most important distinction must be drawn when estimating transition intensities from epidemiological studies (see, for example, Kahn & Sempos (1989), Clayton & Hills (1993), Lilienfeld & Hills (1993), Selvin (1996)).

(a) *Prospective* studies, based on samples of the general population, ought to yield the most reliable estimates of population risk, but are expensive and time-consuming. Moreover, they are rarely even begun until substantial evidence of an effect has been accumulated from other studies.

(b) *Case-based* studies, based on affected persons (and controls) often yield relative risks greatly in excess of the true population risks, precisely because the subjects are affected or at risk. However, early studies into any medical condition are almost inevitably of this type.

Our current knowledge of most genetic disorders is derived from case-based studies; this is certainly true of ApoE and AD (see Section 2.5). It is very likely that estimates of risk conferred by ApoE genotype will fall as more prospective studies are carried out (see the comment on Evans *et al.* (1997) at the end of Section 2.5(e)), but this will take time.

The approach we adopt is as follows:

(a) in Section 4.2, we state assumptions about the general level of mortality;

(b) in Section 4.3, we estimate the aggregate incidence of AD, denoted μ_{x+t}^{AD}, which has been investigated extensively;

(c) in Section 4.4, we estimate the intensity of institutionalisation, following the onset of AD (that is, μ_{x+t}^{i23}) and the force of mortality following the onset of AD (that is, μ_{x+t}^{i24});

(d) in Section 4.5, we estimate the force of mortality for lives institutionalised with AD (that is, μ_{x+t}^{i34});

(e) in Section 5.1, we estimate the population frequencies of the ApoE alleles; and

(f) in Section 5.2, we estimate the incidence of AD for each genotype using odds ratios from the genetic studies: this gives estimates of μ_{x+t}^{i12}.

4.2. Baseline Mortality Tables

For convenience, we choose parametric approximations to the AM80 and AF80 Ultimate mortality tables as bases for mortality assumptions; for use in the model they are adjusted in a variety of ways. Gompertz curves were fitted to μ_{x+t} at ages 65-120, using log-linear least squares (see equation (1)):

$$
\begin{aligned}
{}^{AM80}\mu_{x+t} &= 0.00009411 6e^{0.084554(x+t)} \qquad (1)\\
{}^{AF80}\mu_{x+t} &= 0.000025934e^{0.093605(x+t)}.
\end{aligned}
$$

Experiments with the AM80 and AF80 tables themselves showed that the Gompertz approximations had a negligible effect in long-term care applications; we use them because they are sometimes useful in numerical work. For insurance use, some allowance must be made for future improvements in mortality. No experience is available to help, but following discussion with some actuaries experienced in pricing long-term care insurance, we have chosen 65% of these baseline tables as the aggregate mortality assumptions.

4.3. The Onset of Alzheimer's Disease in the Population

AD has been the subject of some large-scale epidemiological studies, many of them pre-dating the discovery of the ApoE gene. Some of these report incidence rates, or 'occurrence/exposure' rates, which are exactly the estimates we need for transition intensities.

There is general agreement, in this literature, on the shape of the intensity μ_{x+t}^{AD} in the age range 60-85 years; it is very low at ages 60-64 (about 0 to 0.002) and increases rapidly with age, approximately doubling every 5 years. Sayetta (1986) and Hebert *et al.* (1995) found that a Gompertz curve gave the best fit, despite trying a number of more complex models.

A number of studies report the incidence of AD (that is, the intensity μ_{x+t}^{AD}) by age but not by genotype, including Copeland *et al.* (1992), Hagnell *et al.* (1992), Kokmen *et al.* (1993), Letenneur *et al.* (1994), Nilsson (1984), Ott *et al.* (1998), Rocca *et al.* (1998) and Rorsman *et al.* (1986). Of particular interest, however, is the recent meta-analysis of the incidence of AD by Jorm & Jolley (1998):

(a) it draws on 23 studies world-wide, including 13 European studies;
(b) the analysis is carried out separately for Europe, the U.S.A. and East Asia;
(c) the incidence of AD is estimated by severity, categorised as Mild+ and Moderate+ AD, where Mild+ includes all cases classified as mild or worse; and
(d) point estimates of μ_{x+t}^{AD} were obtained for 5-year age groups from 65 to 95, and no *a priori* shape of μ_{x+t}^{AD} was assumed.

We estimated μ_{x+t}^{AD} from Jorm & Jolley (1998) using the figures from the European studies and for Mild+ AD. The estimates, 95% confidence limits and the log-linear least squares Gompertz fit:

$$\mu_{x+t}^{AD} = 1.31275 \times 10^{-7} e^{0.145961(x+t)} \tag{2}$$

are shown in Figure 2. It is clear that a Gompertz curve is a very good fit.

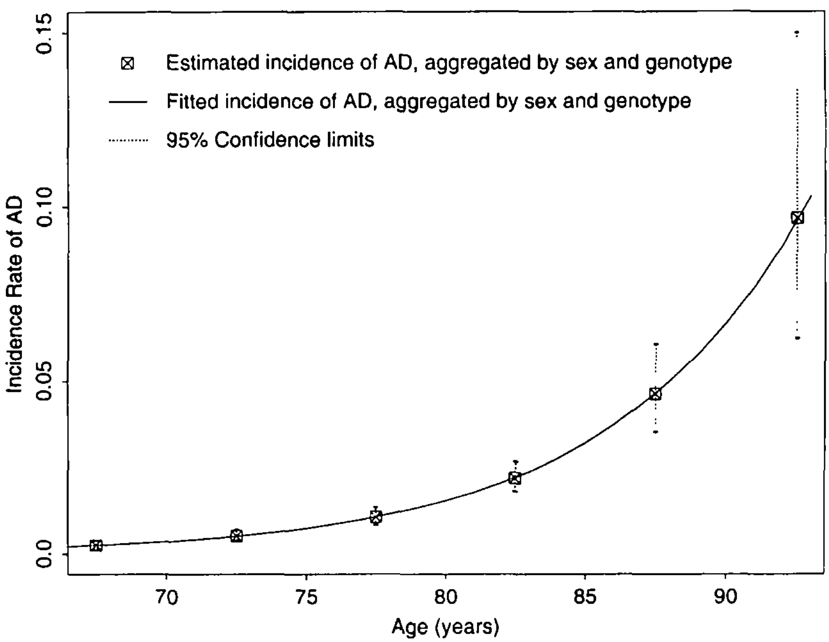


FIGURE 2: Aggregate incidence of Alzheimer's disease: point estimates and 95% confidence intervals. Source: Jorm & Jolley (1998).

Data on the incidence of AD among the very elderly (> 90 years) are sparse, so estimates at these ages have wide confidence intervals and the trend is uncertain. The meta-analysis by Gao *et al.* (1998) found that the rate of increase in μ^{AD}_{x+t} slowed down with age, but other studies found no evidence of this (Hebert *et al.*, 1995; Jorm & Jolley, 1998; Letenneur *et al.*, 1994). We have simply extrapolated the Gompertz formula above to all ages; the effect of this assumption will depend on the particular application, or type of insurance, and this should be investigated when the model is used (see, for example, Macdonald & Pritchard (1999)).

Many studies have found men and women to be at the same risk of AD (Kokmen *et al.*, 1993; Nilsson, 1984; Ott *et al.*, 1998; Rocca *et al.*, 1998) and, when differences have been reported (Gao *et al.*, 1998; Jorm & Jolley, 1998; Letenneur *et al.*, 1994), women were found to be at greater risk only at very old ages. Some experiments (not described here) in applying the model to AD-related long-term care insurance costs using different rates of AD for men and women showed that it made little difference, and here we have used the aggregate rate (equation (2)).

4.4. Time from Onset of Alzheimer's Disease to Institutionalisation or Death

The available data do not allow us to analyse μ^{i23}_{x+t}, μ^{i24}_{x+t} or μ^{i34}_{x+t} by genotype.

Table 1 summarises the literature on time to the first of institutionalisation or death ('first event') for AD patients. Some studies give times from entry to the study rather than from onset, which is usually not observed directly. A striking feature is that few lives die before becoming institutionalised. This may seem surprising as AD patients have generally been reported to suffer higher mortality than healthy lives (see Section 4.5). However, AD's debilitating effects are not sudden, and we may expect patients to be in receipt of informal care between onset and institutionalisation, which might lead μ^{i24}_{x+t} to be relatively light.

We used the data from the study by Jost & Grossberg (1995). Although it is not the largest study, it does have advantages:

(a) it is a brain bank study, so all AD cases were confirmed by autopsy (the only reliable method of diagnosis);
(b) there were no censored cases; and
(c) the time from onset to institutionalisation is estimated.

TABLE 1

MEAN AND MEDIAN TIMES TO INSTITUTIONALISATION (INST'N) OR FIRST EVENT FOR AD PATIENTS

Reference	*Age at*		*Time (years) to*		*% for which 1st event is death*
	Onset	*Entry*	*Inst'n*	*1st Event*	
Berg *et al.* (1988)		71.4 (1)			7.1%
Heyman *et al.* (1997)		72.0 (2)		3.1 (2)	13.1%
Jost *et al.* (1995)	75.1 (3)		4.3 (1)		15.0%
Severson *et al.* (1994)		79.4 (1)	2.5 (2) (4)		10.0%

(1) Mean.
(2) Median.
(3) Mean age at onset of AD, if institutionalised, estimated as (mean age at institutionalisation – mean time to institutionalisation).
(4) Median time from onset estimated as 5.6 years.

Since we cannot distinguish genotypes here, we will just write μ^{23}_{x+t} and μ^{24}_{x+t} instead of μ^{i23}_{x+t} and μ^{i24}_{x+t}, respectively. Guided by these data we derive moment estimates of μ^{23}_{x+t} (the force of institutionalisation) and μ^{24}_{x+t} (the force of mortality of an AD patient prior to institutionalisation). We define below the usual indicator functions (I_j) and sample path functions (N_{jk}) in respect of a single life (see Macdonald (1996b)):

$$I_j(t) = \begin{cases} 1 & \text{if life is in state } j \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

$$dN_{jk}(t) = \begin{cases} 1 & \text{if life transfers from state } j \text{ to state } k \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

$$N_{jk}(T) = \int_0^T dN_{jk}(t) = \text{No. of transfers from state } j \text{ to state } k$$

Also let P^{jj}_{xy} be the probability that a life in state j at age x is in state j at age y. Then equation (3) below is the mean age at onset of AD, given that the life was eventually institutionalised with AD (as in Jost & Grossberg (1995)):

$$\mathrm{E}\left[x + \int_x^\omega I_1(t)dt \,\middle|\, N_{23}(\omega - x) = 1 \text{ and } I_1(x) = 1\right] =$$

$$x + \frac{\int_x^\omega (t-x)\mu^{12}_t P^{11}_{xt}\left\{\int_t^\omega \mu^{23}_s P^{22}_{ts} ds\right\}dt}{\int_x^\omega \mu^{12}_t P^{11}_{xt}\left\{\int_t^\omega \mu^{23}_s P^{22}_{ts} ds\right\}dt}\;; \qquad (3)$$

equation (4) is the mean time from onset of AD to institutionalisation:

$$\mathrm{E}\left[\int_x^{\omega} I_2(t)dt \,\middle|\, N_{23}(\omega - x) = 1 \text{ and } I_1(x) = 1\right] = \frac{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} \left\{\int_t^{\omega} (s-t)\mu_s^{23} P_{ts}^{22} ds\right\} dt}{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} \left\{\int_t^{\omega} \mu_s^{23} P_{ts}^{22} ds\right\} dt};$$

and equation (5) is the probability that an AD patient dies before becoming institutionalised. The upper age bound, denoted ω, is taken to be 120 years:

$$\mathrm{P}\left[N_{24}(\omega - x) = 1 \,\middle|\, N_{12}(\omega - x) = 1 \text{ and } I_1(x) = 1\right] = \frac{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} \left\{\int_t^{\omega} \mu_s^{24} P_{ts}^{22} ds\right\} dt}{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} dt} \qquad (5)$$

Setting equations (3), (4) and (5) equal to their estimated values from Table 1, we obtain 3 equations, which can be solved for at most 3 unknown parameters. The parametric forms we chose were as follows:

(a) $\mu_{x+t}^{12} = A + \mu_{x+t}^{AD}$, where μ_{x+t}^{AD} is given by equation (2). This Makeham term adjusts the incidence of AD to a level that gives the same mean age at onset (for AD patients who become institutionalised).

(b) $\mu_{x+t}^{23} = D$. We felt that the data did not support anything more elaborate than a constant intensity.

(c) $\mu_{x+t}^{24} = P\mu_{x+t}^{14}$. That is, the mortality of an AD patient before becoming institutionalised is a proportion of baseline mortality.

(d) μ_{x+t}^{14}, baseline mortality, was taken as AM80 mortality, using the Gompertz approximation given by equation (1). Although it is appropriate to allow for future improvements in mortality in applications, it is not appropriate to do so in estimation based on past data. The values of D and P do not depend strongly on the baseline mortality.

Solving these equations numerically yields the solutions:

$$A = 0.02025038 \quad D = 0.18895779 \quad P = 0.33502488.$$

The transition intensities μ_{x+t}^{23} and μ_{x+t}^{24} are summarised in Table 3.

The Makeham term, A, is a nuisance parameter used to adjust the incidence of AD so that the mean age at onset in the model is the same as that in the data. Its only purpose here is to improve the estimation of the other terms, as the survival of a cohort of AD patients is strongly related to their mean age at onset. It does not furnish an estimate of the incidence of AD in the whole population, which was described in Section 4.3.

4.5. Mortality of Lives with Alzheimer's Disease

AD patients have been found to suffer higher mortality than the general population (Barclay *et al.*, 1985(b); Bonaiuto *et al.*, 1995; Bracco *et al.*, 1994;

Burns *et al.*, 1991; van Dijk *et al.*, 1991; Evans *et al.*, 1991; Heyman *et al.*, 1996; Mölsä *et al.*, 1986; Treves *et al.*, 1986). However, there is little agreement on the magnitude of the increase, or its dependence on age at onset, duration since onset, sex, race, level of education, marital status, level of cognitive impairment, familial/non-familial AD and level of behavioural impairment. The main factors we need to consider are:

(a) *The magnitude of the increase in mortality for AD lives.* The mortality of lives with AD has been investigated using different methodologies. For example, Evans *et al.* (1991), estimated the relative risk of death for AD patients as 1.44 (95% confidence interval 1.05-1.96) times that of the unaffected. Others have suggested that AD has only a small impact on mortality: Barclay *et al.* (1985a) claimed that well-tended individuals may have life expectancy close to normal, and Sayetta *et al.* (1986) found that survival did not depend on disease acquisition.

(b) *The effect of age at onset on relative mortality.* The mortality of patients with AD increases with age (Bonaiuto *et al.* 1995; Burns *et al.*, 1991). Most studies into survival times have found no relation between age at entry into the study and relative survival (Barclay *et al.*, 1985b; Bracco *et al.*, 1994; Heyman *et al.*, 1996; Stern *et al.*, 1995; Mölsä *et al.*, 1986), except that Barclay *et al.* (1985b) found that younger lives had shorter relative survival times. Diesfeldt *et al.* (1986), investigating survival from onset of AD, found that AD patients with onset before age 76 had reduced survival times, but not those with later onset. Comparing the two methods of investigation, Walsh *et al.* (1990) found that older age at onset affected survival adversely, whereas older age at entry into the study did not; a possible explanation was that older patients have symptoms for a shorter time before presentation. Although no definitive relationship between age at onset and relative survival emerges, it is clear that:

   (1) survival with AD depends on age; and
   (2) if age at onset affects relative mortality, the relationship is only weak, but possibly stronger at younger ages.

   In terms of the model in Figure 1, this suggests that mortality in state *i*3 (institutionalised from AD) could be modelled by the addition of a Makeham term to the normal force of mortality; the latter is age dependent, and the Makeham term will be less significant at older ages.

(c) *The effect of the duration of AD on relative survival.* Perhaps surprisingly, the duration of AD has not been found to be associated with increased mortality (Barclay *et al.*, 1985a; Bracco *et al.*, 1994; Burns *et al.*, 1991; Diesfeldt *et al.*, 1986; Heyman *et al.*, 1996; Sayetta *et al.*, 1986; Walsh *et al.*, 1990). That is, AD patients with long duration of symptoms do not suffer higher mortality than patients, of the same age, with short duration of symptoms. In terms of the model, this means that the mortality of lives in states 2 and 3 (onset of AD and institutionalised from AD) does not depend on the time spent in these states. This is especially convenient, as it allows us to work in a Markov framework.

(d) *The effect of gender on relative survival with AD.* Many researchers have found that the differences in survival between men and women with AD can be explained by the usual mortality differential between men and women (Beard *et al.*, 1994; Bonaiuto *et al.*, 1995; Bracco *et al.*, 1994; Burns *et al.*, 1991; Heyman *et al.*, 1996; Walsh *et al.*, 1990), though Barclay *et al.* (1985a), did find greater differences. In terms of modelling, allowing for the normal differences in mortality between genders should be sufficient.

Table 2 summarises the literature on survival with AD. Since we cannot distinguish genotypes here, we will just write μ^{34}_{x+t} instead of μ^{i34}_{x+t}. As in the previous section, we can write down the mean age at onset (see equation (6)) and the mean survival time (see equation (7)) in the model of Figure 1:

$$\mathrm{E}\left[x + \int_x^{\omega} I_1(t)dt \,\middle|\, N_{12}(\omega - x) = 1 \text{ and } I_1(x) = 1\right] = x + \frac{\int_x^{\omega} (t-x)\mu_t^{12} P_{xt}^{11} dt}{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} dt} \tag{6}$$

$$\mathrm{E}\left[\int_x^{\omega} I_2(t) + I_3(t)dt \,\middle|\, N_{12}(\omega - x) = 1 \text{ and } I_1(x) = 1\right] = \frac{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} \{\int_t^{\omega} (s-t)(\mu_s^{23} + \mu_s^{24}) P_{ts}^{22} ds + \int_t^{\omega} \mu_s^{23} P_{ts}^{23} \int_s^{\omega} (r-s)\mu_r^{34} P_{sr}^{33} dr\, ds\} dt}{\int_x^{\omega} \mu_t^{12} P_{xt}^{11} dt} \tag{7}$$

TABLE 2

SUMMARY STATISTICS ON SURVIVAL TIMES OF AD PATIENTS

Reference	*Mean (Median) Age at Onset*	*Mean (Median) Survival Time*	*Addition to* μ^{34}_{x+t}
Barclay *et al.* (1985a)	(73.3) yrs	(8.1) yrs	0.15829
Bracco *et al.* (1994)	(72.4) yrs	7.3 yrs	0.25259
Diesfeldt *et al.* (1986)	75.6 yrs	7.2 yrs	0.21056
Heyman *et al.* (1996)	(69.2) yrs	(9.7) yrs	0.10993
Jost *et al.* (1995)	75.1 yrs	8.11 yrs	0.13345
Kokmen *et al.* (1988)	80.4 yrs	6.2 yrs	0.26420
Treves *et al.* (1986)	73.9 yrs	(9.3) yrs	0.08135
Average			0.17291

Setting equations (6) and (7) equal to their estimated values in Table 2, and noting that we have estimates of μ^{23}_{x+t} and μ^{24}_{x+t} from the previous section, we obtain 2 equations, which can be solved for at most 2 unknown parameters. The parametric forms we used are as follows:

(a) $\mu^{12}_{x+t} = A + \mu^{AD}_{x+t}$, where μ^{AD}_{x+t} is given by equation (2). This is just the addition of a Makeham term to the force of incidence of AD, shifting the latter to a level that gives the estimated age at onset.

(b) $\mu^{34}_{x+t} = K + {}^{AM80}\mu_{x+t}$. This is a Makeham term as discussed in (d) above.

The estimated values of K for each of the references cited are given in the last column of Table 2. They range from about 0.08 to 0.27, with an average of 0.173. The Makeham term A is, again, only included to improve the estimation of the other terms (see the end of the previous section).

4.6. Summary of the Transition Intensities for the AD Model

For clarity, we summarise the transition intensities estimated here. They all have the form:

$$\mu^{jk}_{x+t} = A + DBe^{C(x+t)} \tag{8}$$

and the calculated values are given in Table 3. Three values are given for μ^{34}_{x+t}, an upper bound, mean value and lower bound to enable a check of how sensitive the results are, in any particular investigation, to this term.

TABLE 3

SUMMARY OF TRANSITION INTENSITIES FOR THE AD MODEL WITH BASELINE MORTALITY 100% (65%) OF AM80 AND AF80

Transition Intensity	*Parameter Values*					
	A	*D*	*B* ($\times 10^{-5}$)		*C* ($\times 10^{-2}$)	
			Male	*Female*	*Male*	*Female*
μ^{24}_{x+t}	0	0.33502 (0.21776)	9.4116	2.5934	8.4554	9.3605
μ^{23}_{x+t}	0.18896	0.00				
μ^{34}_{x+t} (Lower bound)	0.08	1.00 (0.65)	9.4116	2.5934	8.4554	9.3605
μ^{34}_{x+t} (Mean)	0.17291	1.00 (0.65)	9.4116	2.5934	8.4554	9.3605
μ^{34}_{x+t} (Upper bound)	0.27	1.00 (0.65)	9.4116	2.5934	8.4554	9.3605

5. Estimation of Transition Intensities Depending on ApoE Genotype

5.1. Population Frequencies of the ApoE Genotypes

Table 4 shows the population frequencies of the ApoE genotypes estimated in several studies. Others are Corder *et al.* (1994, 1995), Gomez-Isla *et al.* (1996), Lehtovirta *et al.* (1996), Liddell *et al.* (1994), Lopez *et al.* (1998), Nalbantoglu *et al.* (1994), Poirier *et al.* (1993), Roses (1995) and Tsai *et al.* (1994).

Some features are clear: the $\varepsilon 3/\varepsilon 4$ genotype is not uncommon (about 21%) while the $\varepsilon 2/\varepsilon 4$ and $\varepsilon 4/\varepsilon 4$ genotypes are quite uncommon (about 3% and 1% respectively). We might expect to find lower proportions of 'dangerous' genotypes at older ages, because these lives suffer a higher rate of AD onset, but the two age-related studies (Bickeböller *et al.* (1997) and Corder *et al.* (1995)) gave conflicting results. However, there is reasonable agreement on the gene frequencies at around ages 60-70, which is what we need for our modelling.

Farrer *et al.* (1997) is a meta-analysis, combining the results of 40 other studies, including 6,264 Caucasian subjects. As it is the largest study, and differentiates by ethnic group, gender and ascertainment methods, and as the ApoE $\varepsilon 4$ allele was found with the same frequency in respect of AD diagnosed at autopsy and clinically diagnosed probable AD, we use its estimated gene frequencies, namely: $\varepsilon 2/\varepsilon 2$ 0.008; $\varepsilon 2/\varepsilon 3$ 0.127; $\varepsilon 2/\varepsilon 4$ 0.026; $\varepsilon 3/\varepsilon 3$ 0.609; $\varepsilon 3/\varepsilon 4$ 0.213; $\varepsilon 4/\varepsilon 4$ 0.018. The sharp-eyed will notice that these sum to 1.001, because of roundings used in Farrer *et al.* (1997), but we have left this small discrepancy unadjusted.

5.2. Genetic Risk of Alzheimer's Disease

When we have a heterogeneous population, it is often convenient to think of a given intensity in each sub-population as a multiple (not necessarily constant) of a 'baseline' intensity, either in one of the sub-populations or in an aggregated 'average' population. Similarly, if p_1 and p_2 are the probabilities of an event in populations 1 and 2 respectively, the relative risk in population 2 (with respect to population 1) is p_2/p_1. A related quantity is the odds ratio: the odds in populations 1 and 2, respectively, are $p_1/(1-p_1)$ and $p_2/(1-p_2)$, and the odds ratio is:

$$\frac{p_2(1-p_1)}{p_1(1-p_2)} . \tag{9}$$

When intensities are small, the odds ratio is a good approximation to the relative risk. In many studies, the published results are either relative risks or odds ratios.

TABLE 4

ESTIMATED POPULATION FREQUENCY OF APOE GENOTYPES

Reference	*Country / Ethnicity*	*No. of Lives*	*Sex*	*Age Group*	*Allele Frequency*			*Genotype Frequency*					
					$\varepsilon 2$	$\varepsilon 3$	$\varepsilon 4$	$\varepsilon 2/\varepsilon 2$	$\varepsilon 2/\varepsilon 3$	$\varepsilon 2/\varepsilon 4$	$\varepsilon 3/\varepsilon 3$	$\varepsilon 3/\varepsilon 4$	$\varepsilon 4/\varepsilon 4$
Farrer *et al.* (1997)	Caucasian	6,262	M & F	All	0.084	0.779	0.137	0.01	0.13	0.03	0.61	0.21	0.02
	Afr-Amer	240	M & F	All	0.083	0.727	0.190	0.01	0.13	0.02	0.50	0.32	0.02
	Hispanic	267	M & F	All	0.067	0.823	0.110	0.00	0.12	0.01	0.67	0.18	0.02
	Japanese	1,977	M & F	All	0.042	0.869	0.089	0.00	0.07	0.01	0.76	0.16	0.01
Bickeböller *et al.* (1997)	France	1,030	M & F	All	0.085	0.770	0.145	0.01	0.12	0.03	0.59	0.24	0.01
		316	M	All	0.070	0.815	0.105	0.01	0.10	0.02	0.67	0.19	0.00
		40	M	< 60	0.050	0.800	0.150	0.00	0.08	0.02	0.62	0.28	0.00
		93	M	60-69	0.070	0.845	0.085	0.01	0.12	0.00	0.70	0.17	0.00
		80	M	70-79	0.060	0.840	0.090	0.00	0.10	0.02	0.71	0.16	0.00
		103	M	≥ 80	0.081	0.795	0.125	0.01	0.11	0.03	0.64	0.20	0.01
		714	F	All	0.090	0.750	0.170	0.01	0.13	0.03	0.55	0.27	0.02
		47	F	< 60	0.075	0.735	0.190	0.00	0.15	0.00	0.51	0.30	0.04
		75	F	60-69	0.075	0.730	0.195	0.01	0.11	0.02	0.52	0.31	0.03
		143	F	70-79	0.065	0.765	0.160	0.00	0.12	0.01	0.56	0.29	0.01
		449	F	≥ 80	0.095	0.750	0.165	0.01	0.13	0.04	0.56	0.25	0.02
van Duijn *et al.* (1995)	Netherlands	532	M & F	< 65	0.103	0.731	0.165	0.01	0.17	0.02	0.51	0.27	0.02
		228	M	< 65	0.105	0.705	0.190	0.01	0.16	0.03	0.48	0.29	0.03
		304	F	< 65	0.100	0.745	0.155	0.01	0.17	0.01	0.53	0.26	0.02
Evans *et al.* (1997)	E. Boston	490	M & F	≥ 65	0.062	0.854	0.084	0.01	0.09	0.01	0.74	0.14	0.01
Jarvik *et al.* (1996)	not given	310	M & F	48-98	0.098	0.750	0.132	0.01	0.12	0.05	0.59	0.20	0.01
		117	M	48-98	0.068	0.791	0.124	0.02	0.09	0.02	0.64	0.21	0.01
		193	F	48-98	0.117	0.725	0.137	0.01	0.15	0.07	0.55	0.20	0.01
Lambert *et al.* (1998)	not given	308	M & F	not given	0.081	0.805	0.114	0.00	0.13	0.02	0.65	0.18	0.01
Levy-Lahad *et al.* (1996)	not given	304	M & F	not given	0.100	0.765	0.135	0.01	0.13	0.05	0.60	0.21	0.01
Lucotte *et al.* (1997)	France	248	M & F	≥ 65	0.069	0.804	0.127	0.02	0.09	0.01	0.65	0.22	0.01

Few studies report prospectively the incidence of AD by genotype. Two that do are Evans *et al.*, (1997) and Slooter *et al.*, (1998). Both have quite small study populations, and neither provides age specific estimates of AD risk, so they are not appropriate for our purposes.

Table 5 gives the Odds Ratios (ORs) of AD and 95% confidence intervals from a number of genetic studies. The 'reference' populations (also shown in the table) were either the $\varepsilon 3/\varepsilon 3$ genotype or the three non-$\varepsilon 4$ genotypes combined. The estimated ORs vary considerably across studies. For example, estimates of the OR for the $\varepsilon 3/\varepsilon 4$ genotype (relative to the $\varepsilon 3/\varepsilon 3$ genotype) range from 1.8% to 3.7%, and for the $\varepsilon 4/\varepsilon 4$ genotype, from 6.2% to 30.7%. Some of the variation may be explained by the differences between the studies themselves. In particular, differences may arise from: the method of ascertainment of patients, the countries of study, the method of diagnosis of AD, the age structure of the samples, the reference/risk genotypes, and whether they are cross-sectional or prospective studies. We make the following observations:

(a) The study by Lopez *et al.* (1998) suggests that the risk of AD associated with the ApoE $\varepsilon 4$ allele may be different in different countries.

(b) In support of the above, Mayeux *et al.* (1998) found that the association between ApoE and AD may depend on ethnic group and, in particular, may not be present in black populations.

(c) Despite the differences between studies: the presence of one or two $\varepsilon 4$ alleles is consistently reported to be significantly associated with AD; and homozygotes are generally reported to have higher risk of onset of AD than heterozygotes.

(d) The weakest associations between ApoE and AD were reported in the two prospective studies, those by Evans *et al.* (1997) and Slooter *et al.* (1998). This is as expected for the reasons given in Section 4.1.

For our purposes, the genetic risk of AD at different ages is important. Few studies have considered this; the odds ratios from two that have are given in Table 6. Bickeböller *et al.* (1997) is based on hospital admissions, and Corder *et al.* (1994) on autopsy cases; both use $\varepsilon 3/\varepsilon 3$ as the reference population. Although some ORs are missing, because of small sample sizes, the trends are fairly clear:

(a) The odds of AD among the higher risk genotypes ($\varepsilon 3/\varepsilon 4$ and $\varepsilon 4/\varepsilon 4$) fall with age. This may be expected as higher risk genotypes will succumb to AD more rapidly, reducing the proportion of such genotypes within the population.

(b) Conversely, the protection conferred by the lower risk genotype ($\varepsilon 2/\varepsilon 3$) seems to weaken with age, possibly as this genotype becomes more common in the remaining population.

TABLE 5

AGGREGATED ODDS RATIOS OF AD FOR THE APOE $\varepsilon 4$ ALLELE

Reference	*Ascertainment Scheme (1)*	*Reference Genotype (2)*	*Genotype (2)*	*Odds Ratio*	
				Mean	*95% CI*
Evans *et al.* (1997)	P	$\varepsilon 3/\varepsilon 3$	$\varepsilon 3/\varepsilon 4$ & $\varepsilon 4/\varepsilon 4$	2.27	1.1-4.9
Frisoni *et al.* (1995)	C	-/-	$\varepsilon 4$/-	6.6	2.2.-19.5
			$\varepsilon 4/\varepsilon 4$	17.9	4.5-70.5
Jarvik *et al.* (1996)	C	$\varepsilon 3/\varepsilon 3$	$\varepsilon 2/\varepsilon 3$	0.4	0.2-0.96
			$\varepsilon 2/\varepsilon 4$	1.4	0.6-3
			$\varepsilon 3/\varepsilon 4$	3.1	2-4.7
			$\varepsilon 4/\varepsilon 4$	30.7	7-131
Kuusisto *et al.* (1994)	P	-/-	$\varepsilon 4$/-	2.7	1.4-5.2
			$\varepsilon 4/\varepsilon 4$	9.1	3.5-23.4
Lambert *et al.* (1998)	C	-/-	$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	4.66	3.14-6.93
Lehtovirta *et al.* (1995)	C	-/-	$\varepsilon 4$/-	5.1	2.4-11.1
			$\varepsilon 4/\varepsilon 4$	21.4	2.8-166.3
Liddell *et al.* (1994)	C	-/-	$\varepsilon 4$/-	2.2	1.1-4.7
			$\varepsilon 4/\varepsilon 4$	10.7	2.3-48.8
Lopez *et al.* (1998)	C	-/-	$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	2.34 (3)	1.03-5.55
			$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	3.64 (4)	1.78-7.69
Mayeux *et al.* (1993)	P	-/-	$\varepsilon 4$/-	4.2 (5)	1.8-9.5
			$\varepsilon 4/\varepsilon 4$	17.9 (5)	4.6-69.8
			$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	15.3 (6)	3.0-78.1
			$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	0.7 (7)	0.1-6.4
			$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	4.5 (8)	0.7-27.7
Myers *et al.* (1996)	P	$\varepsilon 3/\varepsilon 3$	$\varepsilon 3/\varepsilon 4$	3.7	1.9-7.5
			$\varepsilon 4/\varepsilon 4$	30.1	10.7-84.4
Nalbantoglu *et al.* (1994)	A	-/-	$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	15.5	6.2-38.5
Slooter *et al.* (1998)	P	$\varepsilon 3/\varepsilon 3$	$\varepsilon 2/\varepsilon 3$	0.4	0.1-1.0
			$\varepsilon 2/\varepsilon 4$	1.3	0.2-8.5
			$\varepsilon 3/\varepsilon 4$	1.8	1.0-3.1
			$\varepsilon 4/\varepsilon 4$	6.2	1.4-28.2
Tsai *et al.* (1994)	C	-/-	$\varepsilon 4$/- & $\varepsilon 4/\varepsilon 4$	4.6	1.9-12.3
			$\varepsilon 4$/-	3.6	1.5-9.8

(1) Ascertainment Scheme: C indicates clinic/hospital; P, population/community; and A, autopsy/brainbank.
(2) Dash (-) represents $\varepsilon 2$ or $\varepsilon 3$ alleles.
(3) Study population – Gerona, Spain.
(4) Study population – Pittsburgh, USA.
(5) Mixture of White, Black and Hispanic ethnic groups.
(6) White ethnic group only.
(7) Black ethnic group only.
(8) Hispanic ethnic group only.

TABLE 6

ODDS RATIOS OF AD BY GENOTYPE AND AGE

Bickeböller et al. (1997)				*Corder et al. (1994)*			
Age Group	*Genotype*	*Odds Ratio Mean*	*Odds Ratio 95% CI*	*Age Group*	*Genotype*	*Odds Ratio Mean*	*Odds Ratio 95% CI*
60-69	$\varepsilon2/\varepsilon3$	0.3	0.0-2.3	60-66	$\varepsilon3/\varepsilon3$	0.1	–
	$\varepsilon2/\varepsilon4$	–	–		$\varepsilon2/\varepsilon4$	1.2	–
	$\varepsilon3/\varepsilon4$	3.1	1.4-6.9		$\varepsilon3/\varepsilon4$	11.1	–
	$\varepsilon4/\varepsilon4$	29.1	3.6-239.5		$\varepsilon4/\varepsilon4$	123.8	–
70-79	$\varepsilon2/\varepsilon3$	0.4	0.1-2.3	67-74	$\varepsilon2/\varepsilon3$	0.3	–
	$\varepsilon2/\varepsilon4$	–	–		$\varepsilon2/\varepsilon4$	1.1	–
	$\varepsilon3/\varepsilon4$	3.2	1.5-6.6		$\varepsilon3/\varepsilon4$	4.6	–
	$\varepsilon4/\varepsilon4$	–	–		$\varepsilon4/\varepsilon4$	20.8	–
80+	$\varepsilon2/\varepsilon3$	0.3	0.0-2.6	75-92	$\varepsilon2/\varepsilon3$	0.5	–
	$\varepsilon2/\varepsilon4$	–	–		$\varepsilon2/\varepsilon4$	1.6	–
	$\varepsilon3/\varepsilon4$	1.3	0.5-3.4		$\varepsilon3/\varepsilon4$	3.2	–
	$\varepsilon4/\varepsilon4$	–	–		$\varepsilon4/\varepsilon4$	10.0	–
60+	$\varepsilon2/\varepsilon3$	0.4	0.1-0.9	60+	$\varepsilon2/\varepsilon3$	0.3	–
	$\varepsilon2/\varepsilon4$	1.6	0.5-5.5		$\varepsilon2/\varepsilon4$	1.1	–
	$\varepsilon3/\varepsilon4$	2.2	1.5-3.5		$\varepsilon3/\varepsilon4$	4.4	–
	$\varepsilon4/\varepsilon4$	11.2	4.0-31.6		$\varepsilon4/\varepsilon4$	19.3	–

These trends are supported by the meta-analysis by Farrer *et al.* (1997), and as it is from this study that we take our estimates of the ApoE genotype risks, we cite some relevant details:

(a) The aggregate relative odds from Farrer *et al.* (1997) (relative to the $\varepsilon3/\varepsilon3$ genotype) are shown in Figure 3.

(b) The genotype risks of AD were not significantly different in respect of Caucasian males and females, except in the case of the $\varepsilon3/\varepsilon4$ genotype, for which women had a significantly higher risk of AD. The relative odds of AD by ApoE for Caucasian men and women are shown in Figures 4 and 5. (The authors kindly provided us with the numerical values of the odds ratios; confidence intervals were not available.)

(c) The genotypes $\varepsilon2/\varepsilon2$ and $\varepsilon2/\varepsilon3$ were combined as there were very few $\varepsilon2/\varepsilon2$ genotypes, and the risks associated with the two genotypes appeared to be similar.

(d) Note that the risks associated with the ApoE $\varepsilon4$ allele were considerably higher than those found in the two population-based studies by Evans *et al.*, (1997) and Slooter *et al.*, (1998).

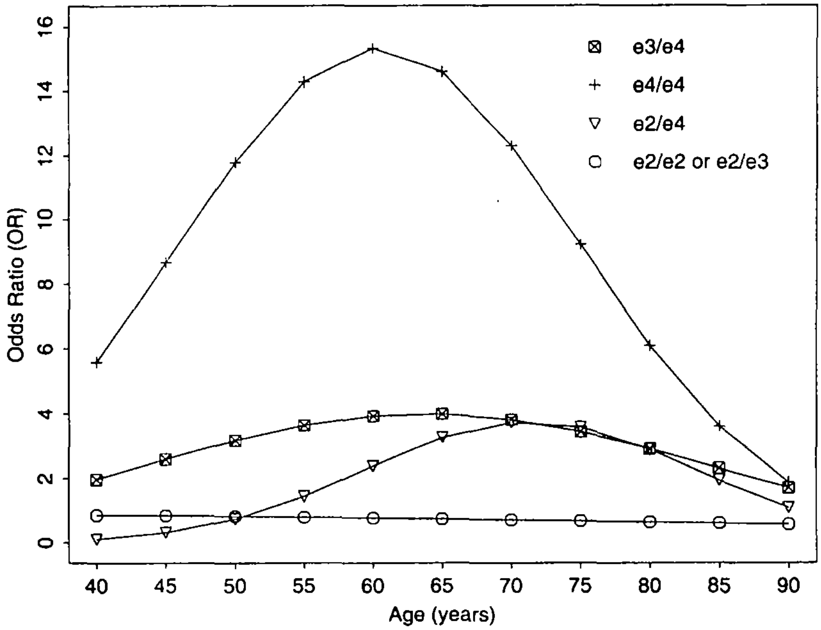


FIGURE 3: Odds ratios (ORs) of AD relative to $\varepsilon3/\varepsilon3$ genotype for males and females combined. Source: Farrer *et al.* (1997).

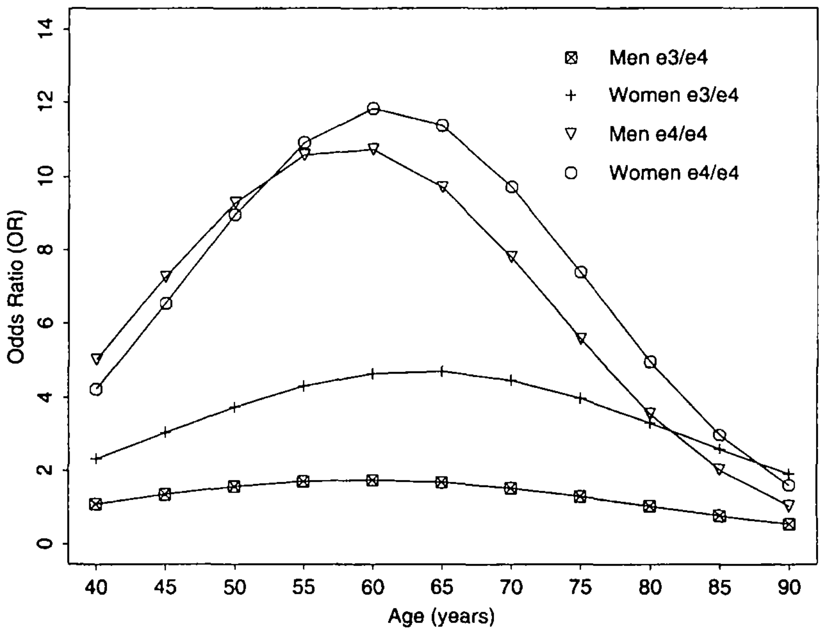


FIGURE 4: Odds ratios (ORs) of AD relative to $\varepsilon3/\varepsilon3$ genotype for $\varepsilon3/\varepsilon4$ and $\varepsilon4/\varepsilon4$ genotypes. Source: Farrer *et al.* (1997).

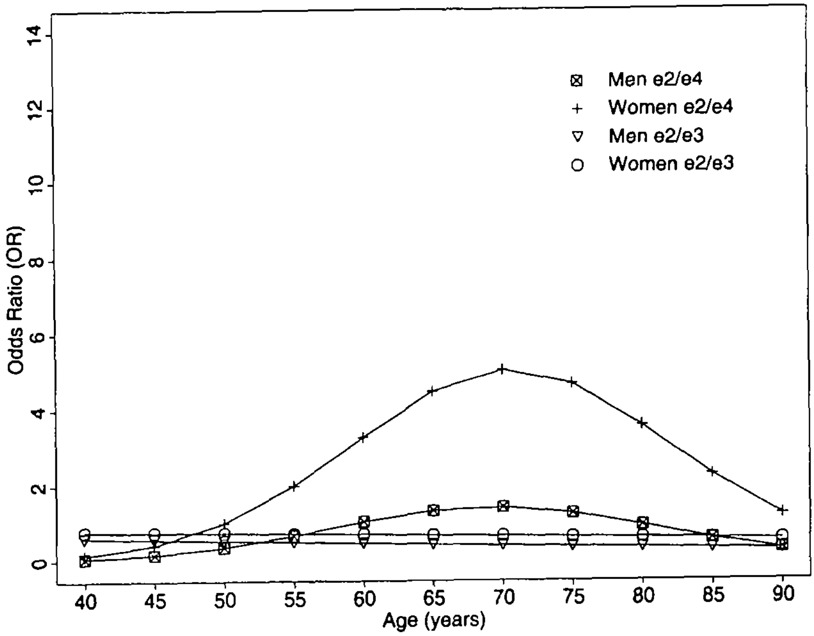


FIGURE 5: Odds ratios (ORs) of AD relative to $\varepsilon3/\varepsilon3$ genotype for $\varepsilon2/\varepsilon2$ or $\varepsilon2/\varepsilon3$ and $\varepsilon2/\varepsilon4$ genotypes. Source: Farrer *et al.* (1997).

For use in our model, these odds ratios have to be converted into relative risks. More precisely, we have to find a plausible set of age- and genotype-dependent transition intensities that are consistent with the odds ratios and together are consistent with the aggregate incidence of AD. There is no unique solution to this problem. The method we used was to model the incidence of AD for the ith genotype as:

$$\mu_{x+t}^{i12} = r_1 f_{x+t}^{i} \mu_{x+t}^{AD} \tag{10}$$

where:

(a) μ_{x+t}^{AD} is the aggregate incidence rate of AD (from Section 4.3);

(b) f_{x+t}^{i} is a parametric function representing the risk relative to the aggregate incidence rate, where we take $f_{x+t}^{i} = 1$ in the case of the $\varepsilon3/\varepsilon3$ genotype; and

(c) r_1 is a constant chosen so that the aggregate incidence of AD based on the modelled intensities is consistent with the aggregate incidence μ_{x+t}^{AD}.

We did this for males and females separately and combined. We confined our attention to ages 60 and over, in order to get a better fit in the age range of interest in applications. The form of the ORs, either rising to a peak and then falling, or gently declining, suggested a similar pattern of relative risks, and we found the following family of functions satisfactory (note that constant relative risks, or proportional hazards, result in odds ratios with exponential growth):

$$f^i_{x+t} = Ee^{-F((x+t)-k_1)^2 - G((x+t)-k_2)} + H. \tag{11}$$

Actuaries will recognize this as a GM(1,3) function, familiar in the graduation of life tables (Forfar, McCutcheon & Wilkie, 1988), although as described below either F or G is set to zero. We found this flexible enough to give a good approximation to the ORs, and also suitable for extrapolating beyond age 90. The fitting procedure was as follows:

(a) by considering the form of the OR, we set either $F = 0$ (giving an exponential function) or $G = 0$ (giving a bell-curve function), and set H equal to 0 or 1;
(b) the best value of k_1 or k_2 was found, to the nearest integer, by inspection;
(c) the resulting ORs were calculated from the model in Figure 1 using the intensities from previous sections; and
(d) the remaining coefficients (either E and F, or E and G) were fitted by least squares.

For the calculations in (c) above we used the following parameters:

(a) $\mu^{14}_{x+t} = 0.65 \times {}^{AM80}\mu_{x+t}$ for males and $\mu^{14}_{x+t} = 0.65 \times {}^{AF80}\mu_{x+t}$ for females and in aggregate, where ${}^{AM80}\mu_{x+t}$ and ${}^{AF80}\mu_{x+t}$ are given by equation 1.
(b) $\mu^{23}_{x+t} = 0.189$, calculated in Section 4.4.
(c) $\mu^{24}_{x+t} = 0.335 \times \mu^{14}_{x+t}$ calculated in Section 4.4.
(d) $\mu^{34}_{x+t} = 0.173 + \mu^{14}_{x+t}$, the mean value calculated in Section 4.5.

The fitted parameters are given in Table 7. Figure 6 shows that the modelled ORs closely reproduce the estimates from Farrer *et al.* (1997) (see Figure 3; only the aggregate ORs are shown, and the modelled ORs start at age 61 because we start with unaffected lives at age 60).

To determine the parameter r_1, we calculated the aggregate incidence of AD in the whole model, and fitted this to μ^{AD}_{x+t} by least squares. If ${}^iP^{11}_{xt}$ is the probability that a life with genotype i, healthy at age x, is unaffected by AD at age $x + t$, and if p^i_x is the population frequency of the ith genotype at age x then the aggregate incidence of AD is:

$$\text{Aggregate incidence of AD at age } (x+t) = r_1 \left\{ \sum_{i=1}^{5} p^i_x \, {}^iP^{11}_{xt} f^i_{x+t} \right\} \mu^{AD}_{x+t} \tag{12}$$

TABLE 7

PARAMETERS FOR THE RELATIVE RISK OF AD FOR MALES, FEMALES AND IN AGGREGATE, BY GENOTYPE

Gender	*Genotype*	*Parameter Values*						
		E	*F*	*G*	*H*	k_1	k_2	r_1
Both	$\varepsilon4/\varepsilon4$	13.5	0.00529	0	1	60	–	0.93
	$\varepsilon3/\varepsilon4$	2.98	0.00312	0	1	62	–	
	$\varepsilon2/\varepsilon4$	2.87	0.00938	0	1	68	–	
	$\varepsilon2/\varepsilon2$ & $\varepsilon2/\varepsilon3$	0.754	0	0.00859	0	–	60	
Female	$\varepsilon4/\varepsilon4$	10.4	0.00504	0	1	60	–	0.88
	$\varepsilon3/\varepsilon4$	3.68	0.00319	0	1	62	–	
	$\varepsilon2/\varepsilon4$	4.21	0.01020	0	1	68	–	
	$\varepsilon2/\varepsilon2$ & $\varepsilon2/\varepsilon3$	0.675	0	0.00692	0	–	60	
Male	$\varepsilon4/\varepsilon4$	8.94	0.00656	0	1	60	–	1.27
	$\varepsilon3/\varepsilon4$	1.92	0.00103	0	0	51	–	
	$\varepsilon2/\varepsilon4$	1.42	0.00506	0	0	67	–	
	$\varepsilon2/\varepsilon2$ & $\varepsilon2/\varepsilon3$	0.434	0	0.01600	0	–	60	

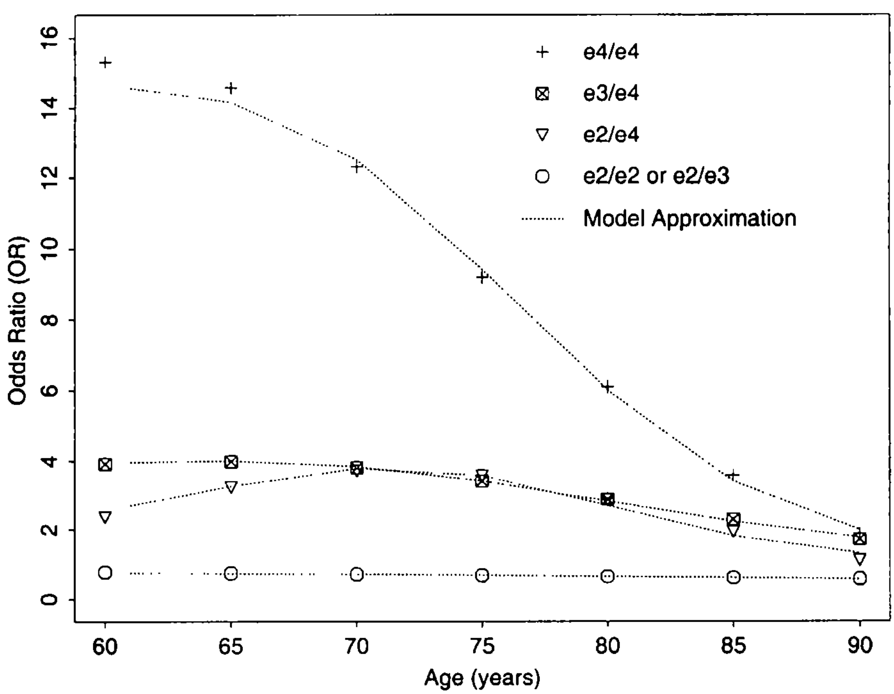


FIGURE 6: Odds ratios (ORs) of AD relative to $\varepsilon3/\varepsilon3$ genotype from Farrer *et al.* (1997), compared with ORs computed using modelled relative risk functions.

We took $x = 60$, and for the p_x^i we used the gene frequencies of the Caucasian control populations in Farrer *et al.* (1997), which were given in Section 5.1. The incidence of AD, μ_{x+t}^{AD}, was taken as that estimated in equation 2 and the occupancy probabilities, ${}^iP_{xt}^{11}$, were calculated by solving Kolmogorov's forward equations numerically using a Runge-Kutta algorithm with step size 0.0005 years (Conte & De Boor, 1972).

The values of r_1 are given in Table 7. The adjustment to the overall level only had a marginal effect on the modelled ORs for the individual genotypes.

The relative risk functions for males and females are given in Figures 7 and 8. For females, the $\varepsilon4/\varepsilon4$, $\varepsilon3/\varepsilon4$ and $\varepsilon2/\varepsilon4$ genotypes are unambiguously high-risk; the relative risks exceed 1.0 at all ages. For males, only the $\varepsilon4/\varepsilon4$ genotype confers higher risks at all ages. The $\varepsilon2$ allele appears to be protective, so the $\varepsilon2/\varepsilon2$ and $\varepsilon2/\varepsilon3$ genotypes are low-risk, while the $\varepsilon3/\varepsilon4$ and $\varepsilon2/\varepsilon4$ genotypes are initially at higher risk but are at lower risk from about age 75. The protection apparently given by the $\varepsilon2$ allele in males means that the $\varepsilon3/\varepsilon3$ genotype confers slightly higher than average risk; this is why, in Table 7, $r_1 > 1$ for males. It must be remembered that data in respect of males are relatively sparse, and data in respect of very old males even sparser, so these effects should be treated with caution; we can have more confidence in the relative risks in respect of females.

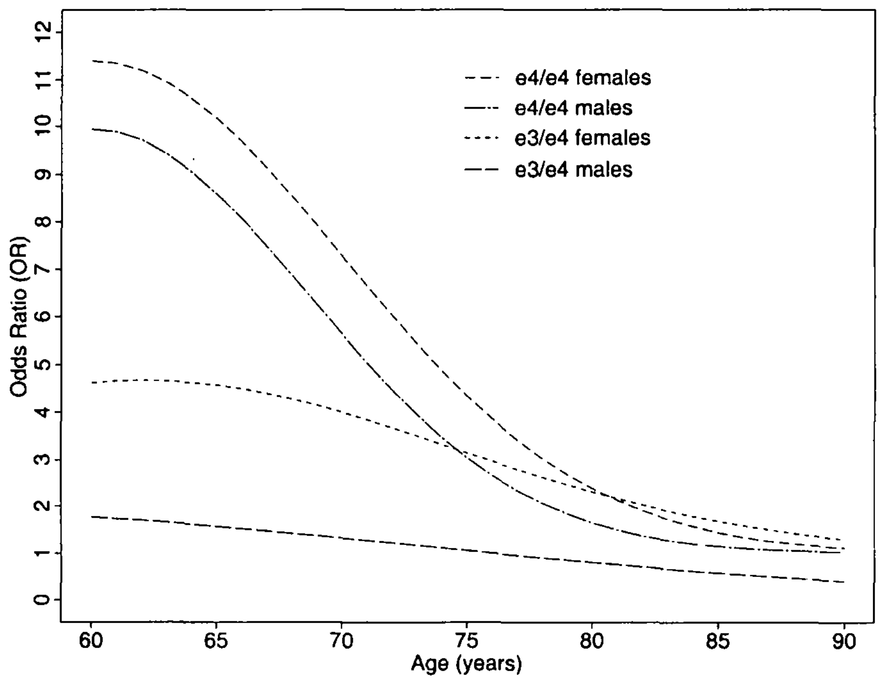


FIGURE 7: Modelled risk of AD, relative to the $\varepsilon3/\varepsilon3$ genotype, for $\varepsilon4/\varepsilon4$ and $\varepsilon3/\varepsilon4$ genotypes. Based on odds ratios from Farrer *et al.* (1997).

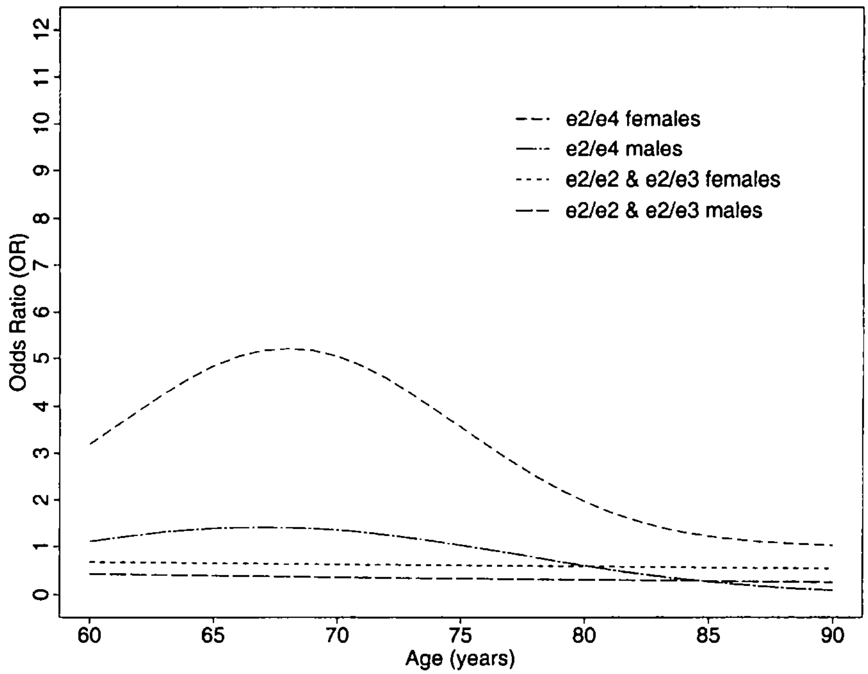


FIGURE 8: Modelled risk of AD, relative to the $\varepsilon 3/\varepsilon 3$ genotype, for $\varepsilon 2/\varepsilon 4$ and $\varepsilon 2/\varepsilon 2$ & $\varepsilon 2/\varepsilon 3$ genotypes. Based on odds ratios from Farrer *et al.* (1997).

These risk estimates probably overstate the true population risks, perhaps quite substantially, as they are from clinic- and autopsy-based studies, which investigate precisely the subjects that are affected or already known to be at risk. To allow for this possibility we will also consider models assuming that the true relative risks are a proportion $m < 1$ of those estimated above. We do this by adjusting equation (10) so that for genotype i:

$$\mu^{i12}_{x+t} = r_m\{(f^i_{x+t} - 1)m + 1\}\mu^{AD}_{x+t} \tag{13}$$

where f^i_{x+t} is as above, and r_m is chosen as above so that the aggregate incidence of AD in the model is consistent with μ^{AD}_{x+t}. Values of $r_{0.5}$ and $r_{0.25}$ are shown in Table 8.

TABLE 8

r_m FOR $m = 1$, 0.5 AND 0.25

Gender	r_1	$r_{0.5}$	$r_{0.25}$
Both	0.93	0.96	0.97
Female	0.88	0.94	0.97
Male	1.27	1.11	1.05

Figure 9 shows that the aggregate incidence of AD in the genetic model for both sexes combined is quite close to μ_{x+t}^{AD}. It also shows that increasing the level of relative risk tends to overstate the incidence of AD at younger ages, and to understate it at older ages; the reason is that higher relative risks deplete the high-risk groups more quickly, leaving a relatively healthier population at older ages.

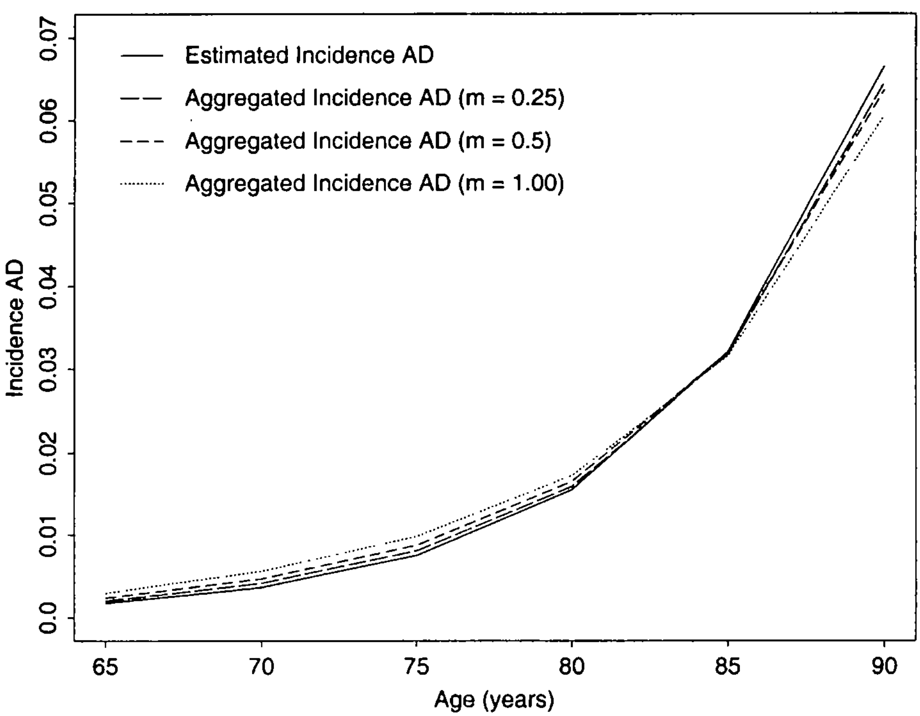


FIGURE 9: Comparison of estimated population incidence of AD μ_{x+t}^{AD} with the aggregated incidence of AD for different levels of relative risk, males and females combined.

Decreasing the level of relative risks for high-risk genotypes means increasing the relative risks for low-risk genotypes. Using a lower value of r_m will diminish any effects of the (possibly anomalous) feature, noted above, that the $\varepsilon3/\varepsilon4$ genotype is low-risk for males.

5.3. Comment on Model Selection

We chose a simple model for the relative risks (equations (10) and (11)). We did consider alternatives, in particular cubic polynomials and Gamma functions, but these gave poorer fits, and were less suitable for extrapolation (cubics to older ages and Gamma functions to younger ages). Also, it is easily seen that if an OR is specified as a function of time, and the transition intensity in the reference population is given, the transition intensity in the second population is determined (as the solution to an ODE); Figure 6, therefore, gives good support for our choice of model. Further refinement seemed somewhat spurious, given the data we were using, and in view of the major sensitivity analysis needed in respect of the dominant parameter m.

6. Results

6.1. Occupancy Probabilities

Figures 10 and 11 show occupancy probabilities up to age 90 for females healthy at age 60, with high ($m = 1$) and low ($m = 0.25$) relative risks, respectively. Each shows:

(a) Occupancy probabilities in respect of each genotype (with $\varepsilon2/\varepsilon2$ and $\varepsilon2/\varepsilon3$ combined).

(b) Occupancy probabilities calculated by aggregating all the genotypes in the model. In the notation of equation (12) the probability of being in state j ($j = 1, 2, 3, 4$) at age $60 + t$ is $\sum_{i=1}^{i=5} p_{60}^{i}\ {}^{i}P_{60,t}^{1j}$, where the sum is over all genotypes. These are labelled 'Aggregated Genotypes'.

(c) Occupancy probabilities based on the aggregate incidence of AD, μ_{x+t}^{AD}. These are labelled 'Aggregate Model'.

With high relative risks ($m = 1$), the effect of the $\varepsilon4/\varepsilon4$ allele is clear; AD cases rise to a peak in the early 70s, by which time over 10% of the original cohort are in one of the AD states, and then fall away. A similar but smaller effect can be seen for the $\varepsilon3/\varepsilon4$ genotype. With low relative risks ($m = 0.25$) these features are all diminished; in particular the peaks in the early 70s disappear.

We omit the corresponding figures for males; the differences are as we would expect, given the modelled relative risks.

For males and females with low relative risks (Figure 11) the aggregated results from the genetic model are very close to the results from the aggregate (population) model. For females with high relative risks, the rate of onset of AD seems to be too low at younger ages and too high at older ages.

6.2. Prevalence Rates

Also of interest are prevalence rates, namely the proportion of those alive at every age who are in each of the three live states. Figures 12 and 13 show these, for females, including, for convenience, the two AD states combined. We omit the corresponding figures for males.

The most striking feature is the prevalence of AD in respect of the $\varepsilon4/\varepsilon4$ genotype under high relative risks (Figure 12); it increases almost linearly. Again, for males and females the aggregated results from the genetic model are quite close to those from the aggregate model. Moreover, they fall within the range of prevalence rates actually observed. Breteler *et al.* (1992)

cite the following rates: 47.2% at ages 85 and over (Evans *et al.*, (1989)); 31.7% at ages 85 and over (Pfeffer *et al.*, 1987); and 28.0% at ages 90 and over (O'Connor *et al.*, 1989); some other studies gave lower figures.

6.3. Gene Frequencies at Higher Ages

We have assumed that the gene frequencies given by Farrer *et al.* (1997) are appropriate for age 60. They will change with age, as higher-risk genotypes die more quickly. We must estimate these if we wish to consider entrants (to a study, or into insurance) at ages over 60. Table 9 shows estimates of the gene frequencies in respect of lives unaffected by AD at ages 65, 70 and 75. Using the notation of equation (12), these are given by:

$$p^i_{60+t} = \frac{p^i_{60}\ {}^iP^{11}_{60,t}}{\sum_{j=1}^{j=5} p^j_{60}\ {}^jP^{11}_{60,t}}. \tag{14}$$

These are not the gene frequencies in respect of the whole population; lives alive but who have AD are omitted (as is appropriate for insurance applications). Nor are they the gene frequencies in respect of the healthy population; lives with disabilities other than AD are included.

Gene frequencies in the whole population at older ages can also be estimated, as:

$$\frac{p^i_{60}\left({}^iP^{11}_{60,t} + {}^iP^{12}_{60,t} + {}^iP^{13}_{60,t}\right)}{\sum_{j=1}^{j=5} p^j_{60}\left({}^jP^{11}_{60,t} + {}^jP^{12}_{60,t} + {}^jP^{13}_{60,t}\right)} \tag{15}$$

but these are not so relevant for insurance applications.

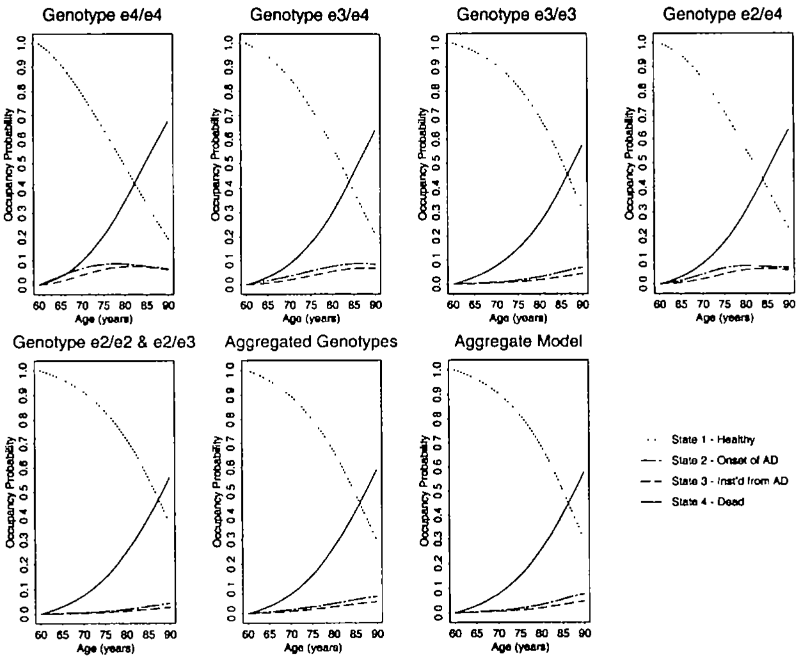


FIGURE 10: Occupancy probabilities for females, healthy at age 60, high relative risks ($m = 1$).

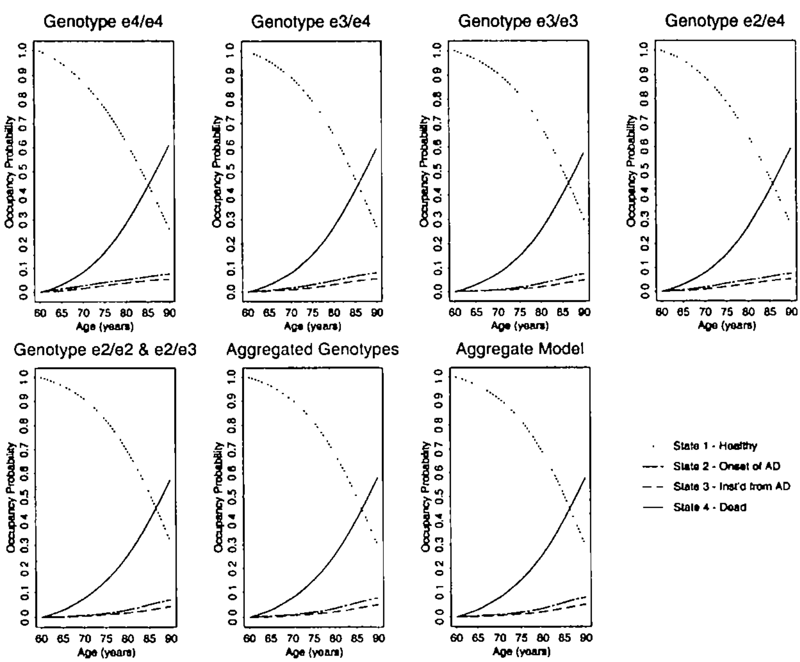


FIGURE 11: Occupancy probabilities for females, healthy at age 60, low relative risks ($m = 0.25$).

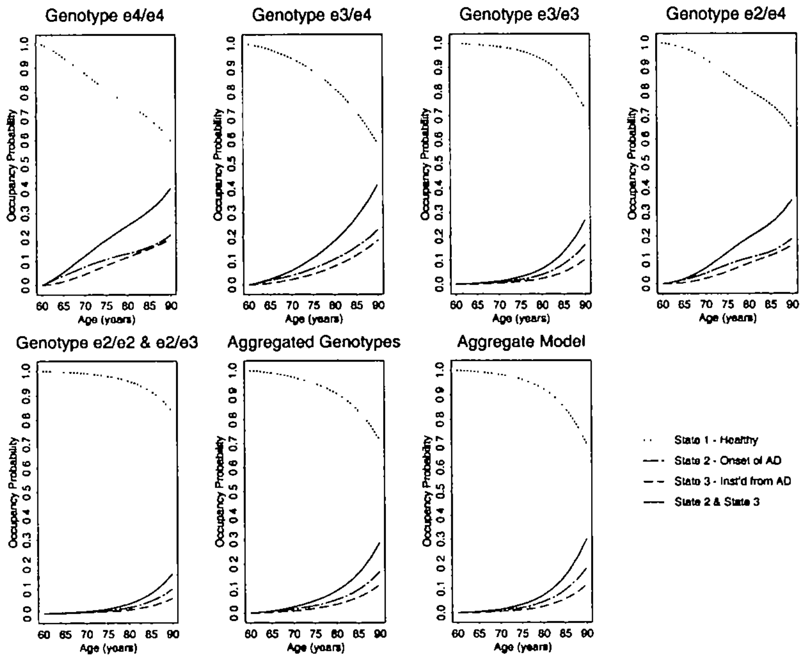


FIGURE 12: Prevalence rate of Alzheimer's disease for females healthy at age 60, high relative risks ($m = 1$).

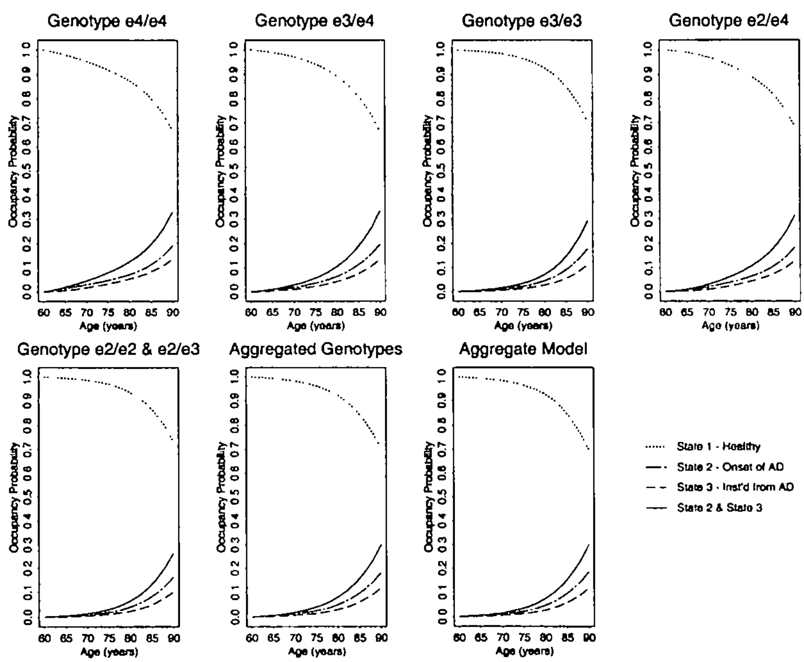


FIGURE 13: Prevalence rate of Alzheimer's disease for females healthy at age 60, low relative risks ($m = 0.25$).

TABLE 9

FREQUENCIES OF APOE GENOTYPES AMONG LIVES FREE OF ALZHEIMER'S DISEASE AT AGES 65, 70 AND 75, ESTIMATED BY SOLVING THE KOLMOGOROV EQUATIONS FORWARD FROM AGE 60

Gender	*Proportion of relative risk, m*	*Age*	*Gene Frequencies in AD-free population*				
			$\varepsilon 4/\varepsilon 4$	$\varepsilon 3/\varepsilon 4$	$\varepsilon 3/\varepsilon 3$	$\varepsilon 2/\varepsilon 4$	$\varepsilon 2/\varepsilon 2$ & $\varepsilon 2/\varepsilon 3$
M & F	1.00	65	0.0168	0.2103	0.6114	0.0258	0.1357
		70	0.0151	0.2055	0.6168	0.0251	0.1374
		75	0.0133	0.1978	0.6246	0.0241	0.1403
	0.50	65	0.0174	0.2115	0.6100	0.0259	0.1353
		70	0.0164	0.2090	0.6128	0.0255	0.1362
		75	0.0154	0.2050	0.6170	0.0250	0.1377
	0.25	65	0.0177	0.2121	0.6092	0.0259	0.1351
		70	0.0172	0.2109	0.6106	0.0258	0.1355
		75	0.0166	0.2088	0.6128	0.0255	0.1363
F	1.00	65	0.0171	0.2097	0.6116	0.0257	0.1358
		70	0.0159	0.2041	0.6176	0.0247	0.1377
		75	0.0145	0.1951	0.6263	0.0232	0.1409
	0.50	65	0.0175	0.2112	0.6101	0.0258	0.1354
		70	0.0168	0.2081	0.6133	0.0253	0.1364
		75	0.0160	0.2033	0.6181	0.0245	0.1381
	0.25	65	0.0177	0.2119	0.6093	0.0259	0.1351
		70	0.0174	0.2104	0.6110	0.0256	0.1357
		75	0.0169	0.2078	0.6135	0.0252	0.1365
M	1.00	65	0.0169	0.2120	0.6094	0.0260	0.1357
		70	0.0153	0.2110	0.6105	0.0258	0.1373
		75	0.0138	0.2097	0.6105	0.0257	0.1403
	0.50	65	0.0175	0.2125	0.6088	0.0260	0.1352
		70	0.0168	0.2120	0.6094	0.0259	0.1359
		75	0.0160	0.2115	0.6094	0.0258	0.1372
	0.25	65	0.0177	0.2126	0.6086	0.0260	0.1350
		70	0.0174	0.2124	0.6089	0.0259	0.1354
		75	0.0170	0.2122	0.6089	0.0259	0.1360

7. Conclusions

7.1. The Model

We have specified and calibrated a simple continuous-time Markov model of AD allowing for the variability of the ApoE gene, suitable for use in insurance and other applications, which will be the subject of further studies (for example, Macdonald & Pritchard (1999)). Much uncertainty remains:

(a) No single study yet exists that would allow all the intensities in the model to be estimated simultaneously. The estimation is based on a number of different studies, some quite small, of different populations, with different research protocols and methods of analysis, and very likely different definitions of 'onset of AD' and 'instititutionalisation'.

(b) The relative risks of the ApoE genotypes are based on case-based studies, not prospective population studies, and the risks associated with the $\varepsilon 4$ allele are almost certain to be lower than those estimated to date. We have been unable to do more than to show what effect this might have.

Nevertheless, certain features of our model ought to be robust. Whatever reduction in relative risks we have used, we have adjusted the genotype-specific incidence rates of AD so that the aggregated (population) incidence rates are close to those actually observed. The latter is one of the few reasonably reliable benchmarks available. Further, our model produces prevalence rates of AD that fall within the range of those observed in many studies.

As well as the intensities, we have estimated the ApoE gene frequencies at ages up to 75, in respect of lives unaffected by AD at these ages. These are needed in (for example) insurance applications.

7.2. Discussion

(a) The model specification is dictated entirely by the events studied in the medical and epidemiological literature, and not by the events that might be of interest in any particular application. If it is the case that actuarial models might, in future, need to incorporate more medical detail, it would be very useful to try to collaborate with medical and other researchers.

(b) The published conclusions of medical papers are usually in the form of summary statistics (means, medians, odds ratios and confidence intervals) and if age-related outcomes are shown they are usually in the form of graphs. These are not ideal for actuarial use. AD is a major condition, much studied, but we have had to make some crude assumptions in order to calibrate the model from published data only. There must be many medical data sets that could furnish age-related estimates of incidence rates, if only they could be re-analysed. Again, closer collaboration between actuaries and other researchers would be valuable.

(c) Another common type of medical statistic is prevalence rates. The difference between prevalence rates and incidence rates is exactly the difference between the Manchester Unity approach to modelling Permanent Health Insurance, and the multiple-state model approach. Prevalence rates are often easier to estimate, as they can be based on census-type surveys, but it would be helpful if the greater versatility of incidence rates (transition intensities) was more widely appreciated.

(d) As a consequence of fitting the intensities using published summary statistics, it is impossible to estimate even crude confidence intervals for them. In any application, therefore, sensitivity analysis is especially important.

(e) Several epidemiological authors have suggested the use of individual patient data rather than summary or published data, partly to avoid publication bias in meta-analyses. Useful references are Piantadosi (1997), Green, Benedetti & Crowley (1997) and Friedman, Furberg & DeMets (1998).

(f) It is now about six years since the rôle of the ApoE gene in AD was confirmed. Since then, the gene has been intensively studied, to the point that meta-analyses including thousands of lives have been published. Even so, little is known about population risk, and data are very scarce in places, so that:
   (1) we have had to reduce relative risks rather arbitrarily to allow for the selectiveness of case-based studies; and
   (2) the relative risks for males are suspect, with the $\varepsilon 2$ allele conferring such strong protection that carriers of the $\varepsilon 4$ allele are not necessarily at higher risk overall.

   If this is typical, it seems likely that the time between the identification of a gene disorder, and the assessment of its impact on insurance, will be of the order of ten years.

(g) Despite the fact that AD is a much-studied condition, many studies reach conflicting conclusions. When setting up an actuarial model, it is necessary to consider the body of medical research in its entirety (hence the large number of references). Inevitably, some studies must be chosen as the basis of the model, but to confine one's attention only to these risks overlooking important features or sources of variation, and could impair the credibility of the results in the eyes of medical experts.

(h) ApoE is a relatively simple gene to consider, since it has only three relevant alleles, hence six genotypes. Other genes are more complex; for example, the BRCA1 gene (predisposing to breast and ovarian cancer) has several hundred known mutations, some of which have only been observed in a single family.

(i) We cannot stress too strongly the speed at which human genetics is developing. This work started in late 1997, and since then the volume of papers on AD and the ApoE gene has increased greatly, as the references show. It is very likely that assessments of the impact of specific genetic tests on insurance will have to be revisited quite frequently, if they are to remain credible.

Points (a) to (c) above suggest ways in which medical data might be made more useful for actuarial models, but there is no *a priori* reason why medical studies should be planned with that in mind. However, actuarial models derived from insurance practice, capable of dealing with fairly general payments while in different states or on transition between different states, could make a useful contribution to health economics, and it would be helpful to pursue collaborations from that point of view.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Faculty and Institute of Actuaries, and one author (DJP) is in receipt of a Ph.D. studentship sponsored by Standard Life Assurance Company. We would like to thank Ross Ainslie, Dr Carol Brayne, Dr Peter Brett, Robert Plumb, Professor Andrew Read, Dr Virginia Warren, Professor Howard Waters and Dr Robert Watson for helpful discussions.

REFERENCES

BARCLAY, L.L., ZEMCOV, A., BLASS, J.P. and MCDOWELL F.H. (1985a) Factors associated with duration of survival in Alzheimer's disease. *Biological Psychiatry* **20**, 86-93.

BARCLAY, L.L., ZEMCOV, A., BLASS, J.P. and SANSONE, J. (1985b) Survival in Alzheimer's disease and vascular dementias. *Neurology* **35**, 834-840.

BASUN, H., GRUT, M., WINBLAD, B. and LANNFELDT, L. (1995) Apolipoprotein ε4 allele and disease progression in patients with late onset Alzheimer's disease. *Neuroscience Letters* **183**, 32-34.

BEARD, C.M., KOKMEN, E., O'BRIEN, P.C. and KURLAND, L.T. (1994) Are patients with Alzheimer's disease surviving longer in recent years? *Neurology* **44**, 1869-1871.

BERG, L., MILLER, J.P., STORANDT, M., DUCHEK, J., MORRIS, J.C., RUBIN, E.H., BURKE, W.J. and COBEN, L.A. (1988) Mild senile dementia of the Alzheimer type: 2. Longitudinal assessment. *Annals of Neurology* **23**, 477-484.

BICKEBÖLLER, H., CAMPION, D., BRICE, A., AMOUYEL, P., HANNEQUIN, D., DIDIERJEAN, O., PENET, C., MARTIN, C., PÉREZ-TUR, J., MICHON, A., DUBOIS, B., LEDOZE, F., THOMAS-ANTERION, C., PASQUIER, F., PUEL, M., DEMONET, J-F., MOREAUD, O., BABRON, M-C., MEULIEN, D., GUEZ, D., CHARTIER-HARLIN, M-C., FREBOURG, T., AGID, Y., MARTINEZ, M. and CLERGET-DARPOUX, F. (1997) Apolipoprotein E and Alzheimer's disease: genotype-specific risks by age and sex. *American Journal of Human Genetics* **60**, 439-446.

BONAIUTO, S., MELE, M., GALLUZO, L. and GIANNANDREA, E. (1995) Survival and dementia: A 7-year follow up of an Italian population. *Archives of Gerontology and Geriatrics* **20**, 105-113.

BRACCO, L., GALLATO, R., GRIGOLETTO, F., LIPPI, A., LEPORE, V., BINO, G., LAZZARO, M.P., CARELLA, F., PICCOLO, T., POZZILLI, C., GIOMETTO, B. and AMADUCCI, L. (1994) Factors affecting course and survival in Alzheimer's disease. *Archives of Neurology* **51**, 1213-1219.

BRETELER, M.B.B., CLAUS, J.J., VAN DUIJN, C.M., LAUNER, L.J. and HOFMAN, A. (1992) Epidemiology of Alzheimer's disease. *Epidemiologic Reviews* **14**, 59-82.

BRINDLE, N., SONG, Y., ROGAEVA, E., PREMKUMAR, S., LEVESQUE, G., YU, G., IKEDA, M., NISHIMURA, M., PATERSON, A., SORBI, S., DUARA, R., FARRER, L. and ST GEORGE-HYSLOP, P. (1998) Analysis of the butyrycholinesterase gene and nearby chromosome 3 markers in Alzheimer's disease. *Human Molecular Genetics* 7, 933-935.

BURNS, A., LEWIS, G., JACOBY, R. and LEVY, R. (1991) Factors affecting survival in Alzheimer's disease. *Psychological Medicine* **21**, 363-370.

CLAYTON, D. and HILLS, M. (1993) *Statistical methods in epidemiology*. Oxford University Press.

CONTE, S.D. and DE BOOR, C. (1972) *Elementary numerical analysis*. McGraw-Hill, New York.

COPELAND, J.R.M., DAVIDSON, I.A., DEWEY, M.E., GILMORE, C., LARKIN, B.A., MCWILLIAM, C., SAUNDERS, P.A., SCOTT, A., SHARMA, V. and SULLIVAN, C. (1992) Alzheimer's disease, other dementias, depression and pseudo-dementia: prevalence, incidence and three-year outcome in Liverpool. *British Journal of Psychiatry* **161**, 230-239.

CORDER, E.H., SAUNDERS, A.M., STRITTMATTER, W.J., SCHMECHEL, D.E., GASKELL, P.C., RIMMLER, J.B., LOCKE, P.A., CONNEALLY, P.M., SCHMADER, K.E., SMALL, G.W., ROSES, A.D., HAINES, J.L. and PERICAK-VANCE, M.A. (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer's disease. *Nature Genetics* 7, 180-184.

CORDER, E.H., SAUNDERS, A.M., STRITTMATTER, W.J., SCHMECHEL, D.E., GASKELL, P.C., RIMMLER, J.B., LOCKE, P.A., CONNEALLY, P.M., SCHMADER, K.E., TANZI, R.E., GUSELLA, J.F., SMALL, G.W., ROSES, A.D., PERICAK-VANCE, M.A. and HAINES, J.L. (1995) Apolipoprotein E, survival in Alzheimer's disease and the competing risks of death and Alzheimer's disease. *Neurology* **45**, 1323-1328.

DIESFELDT, H.F.A, VAN HOUTE, L.R. and MOERKENS, R.M. (1986) Duration of survival in senile dementia. *Acta Psychiatr Scandinavica* **73**, 366-371.

VAN DIJK, P.T.M., DIPPEL, D.W.J. and HABBEMA, J. DIK F. (1991) Survival of patients with dementia. *Journal of the American Geriatric Society* **39**, 603-610.

VAN DUIJN, C.M., DE KNIFF, P., WEHNERT, A., DE VOECHT, J., BRONZOVA, J.B., HAVEKES, L.M., HOFMAN, A. and VAN BROECKHOVEN, C. (1995) The apolipoprotein E $\varepsilon 4$ allele is associated with an increased risk of early-onset Alzheimer's disease and a reduced survival. *Annals of Neurology* **37**, 605-610.

EVANS, D.A., FUNKENSTEIN, H., ALBERT, M.S., SCHERR, P.A., COOK, N.R., CHOWN, M.J., HEBERT, L.E., HENNEKENS, C.H. and TAYLOR, J.O. (1989) Prevalence of Alzheimer's disease in a community population of older persons: Higher than previously reported. *Journal of the American Medical Association* **262**, 2551-2556.

EVANS, D.A., SMITH, L.A., SCHERR, P.A., ALBERT, M.S., FUNKENSTEIN, H.H. and HEBERT, L.E. (1991) Risk of death from Alzheimer's disease in a community population of older persons. *American Journal of Epidemiology* **134**, 403-412.

EVANS, D.A., BECKETT, L.A., FIELD, T.S., FENG, L., ALBERT, M.S., BENNETT, D.A., TYCKO, B. and MAYEUX R. (1997) Apolipoprotein E $\varepsilon 4$ and incidence of Alzheimer's disease in a community population of elder persons. *Journal of the American Medical Association* **277**, 822-824.

FARRER, L.A., CUPPLES, L.A., HAINES, J.L., HYMAN, B., KUKULL, W.A., MAYEUX, R., MYERS, R.H., PERICAK-VANCE, M.A., RISCH, N., VAN DUIJN, C.M. and APOE and ALZHEIMER DISEASE META ANALYSIS CONSORTIUM (1997) Effects of age, gender and ethnicity on the association between apolipoprotein E genotype and Alzheimer's disease. *Journal of the American Medical Association* **278**, 1349-1356.

FORFAR, D.O., MCCUTCHEON, J.J. and WILKIE, A.D. (1988) On graduation by mathematical formula. *J.I.A.* **115**, 1-149 and (with discussion) *T.F.A.* **41**, 97-269.

FRIEDMAN, L.M., FURBERG, C.D. and DEMETS, D.L. (1998) *Fundamentals of clinical trials*. Springer-Verlag, New York.

FRISONI, G.B., GOVONI, S., GEROLDI, C., BIANCHETTI, A., CALABRES, L., FRANCESCHINI, G. and TRABUCCHI, M. (1995) Gene dose of the $\varepsilon 4$ allele of apolipoprotein E and disease progression in sporodic late onset Alzheimer's disease. *Annals of Neurology* **37**, 596-604.

GAO, S., HENDRIE, H.C., HALL, K.S. and HUI, S. (1998) The relationship between age, sex and the incidence of dementia and Alzheimer's disease. *Archives of General Psychiatry* **55**, 809-815.

GOMEZ-ISLA, T., WEST, H.L., REBECK, G.W., HARR, S.D., GOWDON, J.H., LOCASCIO, J.J., PERLS, T.T., LEPSITZ, L.A. and HYMAN, B.T. (1996) Clinical and pathological correlates of apolipoprotein E ϵ4 in Alzheimer's disease. *Annals of Neurology* **39**, 62-70.

GREEN, S., BENEDETTI, J. and CROWLEY, J. (1997) *Clinical trials in oncology*. Chapman and Hall, London.

HAGNELL, O., OJESJO, L. and RORSMAN, B. (1992) Incidence of dementia in the Lundby study. *Neuroepidemiology* **11**, 61-66.

HEBERT, L.E., SCHERR, P.A., BECKETT, L.A., ALBERT, M.S., PILGRIM, D.M., CHOWN, M.J., FUNKENSTEIN, H.H. and EVANS, D.A. (1995) Age specific incidence of Alzheimer's disease in a community population. *Journal of the American Medical Association* **273**, 1354-1359.

HEYMAN, A., PETERSON, B., FILLENBAUM, G. and PIEPER, C. (1996) The consortium to establish a registry for Alzheimer's disease (CERAD) Part XIV: Demographic and clinical predictors of survival in patients with Alzheimer's disease. *Neurology* **46**, 656-660.

HEYMAN, A., PETERSON, B., FILLENBAUM, G. and PIEPER, C. (1997) Predictors of time to institutionalization of patients with Alzheimer's disease: the CERAD experience part XVII. *Neurology* **48**, 1304-1309.

JARVIK, G.P., LARSON, E.B., GODDARD, K., KUKULL, W.A., SCHELLENBERG, G.D. and WIJSMAN, E.M. (1996) Influence of apolipoprotein E genotype on the transmission of Alzheimer's disease in a community-based sample. *American Journal of Human Genetics* **58**, 191-200.

JORM, A.F. and JOLLEY, D. (1998) The incidence of dementia – a meta analysis. *Neurology* **51**, 728-733.

JOST, B.C. and GROSSBERG, G.T. (1995) The natural history of Alzheimer's disease: a brain bank study. *Journal of the American Geriatrics Society* **43**, 1248-1255.

KAHN, H. and SEMPOS, C.T. (1989) *Statistical methods in epidemiology*. Oxford University Press.

KOKMEN, E., CHANDRA, V. and SCHOENBERG, B. (1988) Trends in incidence of dementing illness in Rochester, Minnesota, in three quinquennial periods, 1960-1974. *Neurology* **38**, 975-980.

KOKMEN, E., BEARD, R.N., O'BRIEN, P.C., OFFORD, M.S. and KURLAND, L.T. (1993) Is the incidence of dementing illness changing? *Neurology* **43**, 1887-1892.

KUUSISTO, J., KOIVISTO, K., MYKKÄNEN, L., HELKALA, E-L., VANHANEN, M., HÄNNINEN, T., PYÖRÄLÄ, K., KESÄNIEMI, Y.A., RIEKKINON, P. and LAAKSO, K. (1994) Association of apolipoprotein E phenotypes with late onset Alzheimer's disease: population based study. *British Medical Journal* **309**, 636-638.

LAMBERT, J-C., PASQUIER, F., COTTEL, D., FRIGARD, B., AMOUYEL, P. and CHARTIER-HARLIN, M-C. (1998) A new polymorphism in the APOE promoter associated with risk of developing Alzheimer's disease, *Human Molecular Genetics* **7**, 533-540.

LEHMANN, D.J., JOHNSTON, C. and SMITH, A.D. (1997) Synergy between the genes for butyrylcholinesterase K variant and apolipoprotein E4 in late-onset confirmed Alzheimer's disease. *Human Molecular Genetics* **6**, 1933-1936.

LEHTOVIRTA, M., HELISALMI, S., MANNERMAA, A., SOININEN, H., KOIVISTO, K., RYYNÄNEN, M. and RIEKKINEN SR. P. (1995) Apolipoprotein E polymorphism and Alzheimer's disease in Eastern Finland. *Neuroscience Letters* **185**, 13-15.

LEMAIRE, J., SUBRAMANIAN, K., ARMSTRONG, K. and ASCH, D.A. (2000) Pricing term insurance in the presence of a family history of breast or ovarian cancer. *To appear in North American Actuarial Journal.*

LETENNEUR, L., COMMENGES, D., DARTIGUES, J.F. and BARBERGERGATEAU, P. (1994) Incidence of dementia and Alzheimer's disease in elderly community residents of south-western France. *International Journal of Epidemiology* **23**, 1256-1261.

LEVY-LAHAD, E. and BIRD, T.D. (1996) Genetic factors in Alzheimer's disease: a review of recent advances. *Annals of Neurology* **40**, 829-840.

LIDDELL, M., WILLIAMS, J., BAYER, A., KAISER, F. and OWEN, M. (1994) Confirmation of association between the ϵ4 allele of apolipoprotein E and Alzheimer's disease. *Journal of Medical Genetics* **31**, 197-200.

LILIENFELD, D.E. and STOLLEY, P.D. (1994) *Foundations of epidemiology*. Oxford University Press.

LOPEZ, O.L., LOPEZ-POUSA, S., KAMBOH, M.I., ADROER, R., OLIVA, R., LOZANO-GALLEGO, M., BECKER, T.B. and DEKOSKY, S.T. (1998) Apolipoprotein E polymorphism in Alzheimer's disease: a comparative study of two research populations from Spain and the United States. *European Neurology* **39**, 229-233.

LUCOTTE, G., DAVID, F., LEGRAND, C., TURPIN, J-C. and CLAVEL, C. (1997) Predictive value of apolipoprotein E $\varepsilon 4$ allele genotyping in Alzheimer's disease. *Alzheimer's Research* **3**, 7-9.

MACDONALD, A.S. (1996a) An actuarial survey of statistical models for decrement and transition data I: Multiple state, Poisson and binomial models. *British Actuarial Journal* **2**, 129-155.

MACDONALD, A.S. (1996b) An actuarial survey of statistical models for decrement and transition data III: Counting process models. *British Actuarial Journal* **2**, 703-726.

MACDONALD, A.S. (1997) How will improved forecasts of individual lifetimes affect underwriting?. *Philosophical Transactions of the Royal Society* B, **352**, 1067-1075, and (with discussion) *British Actuarial Journal* **3**, 1009-1025 and 1044-1058.

MACDONALD, A.S. (1999) Modeling the impact of genetics on insurance. *North American Actuarial Journal* **3:1**, 83-101.

MACDONALD, A.S. and PRITCHARD, D.J. (1999) Genetics, Alzheimer's disease and long-term care insurance. *Genetics and Insurance Research Report No. 99/2, Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh.*

MASSULLO, C., DANIELE, A., SERIPA, D., FILIPPINI, V., GRAVINA, C., CARBONE, G., GAINOTTI, G. and FAZIO, V.M. (1998) Apolipoprotein E genotype in sporadic early- and late-onset Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders* **9**, 121-125.

MAYEUX, R., STERN, Y., OTTMAN, R., TATEMICHI, T.K., TANG, M-X., MAESTRE, G., NGAI, C., TYCKO, B. and GINSBERG, H. (1993) The apolipoprotein $\varepsilon 4$ allele in patients with Alzheimer's disease. *Annals of Neurology* **34**, 752-754.

MÖLSÄ, P.K., MARTTILA, R.J. and RINNES, U.K. (1982) Epidemiology of dementia in a Finnish population. *Acta Neurologica Scandinavica* **65**, 541-552.

MÖLSÄ, P.K., MARTTILA, R.J. and RINNES, U.K. (1986) Survival and cause of death in Alzheimer's disease and multi-infarct dementia. *Acta Neurologica Scandinavica* **74**, 103-107.

MYERS, R.H., SCHAEFFER, E.J., WILSON, P.W.F., D'AGOSTINO, R., ORDOVAS, J.M., ESPINO, A., WHITE, R.F., KNOEFEL, J.E., COBB, J.L., MCNULTY, K.A., BEISER, A. and WOLF, P.A. (1996) Apolipoprotein E $\varepsilon 4$ association with dementia in a population based study: the Framingham study. *Neurology* **46**, 673-677.

NALBANTOGLU, J., GILFIX, B.M., BERTRAND, P., ROBITAILLE, Y., GAUTHIER, S., ROSENBLATT, D.S. and POIRIER, J. (1994) Predictive value of apolipoprotein E genotyping in Alzheimer's disease: results of an autopsy series and an analysis of several combined studies. *Annals of Neurology* **36**, 889-895.

NILSSON, L.V. (1984) Incidence of severe dementia in an urban sample followed from 70 to 79 years of age. *Acta Psychiatrica Scandinavica* **70**, 478-486.

NORBERG, R. (1995) Differential equations for moments of present values in life insurance. *Insurance: Mathematics and Economics* **17**, 171-180.

NORRMAN, J., BROOKES, A.J., YATES, C. and CLAIR, D.S. (1995) Apolipoprotein E genotype and its effect on duration and severity of early and late onset Alzheimer's disease. *British Journal of Psychiatry* **167**, 533-536.

O'CONNOR, D.W., POLLITT, P.A., HYDE, J.B. *et al.* (1989) The prevalence of dementia as measured by the Cambridge Mental Disorders of the Elderly Examination. *Acta Psychiatrica Scandinavica* **79**, 190-198.

OTT, A., BRETELER, M.M.B., VAN HARSKAMP, F., STIJNEN T. and HOFMAN, A. (1998) Incidence and risk of dementia. *American Journal of Epidemiology* **147**, 574-580.

PAYAMI, H., SCHELLENBERG, G.D., ZAREPARSI, S., KAYE, J., SEXTON, G.J., HEAD, M.A., MATSUYAMA, S.S., JARVIK, L.F., MILLER, B., MCMANUS, D.Q., BIRD, T.D., KATZMAN, R., HESTON, L., NORMAN, D. and SMALL, G.W. (1997) Evidence for association of HLA-A2 allele with onset age of Alzheimer's disease. *Neurology* **48**, 512-518.

PERICAK-VANCE, M.A., BEBOUT, J.L., GASKELL, P.C., YAMAOKA, L.H., HUNG, W.Y., ALBERTS, M.J., WALKER, A.P., BARTLETT, R.J., HAYNES, C.A., WELSH, K.A., EARL, N.L., HEYMAN, A., CLARK, C.M. and ROSES, A.D. (1991) Linkage studies in familial Alzheimer's disease: evidence for chromosome 19 linkage. *American Journal of Human Genetics* **48**, 1034-1050.

PFEFFER, R.I., AFIFI, A.A. and CHANCE J.M. (1987) Prevalence of Alzheimer's disease in a retirement community. *American Journal of Epidemiology* **125**, 420-436.

PIANTADOSI, S. (1997) *Clinical trials: A methodologic perspective*. John Wiley and Sons, New York.

PODUSLO, S.E., NEAL, M., HERRING, K. and SHELLY, J. (1998) The apolipoprotein CI A allele as a risk factor for Alzheimer's disease. *Neurochemical Research* **23**. 361-367.

POIRIER, J., DAVIGNON, J., BOUTHILLIER, D., KOGAN, S., BERTRAND, P. and GAUTHIER, S. (1993) Apolipoprotein E polymorphism and Alzheimer's disease. *Lancet* **342**, 697-699.

PRITCHARD, D.J. (1997) *Life assurance: financial implications of a change in insuring behaviour resulting from individuals' increased knowledge of their genetic predispositions*. M.Sc. dissertation, Heriot-Watt University, Edinburgh.

ROCCA, W.A., CHA, R.H., WARING, S.C. and KOKMEN, E. (1998) Incidence of dementia and Alzheimer's disease: a reanalysis of data from Rochester, Minnesota, 1975-1984. *American Journal of Epidemiology* **148**, 51-62.

RORSMAN, B., HAGNELL, O. and LANKE, J. (1986) Prevalence and incidence of senile and multi-infarct dementia in the Lundby study: a comparison between the time periods 1947-1957 and 1957-1972. *Neuropsychobiology* **15**, 122-129.

ROSES, A.D. (1995) Apolipoprotein E genotyping in the differential diagnosis, not prediction, of Alzheimer's disease. *Annals of Neurology* **38**, 6-14.

SAYETTA, R.B. (1986) Rates of senile dementia – Alzheimer's type in the Baltimore longitudinal study. *Journal of Chronic Disabilities* **39**, 271-286.

SELVIN, S. (1996) *Statistical analysis of epidemiologic data*. Oxford University Press.

SEVERSON, M.A., SMITH, G.E., TANGALOS, E.G., PETERSON, R.C., KOKMEN, E., IVNIK, R.J., ATKINSON, E.J. and KURLAND, L.T. (1994) Patterns and predictors of institutionalization in community-based dementia patients. *Journal of the American Geriatrics Society* **42**, 181-185.

SINGLETON, A., SMITH, G., GIBSON, A.M., WOODWARD, R., PERRY, R.H., INCE, P.G., EDWARDSON, J.A. and MORRIS, C.M. (1998) No association between the K variant of the butyrylcholinesterase gene and pathologically confirmed Alzheimer's disease. *Human Molecular Genetics* **7**, 937-939.

SLOOTER, A.J.C. and VAN DUIJN, C.M. (1997) Genetic epidemiology of Alzheimer disease. *Epidemiologic Reviews* **19**, 107-119.

SLOOTER, A.J.C., CRUTS, M., KALMIJN, S., HOFMAN, A., BRETELER, M.M.B., VAN BROECKHOVEN, C. and VAN DUIJN, C.M. (1998) Risk estimates of dementia by apolipoprotein E genotype from a population-based incidence study: the Rotterdam study. *Archives of Neurology* **55**, 964-968.

STERN, G., TANG, M.X., DENARO, J. and MAYEUX, R. (1995) Increased risk of mortality in Alzheimer's disease patients with more advanced educational and occupational attainment. *Annals of Neurology* **37**, 590-595.

STERN, Y., BRANDT, J., ALBERT, M., JACOBS, D.M., LIU, X., BELL, K., MARDER, K., SANO, M., ALBERT, S., CASTENADA, C.D-C., BYLSMA, F., TYCKO, B. and MAYEUX, R. (1997) The absence of an apolipoprotein $\varepsilon 4$ allele is associated with a more aggressive form of Alzheimer's disease. *Annals of Neurology* **41**, 615-620.

STRACHAN, T. and READ, A.P. (1999) *Human molecular genetics (second edition)*. BIOS Scientific Publishers, Oxford.

STRITTMATTER, W.J., SAUNDERS, A.M., SCHMECHEL, D., PERICAK-VANCE, M., ENGHILD, J., SALVESEN, G.S. and ROSES, A.D. (1993) Apolipoprotein E: High-avidity binding to β-amyloid and increased frequency of type-4 allele in late-onset Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 1977-1981.

SUBRAMANIAN, K. and LEMAIRE, J. (2000) Estimating adverse selection costs in a market with genetic testing for breast and ovarian cancer: the case of life insurance. *To appear in Journal of Risk and Insurance*.

TAN, K.W. (1997) *The financial impact of genetic testing on annuities.* M.Sc. dissertation, Heriot-Watt University, Edinburgh.

TIERNEY, M.C., FISHER, R.H., LEWIS, A.J., ZORZITTO, M.L., SNOW, W.G., REID, D.W. and NIEUWSTRATEN, P. (1988) The NINCDS-ADRDA work group criteria for the clinical diagnosis of probable Alzheimer's disease: a clinicopathologic study of 57 cases. *Neurology* **38**, 359-364.

TREVES, T., KORCZYN, A.D., ZILBER, N., KAHANA, E., LEIBOWITZ, Y., ALTER, M. and SCHOENBERG, B.S. (1986) Presenile dementia in Israel. *Archives of Neurology* **43**, 26-29.

TSAI, M-S., TANGALOS, E.G., PETERSEN, R.C., SMITH, G.E., SCHAID, D.J., KOKMEN, E., IVNIK, R.J. and THIBODEAU, S.N. (1994) Apolipoprotein E: risk factor for Alzheimer's disease. *American Journal of Human Genetics* **54**, 643-649.

WALSH, J.S., WELCH, H.G. and LARSON, E.B. (1990) Survival of outpatients with Alzheimer-type dementia. *Annals of Internal Medicine* **113**, 429-434.

WARREN, V., BRETT, P., MACDONALD, A.S., PLUMB, R.H. and READ, A.P. (1999) *Genetic tests and future need for long-term care in the UK: report of a Work Group of the Continuing Care Conference Genetic Tests and Long-term Care Study Group.* Continuing Care Conference, London.

WATERS, H.R. (1984) An approach to the study of multiple state models. *Journal of the Institute of Actuaries* **111**, 363-374.

WATERS, H.R. and WILKIE, A.D. (1987) A short note on the construction of life tables and multiple decrement tables. *Journal of the Institute of Actuaries* **114**, 569-580.

WATERS, H.R. (1991) Computational procedures for the model. *Continuous Mortality Investigation Bureau Report* **12**, 79-96.

ANGUS MACDONALD AND DELME PRITCHARD
Department of Actuarial Mathematics and Statistics
Heriot-Watt University
Edinburgh EH14 4AS
United Kingdom

ON MULTIVARIATE VERNIC RECURSIONS

BY

BJØRN SUNDT

University of Bergen

ABSTRACT

In the present paper we extend a recursive algorithm developed by Vernic (1999) for compound distributions with bivariate counting distribution and univariate severity distributions to more general multivariate counting distributions.

1. INTRODUCTION

1A. Panjer (1981) described a procedure for recursive evaluation of a compound distribution when the counting distribution belongs to a certain class. Vernic (1999) developed a bivariate version of this recursion, assuming that the counting distribution is bivariate and the severity distributions univariate. In the present paper we discuss a generalisation of the result of Vernic to a situation with an m-variate counting distribution and a univariate severity distribution.

The recursions of Panjer and Vernic are briefly recapitulated in Sections 2 and 3 respectively, and the multivariate extension is introduced in Section 4. In Section 5 we look at some examples, and, finally, in Section 6 we briefly indicate some possible extensions of the theory.

1B. In the recursions that we study in the present paper, the distributions are expressed through their probability functions. For simplicity we shall therefore normally mean the probability function when referring to a distribution.

We make the convention that a summation over an empty set is equal to zero and multiplication over an empty set is equal to one.

2. THE RECURSION OF PANJER

In the univariate case, a compound distribution is the distribution of the sum of independent and identically distributed random variables where the number of terms is itself a random variable assumed to be independent of the terms. We shall assume that the terms are distributed on the positive

ASTIN BULLETIN, Vol. 30, No 1, 2000, pp. 111-122

integers. Let p be the distribution of the number of terms (the counting distribution), f the distribution of the terms (the severity distribution), and g the compound distribution. Then $g = \sum_{n=0}^{\infty} p(n) f^{n*}$. As f is confined to the positive integers, we must have $f^{n*}(x) = 0$ for all integers $n > x$, and thus

$$g(x) = \sum_{n=0}^{x} p(n) f^{n*}(x); \quad (x = 0, 1, 2, ...)$$

in particular we have $g(0) = p(0)$.

If p satisfies the recursion

$$p(n) = \left(a + \frac{b}{n}\right) p(n-1), \quad (n = 1, 2, ...)$$

then

$$g(x) = \sum_{y=1}^{x} \left(a + b\frac{y}{x}\right) f(y) g(x-y). \quad (x = 1, 2, ...)$$

This recursion was described by Panjer (1981).

3. The recursion of Vernic

When extending the concept of compound distributions to the multivariate case, one can go in two directions:

1. Let the severities be independent and identically distributed random vectors.
2. Let the counting distribution be multivariate and the severities one-dimensional; we consider the distribution of, say, m random variables with compound distributions whose counting variables are dependent whereas the severities are mutually independent and independent of the counting variables.

The two approaches can be combined by letting the severities in Case 2 be random vectors.

For Case 1 recursions have been studied by Ambagaspitiya (1999) and Sundt (1999); for Case 2 by Hesselager (1996) and Vernic (1999) in the bivariate case.

In Case 2 the compound distribution is given by

$$g = \sum_{n_1=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} p(n_1, ..., n_m) \prod_{i=1}^{m} f_i^{n_i *}. \tag{1}$$

When assuming that the severity distributions are restricted to the positive integers, like in the univariate case, we obtain that the infinite summations become finite when we insert an argument in g:

$$g(x_1,...,x_m) = \sum_{n_1=0}^{x_1} \cdots \sum_{n_m=0}^{x_m} p(n_1,...,n_m) \prod_{i=1}^{m} f_i^{n_i *}(x_i); \quad (x_1,...,x_m = 0,1,2,...)$$

in particular we have $g(0,...,0) = p(0,...,0)$.

Let us turn to the bivariate case. Vernic (1999) assumed that

$$p(n_1,n_2) = \psi_{12}(n_1,n_2)p(n_1-1,n_2-1) + \psi_1(n_1,n_2)p(n_1-1,n_2) + \psi_2(n_1,n_2)p(n_1,n_2-1) \tag{2}$$

when at least one of n_1 and n_2 are positive, with

$$\psi_{12}(n_1,n_2) = \begin{cases} a_0 + \dfrac{a_1}{n_1} + \dfrac{a_2}{n_2} + \dfrac{a_{12}}{n_1 n_2} & (n_1,n_2 = 1,2,...) \\ 0 & (\text{otherwise}) \end{cases}$$

$$\psi_1(n_1,n_2) = \begin{cases} b_0 + \dfrac{b_1}{n_1} & (n_1,n_2 = 1,2,...) \\ d_0 + \dfrac{d_1}{n_1} & (n_1 = 1,2,...;\ n_2 = 0) \\ 0 & (n_1 = 0;\ n_2 = 1,2,...) \end{cases}$$

$$\psi_2(n_1,n_2) = \begin{cases} c_0 + \dfrac{c_2}{n_2} & (n_1,n_2 = 1,2,...) \\ e_0 + \dfrac{e_2}{n_2} & (n_1 = 0,\ n_2 = 1,2,...) \\ 0 & (n_1 = 1,2,...;\ n_2 = 0) \end{cases}$$

and showed that then

$$g(x_1,x_2) = \sum_{y_1=1}^{x_1}\sum_{y_2=1}^{x_2} \varphi_{12}(y_1,y_2;x_1,x_2) f_1(y_1) f_2(y_2) g(x_1-y_1,x_2-y_2) +$$
$$\sum_{y_1=1}^{x_1} \varphi_1(y_1;x_1,x_2) f_1(y_1) g(x_1-y_1,x_2) +$$
$$\sum_{y_2=1}^{x_2} \varphi_2(y_2;x_1,x_2) f_2(y_2) g(x_1,x_2-y_2) \tag{3}$$

when at least one of x_1 and x_2 are positive, with

$$\psi_{12}(y_1, y_2; x_1, x_2) = \begin{cases} a_0 + a_1 \dfrac{y_1}{x_1} + a_2 \dfrac{y_2}{x_2} + a_{12} \dfrac{y_1 y_2}{x_1 x_2} & (y_i = 1, \dots, x_i; \\ & x_i = 1, 2, \dots;\ i = 1, 2) \\ 0 & \text{(otherwise)} \end{cases}$$

$$\varphi_1(y_1; x_1, x_2) = \begin{cases} b_0 + b_1 \dfrac{y_1}{x_1} & (y_i = 1, \dots, x_i;\ x_i = 1, 2, \dots;\ i = 1, 2) \\ d_0 + d_1 \dfrac{y_1}{x_1} & (y_1 = 1, \dots, x_1;\ x_1 = 1, 2, \dots;\ x_2 = 0) \\ 0 & \text{(otherwise)} \end{cases}$$

$$\varphi_2(y_2; x_1, x_2) = \begin{cases} c_0 + c_2 \dfrac{y_2}{x_2} & (y_i = 1, \dots, x_i;\ x_i = 1, 2, \dots;\ i = 1, 2) \\ e_0 + e_2 \dfrac{y_2}{x_2} & (y_2 = 1, \dots, x_2;\ x_1 = 0;\ x_2 = 1, 2, \dots) \\ 0. & \text{(otherwise)} \end{cases}$$

Some special cases are studied by Hesselager (1996).

We see that already in the bivariate case the formulae and notation start getting rather messy, and unfortunately it will get even worse when extending the theory to a more general multivariate case. We shall therefore abstain from writing out a general theory in full and rather give a rough outline of what can be done.

4. General results

4A. When considering extension of the Vernic recursions from the bivariate case to the m-variate case, it will be convenient to use some vector notation. We shall denote an m x 1 column vector by a bold-face letter and its elements by the corresponding italic with the number of the element as subscript; subscript $\cdot$ denotes the sum of the elements, e.g. $\mathbf{x} = (x_1, \dots, x_m)'$ and $x_{\cdot} = \sum_{i=1}^{m} x_i$. By $\mathbf{y} \leq \mathbf{x}$ we shall mean that $y_i \leq x_i$ for $i = 1, \dots, m$, and by $\mathbf{y} < \mathbf{x}$ that $\mathbf{y} \leq \mathbf{x}$ with $\mathbf{y} \neq \mathbf{x}$. By $\mathbf{e}_{i_1 \dots i_h}$ we shall mean the vector whose i_jth element is equal to one for $j = 1, \dots, h$, and all other elements are equal to zero. We also introduce the vector $\mathbf{0}$ where all elements are equal to zero.

It is tacitly assumed that all vectors introduced have integer-valued elements.

4B. Let $\mathbf{N}$ be an $m \times 1$ vector of non-negative integer-valued random variables. We introduce positive, integer-valued random variables Y_{ij} ($i = 1, \dots, m$; $j = 1, 2, \dots$), assumed to be independent of $\mathbf{N}$ and mutually independent, and for fixed i identically distributed with common distribution

f_i. Let p denote the distribution of $\mathbf{N}$. We introduce the random vector $\mathbf{X} = (X_1, ..., X_m)'$ with $X_i = \sum_{j=1}^{N_i} Y_{ij}$ for $i = 1, ..., m$. Then the distribution of $\mathbf{X}$ is the compound distribution g given by (1).

4C. When trying to extend (2) and (3) to an m-variate situation, it is natural to look for pairs of functions $(\psi_{i_1...i_h}, \varphi_{i_1...i_h})$ such that

$$p(\mathbf{n}) = \sum_{h=1}^{m} \sum_{1 \le i_1 < ... < i_h \le m} \psi_{i_1...i_h}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{i_1...i_h}) \quad (\mathbf{n} > \mathbf{0}) \tag{4}$$

$$g(\mathbf{x}) = \sum_{h=1}^{m} \sum_{1 \le i_1 < ... < i_h \le m} \sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi_{i_1...i_h}(y_1, ..., y_h;\ \mathbf{x}) \times$$
$$g\left(\mathbf{x} - \sum\nolimits_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}(y_j). \quad (\mathbf{x} > \mathbf{0}) \tag{5}$$

Like in the Vernic recursion, we would normally have that for $i \in \{1, ..., m\} \sim \{i_1...i_h\}$ $\psi_{i_1...i_h}(\mathbf{n})$ and $\varphi_{i_1...i_h}(y_1, ..., y_h; \mathbf{x})$ depend on n_i and x_i respectively only to the extent of whether they are equal to zero or not.

The following lemma describes the relation we need between a ψ and the corresponding φ.

Lemma 1. *If for different integers $i_1, ..., i_h \in \{1, ..., m\}$*

$$\mathrm{E}\left[\varphi(Y_{i_1 1}, ..., Y_{i_h 1};\ \mathbf{x}) \middle| \bigcap_{j=1}^{h} \left(\sum_{r=1}^{n_{i_j}} Y_{i_j r} = x_{i_j}\right)\right] = \psi(\mathbf{n}) \tag{6}$$

for all $\mathbf{x}, \mathbf{n} > \mathbf{0}$ such that $\prod_{i=1}^{m} f_i^{n_i *}(x_i) > 0$, then

$$\sum_{\mathbf{n} > \mathbf{0}} \psi(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{i_1...i_h}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$
$$\sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, ..., y_h;\ \mathbf{x}) g\left(\mathbf{x} - \sum\nolimits_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}(y_j). \quad (\mathbf{x} > \mathbf{0})$$

Proof. We extend the set $\{i_1, ..., i_h\}$ to a permutation $\{i_1, ..., i_m\}$ of $\{1, ..., m\}$. For all $\mathbf{x} > \mathbf{0}$ we have

$$\sum_{\mathbf{n}>\mathbf{0}} \psi(\mathbf{n})p(\mathbf{n} - \mathbf{e}_{i_1...i_h}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$

$$\sum_{\mathbf{n}>\mathbf{0}} p(\mathbf{n} - \mathbf{e}_{i_1...i_h})\mathrm{E}\left[\varphi(Y_{i_1 1}, ..., Y_{i_h 1};\ \mathbf{x}) \middle| \bigcap_{j=1}^{h} \left(\sum_{r=1}^{n_{i_j}} Y_{i_j r} = x_{i_j}\right)\right] \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$

$$\sum_{\mathbf{n}>\mathbf{0}} p(\mathbf{n} - \mathbf{e}_{i_1...i_h}) \sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, ..., y_h;\ \mathbf{x}) \times$$

$$\left(\prod_{j=1}^{h} f_{i_j}(y_j) f_{i_j}^{(n_{i_j}-1)*}(x_{i_j} - y_j)\right) \prod_{j=h+1}^{m} f_{i_j}^{n_{i_j}*}(x_{i_j}) =$$

$$\sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, ..., y_h;\ \mathbf{x}) \left(\prod_{j=1}^{h} f_{i_j}(y_j)\right) \times$$

$$\sum_{\mathbf{n}>\mathbf{0}} p(\mathbf{n} - \mathbf{e}_{i_1...i_h}) \left(\prod_{j=1}^{h} f_{i_j}^{(n_{i_j}-1)*}(x_{i_j} - y_j)\right) \prod_{j=h+1}^{m} f_{i_j}^{n_{i_j}*}(x_{i_j}) =$$

$$\sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, ..., y_h;\ \mathbf{x}) g\left(\mathbf{x} - \sum_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}(y_j).$$

Q.E.D.

In the univariate case Lemma 1 is closely related to Theorem 2 in Sundt & Jewell (1981).

It is clear that if the pairs $(\varphi_1, \psi_1), ..., (\varphi_w, \psi_w)$ satisfy the conditions of Lemma 1, then $\left(\sum_{v=1}^{w} c_v \varphi_v,\ \sum_{v=1}^{w} c_v \psi_v\right)$ also satisfies the conditions of Lemma 1 for all constants $c_1, ..., c_w$.

As the severities are positive, $X_i = 0$ if and only if $N_i = 0$. This implies that if the pairs (φ_1, ψ_1) and (φ_2, ψ_2) satisfy the conditions of Lemma 1, then these conditions are also satisfied by the pair (φ, ψ) given by

$$\varphi(y_1, ..., y_h;\ \mathbf{x}) = \begin{cases} \varphi_1(y_1, ..., y_h;\ \mathbf{x}) & (x_i = 1, 2, ...) \\ \varphi_2(y_1, ..., y_h;\ \mathbf{x}) & (x_i = 0) \end{cases}$$

$$\psi(\mathbf{n}) = \begin{cases} \psi_1(\mathbf{n}) & (n_i = 1, 2, ...) \\ \psi_2(\mathbf{n}). & (n_i = 0) \end{cases}$$

We have already seen one application of such a construction in the Vernic recursion, where the coefficients were allowed to depend on whether some of the variables were equal to zero.

We are now ready to prove our main theorem.

Theorem 1. *If there exist pairs of function* $(\psi_{i_1\ldots i_h}, \varphi_{i_1\ldots i_h})$ *such that* (4) *holds and each pair satisfies* (6) *for all* $\mathbf{x}, \mathbf{n} > \mathbf{0}$ *such that* $\prod_{i=1}^{m} f_i^{n_i *}(x_i) > 0$, *then* (5) *holds*.

Proof. From Lemma 1 we obtain that for all $\mathbf{x} > \mathbf{0}$

$$g(\mathbf{x}) = \sum_{\mathbf{n}>\mathbf{0}} p(\mathbf{n}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$

$$\sum_{\mathbf{n}>\mathbf{0}} \sum_{h=1}^{m} \sum_{1 \le i_1 < \ldots < i_h \le m} \psi_{i_1\ldots i_h}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{i_1\ldots i_h}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$

$$\sum_{h=1}^{m} \sum_{1 \le i_1 < \ldots < i_h \le m} \sum_{\mathbf{n}>\mathbf{0}} \psi_{i_1\ldots i_h}(\mathbf{n}) p(\mathbf{n} - \mathbf{e}_{i_1\ldots i_h}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$

$$\sum_{h=1}^{m} \sum_{1 \le i_1 < \ldots < i_h \le m} \sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi_{i_1\ldots i_h}(y_1, \ldots, y_h;\ \mathbf{x}) g\left(\mathbf{x} - \sum_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}(y_j).$$

Q.E.D.

Our next theorem shows a way to construct additional recursions for g if there are more than one set of recursions that satisfy the conditions of Theorem 1.

Theorem 2. *If for* $v = 1, \ldots, w$ (5) *is satisfied with*

$$\varphi_{i_1\ldots i_h} = \varphi^{(v)}_{i_1\ldots i_h}, \quad (1 \le i_1 < \ \ldots\ < i_h \le m;\ h = 1, \ldots, m)$$

then (5) *is satisfied with*

$$\varphi_{i_1\ldots i_h}(y_1, \ldots, y_h;\ \mathbf{x}) = \sum_{v=1}^{w} c_v(\mathbf{x}) \varphi^{(v)}_{i_1\ldots i_h}(y_1, \ldots, y_h);\ \mathbf{x}),$$

$$(y_j = 1, \ldots, x_{i_j};\ j = 1, \ldots, h;\ 1 \le i_1 < \ \ldots\ < i_h \le m;\ h = 1, \ldots, m;\ \mathbf{x} > \mathbf{0})$$

where the weight functions c_v *are chosen such that* $\sum_{v=1}^{w} c_v(\mathbf{x}) = 1$ *for all* $\mathbf{x} > \mathbf{0}$.

Proof. By assumption we have

$$g(\mathbf{x}) = \sum_{h=1}^{m} \sum_{1 \le i_1 < \ldots < i_h \le m} \sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi^{(v)}_{i_1\ldots i_h}(y_1, \ldots, y_h;\ \mathbf{x}) g\left(\mathbf{x} - \sum_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}(y_j),$$

$$(\mathbf{x} > \mathbf{0};\ v = 1, \ldots, w)$$

and the theorem follows by multiplication by $c_v(\mathbf{x})$ and summation over v.

Q.E.D.

In Section 5 we shall consider an application of Theorem 2.
The condition (6) in Lemma 1 is satisfied by the pairs

$$\varphi(y_1, \ldots, y_h;\ \mathbf{x}) = \prod_{j=1}^{q} \frac{y_j}{x_{i_j}};\quad \psi(\mathbf{n}) = \frac{1}{\prod_{j=1}^{q} n_{i_j}}\quad (q = 0, 1, \ldots, h)$$

and consequently by

$$\varphi(y_1, \ldots, y_h;\ \mathbf{x}) = a + \sum_{q=1}^{h} \sum_{1 \le s_1 < \ldots < s_q \le h} b_{i_{s_1} \ldots i_{s_q}} \prod_{j=1}^{q} \frac{y_{s_j}}{x_{i_{s_j}}} \tag{7}$$

$$\psi(\mathbf{n}) = a + \sum_{q=1}^{h} \sum_{1 \le s_1 < \ldots < s_q \le h} \frac{b_{i_{s_1} \ldots i_{s_q}}}{\prod_{j=1}^{q} n_{i_{s_j}}}. \tag{8}$$

Like in the Vernic recursion the coefficients could depend on whether $x_i = n_i = 0$ for some i's; in particular this should be done to avoid division by zero. To give a general expression for (5) based on these functions would be notationally rather messy, and we shall therefore abstain from that and rather suggest that one develops the formulae in special cases.

In the univariate case, (7) and (8) reduce to

$$\varphi(y; x) = a + b\frac{y}{x};\quad \psi(n) = a + \frac{b}{n}\ .$$

From Theorem 3 in Sundt & Jewell (1981) follows that these are the only (ψ, φ)'s for which (6) is satisfied for every possible choice of severity distribution. The present author believes that also in the multivariate case (7) and (8) give the only (ψ, φ)'s that satisfy the condition (6) of Lemma 1 for every possible choice of severity distributions.

5. Examples

5A. The following model is discussed by Hesselager (1996) in the bivariate case. We assume that the distribution $p.$ of $N.$ satisfies the Panjer recursion

$$p.(n.) = \left(a + \frac{b}{n.}\right) p.(n. - 1),\quad (n. = 1, 2, \ldots)$$

and that the conditional distribution of $\mathbf{N}$ given that $N. = n.$ is the multinominal distribution

$$q(\mathbf{n}) = n.! \prod_{i=1}^{m} \frac{w_i^{n_i}}{n_i!}\ .$$

We have $q = q_1^{n.*}$ with

$$q_1(\mathbf{y}) = \begin{cases} w_i & (\mathbf{y} = \mathbf{e}_i;\ i = 1, 2, ...) \\ 0. & (\text{otherwise}) \end{cases} \tag{9}$$

Hence p is the compound distribution $p = \sum_{n.=0}^{\infty} p.(n.)q_1^{n.*}$ with univariate counting distribution $p.$ and multivariate severity distribution q_1. Such compound distributions are discussed by Sundt (1999). From this Theorem 1 follows that for $h = 1, ..., m$ and $\mathbf{n} > \mathbf{0}$ we have the recursion

$$n_h p(\mathbf{n}) = \sum_{\mathbf{0}<\mathbf{u}\leq\mathbf{n}} (an_h + bu_h)q_1(\mathbf{u})p(\mathbf{n} - \mathbf{u}),$$

and insertion of (9) gives

$$n_h p(\mathbf{n}) = bw_h p(\mathbf{n} - \mathbf{e}_h) + an_h \sum_{i=1}^{m} w_i p(\mathbf{n} - \mathbf{e}_i)\ .$$

When $\mathbf{n} \geq \mathbf{e}_h$, we can divide by n_h, and we then obtain

$$p(\mathbf{n}) = b\frac{w_h}{n_h}p(\mathbf{n} - \mathbf{e}_h) + a\sum_{i=1}^{m} w_i p(\mathbf{n} - \mathbf{e}_i) =$$

$$\left(a + \frac{b}{n_h}\right)w_h p(\mathbf{n} - \mathbf{e}_h) + a\sum_{i\neq h} w_i p(\mathbf{n} - \mathbf{e}_i)\ .$$

Hence

$$g(\mathbf{x}) = b\frac{w_h}{x_h}\sum_{y_h=1}^{x_h} y_h f_h(y_h)g(\mathbf{x} - y_h\mathbf{e}_h) + a\sum_{i=1}^{m} w_i \sum_{y_i=1}^{x_i} f_i(y_i)g(\mathbf{x} - y_i\mathbf{e}_i) =$$

$$w_h \sum_{y_h=1}^{x_h}\left(a + b\frac{y_h}{x_h}\right)f_h(y_h)g(\mathbf{x} - y_h\mathbf{e}_h) + a\sum_{i\neq h} w_i \sum_{y_i=1}^{x_i} f_i(y_i)g(\mathbf{x} - y_i\mathbf{e}_i)\ . \tag{10}$$

Formula (10) gives m recursions for g. We shall now combine these recursions by using Theorem 2. Multiplying (10) by $x_h/x.$ and summing over those values of h where $x_h > 0$, gives

$$g(\mathbf{x}) = \sum_{h=1}^{m} w_h \sum_{y_h=1}^{x_h}\left(a + b\frac{y.}{x.}\right)f_h(y_h)g(\mathbf{x} - y_h\mathbf{e}_h)\ . \quad (\mathbf{x} > \mathbf{0}) \tag{11}$$

Compared to (10), this recursion has the advantage that it holds for all $\mathbf{x} > \mathbf{0}$. On the other hand, as it involves more algebraic operations, it would presumably be more time-consuming.

As a special case of (11) we obtain

$$p(\mathbf{n}) = \left(a + \frac{b}{n.}\right) \sum_{h=1}^{m} w_h p(\mathbf{n} - \mathbf{e}_h) . \quad (\mathbf{n} > \mathbf{0})$$

This recursion was also given by Sundt (1999).

5B. Teicher (1954) discusses a class of multivariate Poisson distributions that satisfy the recursion

$$p(\mathbf{n}) = \frac{1}{n_m}\left(cp(\mathbf{n} - \mathbf{e}_m) + \sum_{h=1}^{m-1} \sum_{1 \leq i_1 < \ldots < i_h \leq m-1} d_{i_1 \ldots i_h} p(\mathbf{n} - \mathbf{e}_{i_1 \ldots i_h m})\right). \quad (\mathbf{n} > \mathbf{e}_m)$$

as well as analogous recursions where we divide by n_k instead of n_m; $k = 1, \ldots, m-1$. In the bivariate case the corresponding compound distributions are discussed by Hesselager (1996) and Vernic (1999).

6. Extensions

6A. In the univariate case Sundt (1992) gave the following extension to Panjer's (1981) recursion.

Theorem 3. *If p satisfies the recursion*

$$p(n) = \sum_{i=1}^{k} \left(a_i + \frac{b_i}{n}\right) p(n-i), \quad (n = 1, 2, \ldots)$$

then

$$g(x) = \sum_{y=1}^{x} g(x-y) \sum_{i=1}^{k} \left(a_i + \frac{b_i}{i}\frac{y}{x}\right) f^{i*}(y) . \quad (x = 1, 2, \ldots)$$

An analogous extension of the theory in Section 4 would mean to allow the recursion for p to go k steps back. In that connection we would need the following extension of Lemma 1.

Lemma 2. *If for different integers $i_1, \ldots, i_h \in \{1, \ldots, m\}$ and positive integers $k_1, \ldots, k_h$*

$$\mathrm{E}\left[\varphi\left(\sum_{j=1}^{k_1} Y_{i_1 j}, \ldots, \sum_{j=1}^{k_h} Y_{i_h j}; \mathbf{x}\right) \middle| \bigcap_{j=1}^{h} \left(\sum_{r=1}^{n_{i_j}} Y_{i_j r} = x_{i_j}\right)\right] = \psi(\mathbf{n}) \qquad (12)$$

for all $\mathbf{x}, \mathbf{n} > \mathbf{0}$ *such that* $\prod_{i=1}^{m} f_i^{n_i *}(x_i) > 0$, *then*

$$\sum_{\mathbf{n}>0} \psi(\mathbf{n}) p\left(\mathbf{n} - \sum\nolimits_{j=1}^{h} k_j \mathbf{e}_{i_j}\right) \prod_{i=1}^{m} f_i^{n_i *}(x_i) =$$
$$\sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, \ldots, y_h;\ \mathbf{x}) g\left(\mathbf{x} - \sum\nolimits_{j=1}^{h} y_j \mathbf{e}_{i_j}\right) \prod_{j=1}^{h} f_{i_j}^{k_j *}(y_j) . \quad (\mathbf{x} > \mathbf{0})$$

Theorem 1 can be extended analogously.

The condition (12) in Lemma 2 is in particular satisfied for

$$\varphi(y_1,\ \ldots,\ y_h;\ \mathbf{x}) = \prod_{j=1}^{q} \frac{y_j}{k_j x_{i_j}}; \quad \psi(\mathbf{n}) = \frac{1}{\prod_{j=1}^{q} n_{i_j}} . \quad (q = 0, 1, \ldots, h)$$

6B. Analogous to the extensions by Ambagaspitiya (1999) and Sundt (1999) of Panjer's recursion to Case 1 of Section 3, we could extend the results of the present paper to the case where the severity distributions are multivariate.

6C. In the present paper we have concentrated on recursions for multivariate distributions. In practice one will often approximate distributions by functions that are not necessarily distributions themselves, and thus it can be of interest to have recursions for more general functions. In the univariate case some recursions originally developed for distributions have been extended to more general functions by Dhaene & Sundt (1998) and Sundt, Dhaene & De Pril (1998); Dhaene, Willmot & Sundt (1999) discuss recursions for some classes of functions related to distributions, in particular cumulative distribution functions. Some multivariate extensions have been given in Sundt (1998). Analogously, the recursions of the present paper could be extended to more general functions. However, as the conditional expectation in (6) does not make sense if we leave the realm of distributions, we have to reformulate that formula. We rewrite it as

$$\sum_{s=1}^{h} \sum_{y_s=1}^{x_{i_s}} \varphi(y_1, \ldots, y_h;\ \mathbf{x}) \left(\prod_{j=1}^{h} f_{i_j}(y_j) f_{i_j}^{(n_{i_j}-1)*}(x_{i_j} - y_j) \right) \prod_{j=h+1}^{m} f_{i_j}^{n_{i_j}*}(x_{i_j}) =$$
$$\psi(\mathbf{n}) \prod_{i=1}^{m} f_i^{n_i *}(x_i) .$$

This is the relation that we need between ψ and φ in the general case, and as this is the relation applied in the proof of the lemma, the proof still holds in the general case. Analogous for Lemma 2.

REFERENCES

AMBAGASPITIYA, R.S. (1999). On the distributions of two classes of correlated aggregate claims. *Insurance: Mathematics and Economics* **24**, 301-308.

DHAENE, J. & SUNDT, B. (1998). On approximating distributions by approximating their De Pril transforms. *Scandinavian Actuarial Journal*, 1-23.

DHAENE, J., WILLMOT, G. & SUNDT, B. (1999). Recursions for distribution functions and their stop-loss transforms. *Scandinavian Actuarial Journal*, 52-65.

HESSELAGER, O. (1996). Recursions for certain bivariate counting distributions and their compound distributions. *ASTIN Bulletin* **26**, 35-52.

PANJER, H.H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* **12**, 22-26.

SUNDT, B. (1992). On some extensions of Panjer's class of counting distributions. *ASTIN Bulletin* **22**, 61-80.

SUNDT, B. (1998). The multivariate De Pril transform. To appear in *Insurance: Mathematics and Economics*.

SUNDT, B. (1999). On multivariate Panjer recursions. *ASTIN Bulletin* **29**, 29-45.

SUNDT, B., DHAENE, J. & DE PRIL, N. (1998). Some results on moments and cumulants. *Scandinavian Actuarial Journal*, 24-40.

SUNDT, B. & JEWELL, W.S. (1981). Further results on recursive evaluation of compound distributions. *ASTIN Bulletin* **12**, 27-39.

TEICHER, H. (1954). On the multivariate Poisson distribution. *Skandinavisk Aktuarietidskrift* **37**, 1-9.

VERNIC, R. (1999). Recursive evaluation of some bivariate compound distributions. *ASTIN Bulletin* **29**, 315-325.

BJØRN SUNDT
Vital Forsikring ASA
P.O. Box 250
N-1326 Lysaker
Norway

LONG-TERM RETURNS IN STOCHASTIC INTEREST RATE MODELS: APPLICATIONS

BY

GRISELDA DEELSTRA

ENSAE, CREST and VUB

ABSTRACT

We extend the Cox-Ingersoll-Ross (1985) model of the short interest rate by assuming a stochastic reversion level, which better reflects the time dependence caused by the cyclical nature of the economy or by expectations concerning the future impact of monetary policies. In this framework, we have studied the convergence of the long-term return by using the theory of generalised Bessel-square processes. We emphasize the applications of the convergence results. A limit theorem proves evidence of the use of a Brownian motion with drift instead of the integral $\int_0^t r_u du$. For practice, however, this approximation turns out to be only appropriate when there are no explicit formulae and calculations are very time-consuming.

KEYWORDS

Interest rates; Cox-Ingersoll-Ross model; Stochastic reversion level; Generalised Bessel-square processes; Convergence; Bond prices; Life insurance.

1. INTRODUCTION

In this paper, which has been presented at the 5th AFIR International Colloquium, we concentrate on the convergence of the long-term return $t^{-1} \int_0^t r_u du$, using a very general two-factor model, which is an extension of the Cox-Ingersoll-Ross (1985) model. Cox, Ingersoll & Ross (1985) express the short interest rate dynamics as

$$dr_t = \kappa(\gamma - r_t)dt + \sigma\sqrt{r_t}dB_t$$

with $(B_t)_{t\geq 0}$ a Brownian motion and κ, γ and σ positive constants. This model has some realistic properties. First, negative interest rates are precluded. Second, the absolute variance of the interest rate increases when the interest rate itself increases. Third, the interest rates are elastically pulled to the long-term value γ, where κ determines the speed of adjustment. Empirical studies like Chan, Karolyi, Longstaff & Sanders (1992) or Brown & Schaefer (1994), however, have shown that there is only weak evidence for the existence of a constant long run level of reversion.

We stress the long-term reversion level and the long-term interest rates since they are important in several issues in finance and insurance. For instance, for pricing an option to exchange a long bond for a short bond; or for mortgage pricing where the long rate determines when homeowners refinance their mortgages. In insurance, whole-life insurances are long-term products and the long-term interest rates play a dominant role.

We therefore follow the idea of Brennan & Schwartz (1982), who introduced a two-factor model by using short-term interest rates and consol rates (see Hogan (1993) for comments on this model).

In this paper, we assume that the short interest rate X is governed by the stochastic differential equation

$$dX_s = (2\beta X_s + \delta_s)ds + v\sqrt{X_s}dB_s$$

with the drift rate parameter $\beta < 0$, v a constant and δ a non-negative predictable stochastic process such that $\int_0^t \delta_u du < \infty$ a.e. for all $t \in \mathbb{R}^+$. This stochastic differential equation has a unique (non-negative) strong solution.

It should be noted that the stochastic process $(\delta_s)_{s\geq 0}$ determines a reversion level. If it is chosen to be a constant and if $v = 2$, the process $(X_s)_{s\geq 0}$ is a Bessel-square process with drift, a process which is studied in great detail by for example Pitman & Yor (1982) and Revuz & Yor (1991). As the model is a generalisation of Bessel-square processes with drift, it is fearly easy to treat.

In Section 2, we concentrate on the convergence almost everywhere of the long-term return $t^{-1}\int_0^t r_u du$. We are interested in this limit as $\left(\exp\left(\int_0^t r_u du\right)\right)^{1/t}$ is the average of the accumulating factor (also called return) which can be useful in the determination of models of participation in the benefit or of saving products with a guarenteed minimum return. Using the results of Deelstra & Delbaen (1995a), we found that in most existing interest rate models, $\left(\exp\left(\int_0^t r_u du\right)\right)^{1/t}$ converges almost everywhere to a constant independent of the current market, as the observing period tends to infinity. We then say that the model has the "strong convergence property" (SCP), whereas we refer to models with the "weak convergence property" when the returns converge to a constant, that will generally depend upon the current economic environment and that may change in a stochastic fashion over time. This terminology appeared in a preliminary version entitled "Do interest rates converge" (1986) of Dybvig, Ingersoll & Ross (1996).

Dybvig, Ingersoll & Ross (1996) proved that the assumption of no-arbitrage implies that the long forward rate and the asymptotic zero-coupon rate never fall and moreover, they show that nearly all models have the surprising implication that long run forward rates and zero coupon rates converge to a constant, which is independent of the current state of the economy. El Karoui, Frachot & Geman (1998) discuss the theoretical and practical consequences of this observation for existing models. They also focus on some issues encountered in empirical work which can be related to the behavior of the long-term yield structure of interest rates.

As noted by El Karoui, Frachot & Geman (1998) and Pearson & Sun (1994), parameter estimates are generally very unstable over time and this fact can be interpreted as an indicator of misspecification: the parameters have to capture the remaining uncertainty due to the stochastic long-term rates. As illustrated by Pearson & Sun (1994) and Chen & Scott (1992), the estimation of multi-factor versions with no stochastic long-term reversion level, show low mean-reversion for one of the state variables. El Karoui, Frachot & Geman (1998) argue that this low mean-reversion reflects the fact that the long-term yield is not constant over time.

Using the almost everywhere convergence theorem of Deelstra & Delbaen (1995a), we show that it is possible to build a model with the WCP in which the long-term return converges almost surely to a reversion level which is random itself. As an example we adapt the model of Tice & Webber (1997).

In Deelstra & Delbaen (1995b), we found conditions necessary to prove the convergence in law of a sequence of transformations of the long-term return to a Brownian motion. In Section 3, we propose a generalised theorem with measure-invariant hypotheses and we recall the idea of approximating $\int_0^t r_u du$ for t large enough. If the objective is to approximate the distribution of the long-term return of an investment made at time 0, it is appropriate to approximate $\int_0^t r_u du$ by a scaled Brownian motion with drift for t going to infinity. In the past, many authors have proposed Wiener models since in the long term, the Central Limit Theorems are applicable. In insurance, e.g. Beekman & Fuelling (1991), Dufresne (1990), Giacotto (1986), Goovaerts et al. (1994, 1995) and Milevsky (1997) modeled the accumulating factor $\exp\left(\int_0^t r_u du\right)$ by the exponential of a Brownian motion with drift for the derivation of prices of different insurance products like annuities and perpetuities.

For practical reasons, we are interested in an approximation of $\int_0^t r_u du$ for all values of t. Therefore we suggest an improved approximation, which is discussed and evaluated by looking at bond prices. The results show that one should be very careful by replacing the integral $\int_0^t r_u du$ by a Brownian motion with drift. This approximation should only be used if no exact formulae are available and the exact computations are very time consuming like could be the case in the derivation of annuities.

In Section 4, we turn to the pricing of n-year temporary life assurances, whole-life assurances and endowment assurances. We calculate the present value and the variance and skewness of this present value of the benefit under these contracts by using on one hand the Cox-Ingersoll-Ross (1985) model and on the other hand a Brownian motion with drift which is suggested by the Central Limit Theorem. The results show that in general, it is inappropriate to use the Brownian motion with drift instead of the Cox-Ingersoll-Ross (1985) model or its extensions.

Without further notice we assume that a probability space $\left(\Omega, (\mathcal{F}_t)_{0\leq t}, \mathbb{P}\right)$ is given and that the filtration $(\mathcal{F}_t)_{0\leq t}$ satisfies the usual assumptions with respect to $\mathbb{P}$, a fixed probability on the sigma-algebra $\mathcal{F}_\infty = \vee_{t\geq 0}\mathcal{F}_t$. Also B is a continuous process which is a Brownian motion with respect to $(\mathcal{F}_t)_{0\leq t}$.

2. Factor models with SCP and WCP

In this section, we show by using a theorem obtained in Deelstra & Delbaen (1995a) that it is easy to verify that existing generalisations of the Cox-Ingersoll-Ross model have the strong convergence property, which means that the long-term return converges to a constant, which is independent of the earlier shape of the term structure and of the current state of the economic environment. By looking at anologous convergence theorems in e.g. a Gaussian setting, we could as a matter of fact prove that most existing interest rate models have the SCP, but this will not be done within this paper.

Afterwards, we use the model of Tice and Webber (1997) to show that multifactor models do not necessarily imply that the strong convergence property holds.

It should be noted that the almost everywhere convergence limit of $t^{-1}\int_0^t r_u du$ is interesting to study since economists and actuaries work with the multiplicative accumulating factor (return) over t years, namely $\exp\left(\int_0^t r_u du\right)$. The average return in one year, where the average is taken over t years, is denoted by $\left(\exp\left(\int_0^t r_u du\right)\right)^{1/t}$. If the observing period goes to infinity, it converges to the exponent of the almost everywhere limit of $t^{-1}\int_0^t r_u du$.

We recall from Deelstra & Delbaen (1995a) that if X is defined by

$$dX_s = (2\beta X_s + \delta_s)ds + v\sqrt{X_s}dB_s$$

with $(B_s)_{s\geq 0}$ a Brownian motion, $\beta < 0$, v a constant and δ a positive, predictable stochastic process such that $s^{-1}\int_0^s \delta_u du \xrightarrow{a.e.} \bar{\delta}$ with $\bar{\delta} : \Omega \to \mathbb{R}^+$, then the following convergence almost everywhere holds:

$$\frac{1}{s}\int_0^s X_u du \xrightarrow{a.e.} \frac{-\bar{\delta}}{2\beta}.$$

It is easy to show that for $r_t = \sigma^2 X_t/4$, $v = 2$, $\beta = -\kappa/2$ and $\delta_t = 4\kappa\gamma_t/\sigma^2$, we obtain a generalised two-factor Cox-Ingersoll-Ross (1985) model

$$dr_t = \kappa(\gamma_t - r_t)dt + \sigma\sqrt{r_t}dB_t$$

with $(\gamma_s)_{s\geq 0}$ a positive stochastic reversion level process. To ensure that the interest rate process $(r_t)_t$ remains a.s. strictly positive, we should add some hypotheses. Comparison theorems for Bessel-square processes with stochastic reversion level (see Deelstra (1995)) can be used to obtain some. Indeed, if $X^{(1)}$ and $X^{(2)}$ are two Bessel-square processes with respectively stochastic reversion level $\delta^{(1)}$, $\delta^{(2)}$ and issued from $x^{(1)}$, $x^{(2)}$ with $x^{(2)} \geq x^{(1)}$ and $\delta^{(2)} \geq \delta^{(1)}$ a.s. for all $t \in \mathbb{R}^+$, then

$$\mathbb{P}\left[X_t^{(2)} \geq X_t^{(1)} \text{ for all } t \geq 0\right] = 1.$$

Now, it is well-known that if $X^{(1)}$ is a Bessel-square process with constant dimension $\delta^{(1)} \geq 2$, then $X_t^{(1)} > 0$ a.s. Therefore, hypotheses like $4\kappa\gamma_t/\sigma^2 \geq 2$ a.s. for all $t \in \mathbb{R}^+$, imply the strict positivity of $(r_t)_t$ a.s. Remark that this is the generalisation of the constraint in case of the Cox-Ingersoll-Ross model.

In this paper, we further choose the process $(\gamma_s)_{s\geq 0}$ such that $t^{-1}\int_0^t \gamma_s ds$ converges almost everywhere to a random variable $\gamma^* = \sigma^2\bar{\delta}/4\kappa : \Omega \to \mathbb{R}^+$. The central tendency process $(\gamma_s)_{s\geq 0}$ may be dependent or independent of the short interest rate process.

We stress this fact since if the reversion level process $(\gamma_s)_{s\geq 0}$ is independent of the short-term interest rates it is possible to derive (quasi-)explicit formulae for bond prices by using scaling properties of Bessel-square processes. This approach has been used in the papers by e.g. Maghsoodi (1996), Delbaen & Shirakawa (1996) and Deelstra (2000), who consider time-dependent but deterministic $(\gamma_s)_{s\geq 0}$. However if the reversion level process $(\gamma_s)_{s\geq 0}$ is dependent on the short interest rate process, no such formulae can be obtained.

As an example, let us describe the stochastic reversion level process $(\gamma_s)_{s\geq 0}$ by a Cox-Ingersoll-Ross (1985) square root process

$$d\gamma_t = \tilde{\kappa}(\gamma^* - \gamma_t)dt + \tilde{\sigma}\sqrt{\gamma_t}d\tilde{B}_t$$

or by a Courtadon (1982) process

$$d\gamma_t = \tilde{\kappa}(\gamma^* - \gamma_t)dt + \tilde{\sigma}\gamma_t d\tilde{B}_t \quad \text{with} \quad \tilde{\sigma}^2 \leq 2\tilde{\kappa}$$

with $(\tilde{B}_s)_{s\geq 0}$ a Brownian motion and with $\tilde{\kappa}$, γ^* and $\tilde{\sigma}$ positive constants. The Brownian motion $(\tilde{B}_s)_{s\geq 0}$ may be correlated with the Brownian motion $(B_s)_{s\geq 0}$ of the short rate process and this correlation may be in a random way. As mentioned above, we do not need the technical assumption of fixed correlation or independence between the two factors of the model: for example, as in Brennan & Schwartz (1982).

The two proposed reversion level processes are from the same family. They both remain positive for $\tilde{\kappa}, \gamma^* \geq 0$, a property which is necessary if one wants to work with nominal interest rates. For $\tilde{\kappa}, \gamma^* > 0$, these processes are mean-reverting to the long-term constant value γ^*, where $\tilde{\kappa}$ represents the speed of adjustment. The volatility increases in both cases with the reversion level.

For this class of stochastic reversion levels, $t^{-1}\int_0^t \gamma_s ds \xrightarrow{a.e.} \gamma^*$ and since $\delta_s = 4\kappa\gamma_s/\sigma^2$, $t^{-1}\int_0^t \delta_s ds \xrightarrow{a.e.} 4\kappa\gamma_s/\sigma^2$. By the theorem mentioned above (see Deelstra & Delbaen (1995a)), the long-term return is shown to converge almost everywhere to a constant:

$$\frac{1}{t}\int_0^t r_s ds = \frac{1}{t}\int_0^t \frac{\sigma^2}{4} X_s ds \xrightarrow{a.e.} \gamma^*.$$

We conclude that the long-term return in these two-factors model of short interest rates satisfies the strong convergence property. The average accumulating factor, where the average is taken over a period t, is found to converge almost everywhere to a constant as the period t tends to infinity, and this constant is independent of the current state of the economy:

$$\left(e^{\int_0^t r_u du}\right)^{1/t} \xrightarrow{a.e.} e^{\gamma^*}.$$

As another example, we treat the two-factor model proposed by Cox, Ingersoll & Ross (1985). They assumed a stochastic reversion level process depending on Y, the state variable which describes the change in the production opportunities, namely

$$\begin{aligned}
dr_t &= \kappa(\gamma_t - r_t)dt + \sigma\sqrt{r_t}dB_t \\
d\gamma_t &= \tilde{\kappa}(Y_t - \gamma_t)dt \\
dY_t &= -\xi\left(\frac{-\zeta}{\xi} - Y_t\right)dt + \tilde{\sigma}\sqrt{Y_t}dB'_t,
\end{aligned}$$

with κ, $\tilde{\kappa}$, σ, ζ and $\tilde{\sigma}$ strictly positive constants. We assume that ξ is a strictly negative constant. We here only theoretically show that this model also has the SCP for the long-term return: since $t^{-1}\int_0^t Y_s ds \xrightarrow{a.e.} -\zeta/\xi$, we have that $t^{-1}\int_0^t \gamma_s ds \xrightarrow{a.e.} -\zeta/\xi$ and by the same reasoning as above, we obtain

$$\frac{1}{t}\int_0^1 r_s ds \xrightarrow{a.e.} \frac{-\zeta}{\xi}.$$

As a consequence of the convergence of the long-term return to a constant, we can conclude that the long-term yield $\lim_{T\to\infty} Y(t,T)$ is uniformly bounded above as by Jensen's inequality (see also Yao (1998))

$$Y(t,T) = \mathbb{E}\left[\exp\left(\frac{1}{T-t}\int_t^T r_u du\right)\right] \leq \exp\left(\frac{1}{T-t}\int_t^T \mathbb{E}[r_u]du\right).$$

It is not surprising that the previous examples satisfy the SCP since in each model, the reversion level process itself is elastically pulled to a constant independent of the economic state. We recall that the convergence theorem from Deelstra & Delbaen (1995a) has no such strong hypothesis; on the contrary, the assumptions are very general. For example, the reversion level process does not have to be continuous. The convergence theorem only assumes a positive, predictable reversion $(\delta_u)_{u\geq 0}$ such that $s^{-1}\int_0^s \delta_u du \xrightarrow{a.e.} \bar{\delta}$, where $\bar{\delta}$ may be a random variable. Models in which this $\bar{\delta}$ really is a random variable, would imply that the long-term return converges to a random variable which will generally depend on the economic environment.

As an example, let us look at the general dynamic mean interest rate model in Tice & Webber (1997)

$$\begin{aligned} dr &= a(\gamma - r)dt + \sigma_r dz_r \\ d\gamma &= b\big(\mu_\gamma(t,r,Y) - \gamma\big)dt + \sigma_\gamma dz_\gamma \\ dY &= c(\mu_Y(t,r,\gamma,Y) - Y)dt + \sigma_Y dz_Y \end{aligned}$$

where z_r, z_γ and z_Y denote Brownian motions, r is the short rate and γ the level to which the short rates revert. Y is assumed to be a vector process summarizing the remainder of the dynamics in the model. Tice & Webber (1997) have interpreted this model within the IS-LM framework, which is a standard model in macroeconomics (see e.g. Hicks (1937)). As a particular case, Tice & Webber (1997) study a three factor model with the third factor related to the availability of transactions credit within the economy. To simplify the notations, Tice & Webber (1997) restrict themselves to σ_r, σ_γ and σ_Y being constant but it is possible to consider e.g. $\sigma_r = \sigma\sqrt{r}$.

In that case, it is clear that we are dealing with an extension of the Cox-Ingersoll-Ross model with a stochastic reversion level. This model has the weak convergence property if the process is not recurrent.

3. Approximation of the Long-Term Return and of Bond Prices

In this section, we give a generalised version of the Central Limit Theorem from Deelstra & Delbaen (1995b). We study the convergence in law since it is always useful to know how the long-term return is distributed in the limit so that approximations can be deduced. We are particularly interested in an approximation of $\int_0^t r_u du$ since this term appears in discounting factors, bond prices, annuities, perpetuities, etc. As a natural candidate appears a Brownian motion with drift. This process has been used in insurance before for modeling the integral $\int_0^t r_u du$, e.g. in Beekman & Fuelling (1991), Dufresne (1990), Giacotto (1986), Goovaerts et al. (1994, 1995) and Milevsky (1997). In order to evaluate this approximation, we compare in the settings of the Cox-Ingersoll-Ross model bond prices calculated by using the approximating Brownian motion with exact values.

In order to obtain convergence in law, we have to make some more assumptions about our family of processes:

Theorem: Suppose that a probability space $\left(\Omega, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}\right)$ is given and that a stochastic process $X : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$ is defined by the stochastic differential equation

$$dX_s = (2\beta X_s + \delta_s)ds + v\sqrt{X_s}dB_s \quad \forall s \in \mathbb{R}^+$$

with $(B_s)_{s\geq 0}$ a Brownian motion with respect to $(\mathcal{F}_t)_{t\geq 0}$, v a constant and $\beta < 0$.
Let us make the following assumptions about the adapted and measurable process $\delta : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$:

- $\frac{1}{s}\int_0^s \delta_u du \xrightarrow{a.e.} \bar{\delta}$ where $\bar{\delta}$ is a strictly positive real number;
- $\sup_{t\geq 1} \frac{1}{t}\int_0^t \delta_u^2 du < \infty$ a.e.;
- For all $a \in \mathbb{R}^+$ $\frac{1}{t}\int_{t-a}^t \delta_u^2 du \xrightarrow{\mathbb{P}} 0$.

Under these conditions, the following convergence in distribution holds:

$$\left(\sqrt{\frac{-8\beta^3}{v^2\bar{\delta}n}}\int_0^{nt}\left(X_u + \frac{\delta_u}{2\beta}\right)du\right)_{t\geq 0} \xrightarrow{\mathcal{L}} (B_t)_{t\geq 0}$$

where $(B_t)_{t\geq 0}$ denotes a Brownian motion and where '$\xrightarrow{\mathcal{L}}$' denotes convergence in law.

Since the proof of this theorem follows more or less the lines of the result in Deelstra & Delbaen (1995b), the proof is omitted and we immediately turn to the applications.

Inspired by this theorem, we estimate $\int_0^t X_u du$ with X as in the settings of the theorem by

$$\int_0^t \frac{-\delta_u}{2\beta}du + \sqrt{\frac{-v^2\bar{\delta}}{8\beta^3}}B_t$$

for t large enough. In Deelstra & Delbaen (1995b), we used the hypothesis $t^{-1}\int_0^t \delta_u du \xrightarrow{a.e.} \bar{\delta}$, to approximate $\int_0^t X_u du$ by the sum of the long-term constant $-\bar{\delta}/2\beta$, to which the long-term return a.e. converges, multiplied by t and a scaled Brownian motion:

$$\int_0^t X_u du \quad \text{by} \quad \frac{-\bar{\delta}}{2\beta}t + \sqrt{\frac{-v^2\bar{\delta}}{8\beta^3}}B_t. \tag{1}$$

It should be noted that in the case that $(\delta_u)_u$ is a stochastic process we replace the stochastic term $(-2\beta)^{-1}\int_0^t \delta_u du$ by a constant times t.

Another drawback of this estimator is that the moments of $\int_0^t X_u du$ do not equal those of the estimator, although they are the same asymptotically. If the period observed is large enough, this is satisfactory. If the objective is to approximate the distribution of the long-term return of an investment made at time 0, it seems to be appropriate to approximate $\int_0^t X_u du$ by a scaled Brownian motion with drift since the Central Limit Theorems are applicable on long-term.

However, one of our objectives is to look at the approximation

$$\int_0^t X_u du \quad \text{by} \quad \frac{-\bar{\delta}}{2\beta}t + \sqrt{\frac{-v^2\bar{\delta}}{8\beta^3}}B_t$$

to find estimations of bond prices for all maturities. Therefore, the moments of $\int_0^t X_u du$ and of the estimator should be equal for all t. A second drawback of the approximation immediately appears in the bond price, namely

$$P(0,t) = \mathbb{E}_{X_0}\left[e^{-\int_0^t X_u du}\right] \sim \exp\left(\frac{\bar{\delta}}{2\beta}t - \frac{v^2\bar{\delta}}{16\beta^3}t\right).$$

It is not realistic that the estimating bond price is independent of the current short interest rate X_0. Remark that we work with the default-free bond prices. In the sequel, we omit without notice the adjective "default-free". We further assume that there is no market price of risk, since we only want to compare different approximations theoretically.

In case of the Cox-Ingersoll-Ross (1985) square root process, the approximating bond price equals:

$$\mathbb{E}_{r_0}\left[e^{-\int_0^t r_u du}\right] \sim \exp\left(\gamma t\left(\frac{\sigma^2}{2\kappa^2} - 1\right)\right).$$

This estimating bond price is a decreasing function of the speed of adjustment parameter κ, where in case of the Cox-Ingersoll-Ross (1985) model, two cases are distinguished: for $r_0 < \gamma$, the bond price is a decreasing function of the parameter κ, and for $r_0 > \gamma$, it is an increasing function of κ. In Deelstra & Delbaen (1995b), we compared these approximating bond prices with values obtained in the Cox-Ingersoll-Ross setting and found that there is an underestimation of bond prices if $r_0 < \gamma$ and an overestimation if $r_0 > \gamma$.

Trying to motivate the approximation of the integral of the short-term interest rates by a Brownian motion with drift, we searched for an improved approximation. It seems logical to propose the approximation

$$\int_0^t X_u du \sim \int_0^t \mathbb{E}[X_u]du + \sqrt{\frac{-\upsilon^2\bar{\delta}}{8\beta^3}}B_t.$$

Then the expectation is equal for all t and the variance is still asymptotically equal.

Since $(X_u)_{u\geq 0}$ is defined by the stochastic differential equation

$$dX_s = (2\beta X_s + \delta_s)ds + \upsilon\sqrt{X_s}dB_s,$$

the expectation value of X_s equals:

$$\mathbb{E}[X_s] = e^{2\beta s}X_0 + e^{2\beta s}\int_0^s e^{-2\beta u}\mathbb{E}[\delta_u]du,$$

which can only be calculated if $\mathbb{E}[\delta_u]$ is known and $\int_0^s \mathbb{E}[\delta_u]du < \infty$. As above, it should be noted that in the case of $(\delta_u)_u$ being a stochastic process, we replace the stochastic term $(2\beta)^{-1}\int_0^t \delta_u du$ by a deterministic time-dependent term. But at least in this way, the current state X_0 is introduced in the approximation.

As an example of the approximation, let us look again at the Cox-Ingersoll-Ross (1985) two-factor model:

$$dr_t = \kappa(\gamma_t - r_t)dt + \sigma\sqrt{r_t}dB_t,$$
$$d\gamma_t = \tilde{\kappa}(\gamma^* - \gamma_t)dt + \tilde{\sigma}\sqrt{\gamma_t}d\tilde{B}_t.$$

The approximation becomes

$$\begin{aligned}\int_0^t r_u du &\sim \int_0^t \mathbb{E}[r_u]du + \sqrt{\frac{\sigma^2\gamma^*}{\kappa^2}}B_t \\ &\sim \gamma^* t + \frac{1-e^{-\kappa t}}{\kappa}\left(r_0 - \gamma^* - \frac{\gamma_0 - \gamma^*}{\kappa - \tilde{\kappa}}\kappa\right) \\ &+ \frac{1-e^{-\tilde{\kappa}t}}{\tilde{\kappa}}\left(\frac{\gamma_0 - \gamma^*}{\kappa - \tilde{\kappa}}\right)\kappa + \sqrt{\frac{\sigma^2\gamma^*}{\kappa^2}}B_t.\end{aligned}$$

The bond price is estimated by:

$$\mathbb{E}_{r_0}\left[e^{-\int_0^t r_u du}\right] \sim \exp\left(\gamma^* t\left(\frac{\sigma^2}{2\kappa^2} - 1\right)\right) \cdot$$
$$\exp\left(-\frac{1-e^{-\kappa t}}{\kappa}\left(r_0 - \gamma^* - \frac{\gamma_0 - \gamma^*}{\kappa - \tilde{\kappa}}\kappa\right) - \frac{1-e^{-\tilde{\kappa}t}}{\tilde{\kappa}}\left(\frac{\gamma_0 - \gamma^*}{\kappa - \tilde{\kappa}}\right)\kappa\right).$$

Let us evaluate the approximation in case of the Cox-Ingersoll-Ross (1985) single-factor model:

$$dr_t = \kappa(\gamma - r_t)dt + \sigma\sqrt{r_t}dB_t.$$

An anonymous referee remarked (see Deelstra & Delbaen (1995b)) that in this case, the moments of the first proposal (1) are equal for all t, as soon as the current short interest rate r_0 is distributed according to the steady state distribution of the square root process, namely the gamma-function with parameters $\alpha = 2\kappa\gamma/\sigma^2$ and $\beta = 2\kappa/\sigma^2$:

$$\mathbb{E}_{r_0}\left[\int_0^t r_u du\right] = \gamma t = \mathbb{E}_0\left[\gamma t + \sqrt{\frac{\sigma^2}{2\kappa^2}}B_t\right].$$

In reality, r_0 is not distributed this way, so an improvement is also necessary here to obtain good estimations of bond prices:

$$\int_0^t r_u du \sim \int_0^t \mathbb{E}[r_u]du + \sqrt{\frac{\sigma^2\gamma}{\kappa^2}}B_t.$$

Substituting the mean of the short interest rate, gives the expression

$$\int_0^t r_u du \sim \gamma t + \frac{1 - e^{-\kappa t}}{\kappa}(r_0 - \gamma) + \sqrt{\frac{\sigma^2\gamma}{\kappa^2}}B_t$$

and the estimating bond price is found to be

$$\mathbb{E}_{r_0}\left[e^{-\int_0^t r_u du}\right] \sim \exp\left(\gamma t\left(\frac{\sigma^2}{2\kappa^2} - 1\right) - \frac{1 - e^{-\kappa t}}{\kappa}(r_0 - \gamma)\right).$$

In case of the previous approximation (1), we found for $r_0 < \gamma$ an underestimation of the bond prices. The approximation in this paper is larger since for $r_0 < \gamma$, a positive term is added to the exponent, namely $-\frac{1 - e^{-\kappa t}}{\kappa}(r_0 - \gamma)$. In the same way, the underestimation in case of $r_0 > \gamma$ is reduced.

For the Cox-Ingersoll-Ross (1985) square root process, an explicit formula for the bond price is given by Pitman & Yor (1982) and Cox, Ingersoll & Ross (1985). We recall the bond price from Pitman & Yor (1982):

$$\mathbb{E}_{r_0}\left[\exp\left(-\int_0^t r_u du\right)\right] = \frac{\exp\left\{-\frac{r_0}{\sigma^2}w\frac{1 + \kappa/w\coth(wt/2)}{\coth(wt/2) + \kappa/w}\right\}e^{\kappa x/\sigma^2}e^{\kappa^2\gamma^* t/\sigma^2}}{(\cosh(wt/2) + \kappa/w\sinh(wt/2))^{\frac{2\kappa\gamma^*}{\sigma^2}}}$$

with $w = \sqrt{\kappa^2 + 2\sigma^2}$.

Using various values for the parameters, we have calculated this exact bond price and the improved approximation, for a large range of maturities. The deviations are always very small. The largest absolute deviations appear when the bond price has a value about 0.5. The reason therefore is that the bond price is a decreasing convex function of maturity and that the endpoints are fixed, namely for $t = 0$, the bond price equals 1, and for $t = \infty$, the bond price converges to 0. Consequently, the largest deviations are to be expected around one half.

In Table 1, the exact bond prices and the estimating bond prices are calculated with the parameters estimated by Chan, Karolyi, Longstaff & Sanders (1992), namely $\kappa = 0.23394$, $\gamma = 0.0808$ and $\sigma = 0.854$. The results are given for $r_0 = 0.04$ and for $r_0 = 0.1$. We present the maturities between 6 and 10 since then, the bond price is approximately 0.5 and the largest absolute deviations appear. Although the absolute error as presented in Table 1 is not a monotonic function, one should note that the error in the rate $-\ln P(0,t)/t$ does reduce for large values of t.

In comparison with the first approximation (1), the underestimation and overestimation are reduced but the difference between the exact result and the approximation remains too large to be useful in practice. This approximation should only be used if no exact formulae are available and the exact computations are very time-consuming like could be the case in the derivation of annuities.

TABLE 1

BOND PRICES EXACT VALUES AND APPROXIMATIONS.

t	$r_0 = 0.04$			$r_0 = 0.1$		
	Exact	*Approx*	*Approx-Exact*	*Exact*	*Approx*	*Approx-Exact*
1	.9565	.9617	.0051	.9068	.9116	.0048
6	.7061	.7254	.0192	.5843	.5978	.0134
7	.6587	.6788	.0200	.5386	.5521	.0134
8	.6135	.6339	.0204	.4970	.5102	.0131
9	.5708	.5912	.0204	.4591	.4719	.0127
10	.5305	.5507	.0201	.4244	.4367	.0122
20	.2503	.2630	.0127	.1968	.2040	.0071
30	.1171	.1239	.0068	.0919	.0959	.0039
40	.0547	.0582	.0035	.0430	.0451	.0021

4 Applications in life assurance

In this section, we follow the lines of Parker (1993, 1994) for deriving the net single premium and the variance and the skewness of the present value of the benefit payable under some insurance contracts. If the short-term interest rates are determined by a Cox-Ingersoll-Ross model, the exact formulae follow from the result of Pitman & Yor (1982). We compare these values with the approximation derived in Section 3.

Following the notation of Parker (1992), we denote by K the integer-valued discrete random variable which represents the number of completed years to be lived by a life assured, whose age is exactly x year at the issue of the contract. We let Z be the present value of the benefit payable under a given assurance contract. As the precise definition of Z depends on the specific assurance under consideration, we look at some examples: the n-year temporary assurance, the whole-life assurance and the endowment assurance (see e.g. Bowers et al. (1986)).

Under the n-year temporary assurance, the benefit of 1 is payable at the end of the year of death of a life assured, if the death occurs within n years from the date of issue. Thus Z is defined to be:

$$Z = \begin{cases} \exp\left(-\int_0^{K+1} X_u du\right) & K = 0, 1, \ldots, n-1 \\ 0 & K = n, n+1, \ldots \end{cases}$$

where $(X_u)_{u \geq 0}$ denotes as before the short interest rate, defined by the stochastic differential equation

$$dX_t = (2\beta X_t + \delta_t)dt + v\sqrt{X_t}d\tilde{B}_t.$$

The m-th non-centered moment of Z is given by

$$\mathbb{E}[Z^m] = \sum_{k=0}^{n-1} \mathbb{E}\left[\exp\left(-m\int_0^{k+1} X_u du\right)\right] {}_{k|}q_x,$$

where ${}_{k|}q_x$ denotes the probability that the life assured dies between his $(x+k)$-th and his $(x+k+1)$-th birthday.

Remark that for a whole-life assurance, the benefit certainly will be paid once, namely at the end of the year of death. Consequently,

$$Z = \exp\left(-\int_0^{K+1} X_u du\right) \qquad K = 0, 1, \ldots, \omega - x - 1,$$

where ω is the least age so that $l_x = 0$. The m-th non-centered moment is given by:

$$\mathbb{E}[\mathcal{Z}^m] = \sum_{k=0}^{\omega-x-1} \mathbb{E}\left[\exp\left(-m\int_0^{k+1} X_u du\right)\right]\ {}_{k|}q_x.$$

Under the endowment assurance contract, the benefit is payable at the end of the year of death if death occurs within n years of the issue date or, if the insured person survives n years, the benefit is payable at time n. Consequently, the present value $\mathcal{Z}$ of an endowment assurance is defined as:

$$\mathcal{Z} = \begin{cases} \exp\left(-\int_0^{K+1} X_u du\right) & K = 0,\ 1,\ \ldots,\ n-1 \\ \exp\left(-\int_0^n X_u du\right) & K = n,\ n+1,\ \ldots \end{cases}$$

The m-th non-centered moment of the present value is given by:

$$\mathbb{E}[\mathcal{Z}^m] = \sum_{k=0}^{n-1} \mathbb{E}\left[\exp\left(-m\int_0^{k+1} X_u du\right)\right]\ {}_{k|}q_x + \mathbb{E}\left[\exp\left(-m\int_0^{n} X_u du\right)\right]\ {}_{n}p_x.$$

Approximations of the net single premium of each contract are easily calculated. Indeed, approximations of the expected value of $\mathcal{Z}$ are obtained by taking $m = 1$ and by substituting the estimating bond price, proposed in the previous section.

We have evaluated this approximation in case of the Cox-Ingersoll-Ross single factor model, with the parameters estimated within Chan, Karolyi, Longstaff & Sanders (1992) and with $r_0 = 0.07$. We used the mortality table HD (1968-72), which is commonly used in Belgium and which is based in Makeham's formula $l_x = ks^x g^{c^x}$ with for the ages between 0 and 69: $k = 1{,}000{,}268$, $s = 0.999147835528$, $g = 0.999731696667$ and $c = 1.115094352734$; and otherwise $k = 1{,}292{,}726$, $g = 0.995564574228$, $c = 1.077130677635$ and the same value of s.

In Table 2, the exact values and the approximations are given for the net single premiums of n-year temporary life assurances and endowment contracts. Remark that for n larger than 60 years, both assurances become whole-life assurances since the life assured is aged $x = 30$ at the date of issue. We conclude that the approximations of the single net premiums are not encouraging.

The variance and the skewness of $\mathcal{Z}$ also are easy to find since the variance is defined as

$$\text{var}[\mathcal{Z}] = \mathbb{E}[\mathcal{Z}^2] - \mathbb{E}[\mathcal{Z}]^2,$$

and the skewness is defined as

$$\begin{aligned} \text{sk}[\mathcal{Z}] &= \frac{\mathbb{E}\left[(\mathcal{Z} - \mathbb{E}[\mathcal{Z}])^3\right]}{\text{var}[\mathcal{Z}]^{3/2}} \\ &= \frac{\mathbb{E}[\mathcal{Z}^3] - 3\mathbb{E}[\mathcal{Z}^2]\mathbb{E}[\mathcal{Z}] + 2\mathbb{E}[\mathcal{Z}]^3}{\text{var}[\mathcal{Z}]^{3/2}} . \end{aligned}$$

Each of these terms can be calculated by substituting $m = 1, 2$ or 3 in $\mathbb{E}[\mathcal{Z}^m]$ and by using the approximation of the m-th non-centered moment of the discounting factor, namely

$$\mathbb{E}\left[\exp\left(-m \int_0^t X_u du\right)\right] \sim \exp\left(-m \int_0^t \mathbb{E}[X_u] du - \frac{m^2 \bar{\delta} v}{16\beta^3} t\right) .$$

TABLE 2

NET SINGLE PREMIUMS: EXACT VALUES AND APPROXIMATIONS

n	*life assurance*			*endowment assurance*		
	Exact	*Approx*	*Approx-Exact*	*Exact*	*Approx*	*Approx-Exact*
1	.00154	.00155	.000008	.9313	.9363	.0049
10	.01453	.01484	.000314	.4785	.4944	.0158
20	.02896	.02985	.000887	.2354	.2453	.0098
40	.06222	.06479	.002572	.0894	.0935	.0041
60	.07635	.07979	.003439	.0767	.0801	.0034
80	.07664	.08010	.003459	.0766	.0801	.0034

In Tables 3 and 4, the variance and the skewness of $\mathcal{Z}$ are calculated, for $\mathcal{Z}$ being the present value of the benefit under an n-year temporary life-assurance, an endowment assurance and a whole-life assurance (if n is very large). Again, we used the formula of Makeham and the Cox-Ingersoll-Ross (1985) model with the same parameters as above. These results seem to be an indicator that the appproximation by a Brownian motion with drift can only be used in practice when there are no explicit formulae or when the calculation is very time-consuming.

We further admit that the major problem of taking into account stochastic interest rates in long-term life insurance products, is that the policies become dependent. With regard to the problems of setting contingency reserves and assessing the solvency of life assurance companies, it is therefore interesting to study portfolios of assurance policies (see e.g. Parker (1992, 1997)).

TABLE 3

THE VARIANCES: EXACT VALUES AND APPROXIMATIONS.

n	*life assurance*			*endowment assurance*		
	Exact	*Approx*	*Exact-Approx*	*Exact*	*Approx*	*Exact-Approx*
1	.00153	.00147	.00006	.00949	.05587	.04638
10	.01182	.01071	.00111	.02844	.07711	.04867
20	.01763	.01587	.00175	.01849	.02994	.01148
40	.02021	.01796	.00225	.01567	.01833	.00266
60	.01902	.01658	.00243	.01653	.01897	.00244
80	.01898	.01654	.00244	.01654	.01898	.00244

TABLE 4

THE SKEWNESS: EXACT VALUES AND APPROXIMATIONS

n	*life assurance*			*endowment assurance*		
	Exact	*Approx*	*Exact-Approx*	*Exact*	*Approx*	*Exact-Approx*
1	25.245	24.866	0.379	−3.685	0.313	−3.998
10	7.540	7.421	0.119	−0.957	1.048	−2.005
20	5.019	5.002	0.017	0.433	2.046	−1.602
40	3.655	3.719	−0.064	3.668	3.961	−0.293
60	3.724	3.891	−0.166	3.733	3.904	−0.171
80	3.731	3.902	−0.170	3.731	3.902	−0.170

ACKNOWLEDGEMENTS

The author wants to thank G. Parker for his suggestions and for giving a copy of his Ph.D. thesis and his articles. The author is also grateful to F. Delbaen for his interesting comments. P. Dybvig is acknowledged for sending his working papers. An anonymous referee of Deelstra (2000) is acknowledged for pointing out the use of comparison theorems to ensure the strict positivity of the generalised Cox-Ingersoll-Ross model.

REFERENCES

BEEKMAN, J. and FUELLING, C. (1991) Extra randomness in certain annuity modes. *Insurance: Mathematics and Economics* **10**, 275-287.

BOWERS, N.L., GERBER, H.U., HICKMAN, J.C., JONES, D.A. and NESBITT, C.J. (1986) *Actuarial Mathematics*. Society of Actuaries, Schaumburg.

BRENNAN, M.J. and SCHWARTZ, E.S. (1982) An equilibrium Model of Bond Pricing and a Test of Market Efficiency. *Journal of Financial and Quantitative Analysis* **17 (3)**, 301-329.

BROWN, R.H. and SCHAEFER, S.M. (1994) The term structure of real interest rates and the Cox, Ingersoll and Ross model. *Journal of Financial Economics* **35**, 3-42.

CHAN, K.C., KAROLYI, G.A., LONGSTAFF, F.A. and SANDERS, A.B. (1992) An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *Journal of Finance* **47**, 1209-1227.

CHEN, R. and SCOTT, L. (1992) Pricing interest rate options in a two-factor Cox-Ingersoll-Ross model of the term structure. *Review of Financial Studies* **5**, 613-636.

COX, J.C., INGERSOLL, J.E. and ROSS, S.A. (1985) A theory of the term structure of interest rates. *Econometrica* **58**, 385-407.

COURTADON, G. The Pricing of Options on Default-Free Bonds. *Financial and Quantitative Analysis* **17**, 75-100.

DEELSTRA, G. (1995) Long-term returns in stochastic interest rate models. Ph.D. thesis, *VUB Brussels*.

DEELSTRA, G. (2000) Yield Options in the Generalised Cox-Ingersoll-Ross Model. *In revision*.

DEELSTRA, G. and DELBAEN, F. (1995a) Long-term returns in stochastic interest rate models. *Insurance: Mathematics and Economics* **17**, 163-169.

DEELSTRA, G. and DELBAEN, F. (1995b) Long-term returns in stochastic interest rate models: Convergence in law. *Stochastics and Stochastics Reports* **55**, 253-277.

DELBAEN, F. and SHIRAKAWA, H. (1996) Squared Bessel processes and their applications to the square root interest rate model. *Working paper*.

DUFRESNE, D. (1990) The distribution of a perpetuity, with applications to risk theory and pension funding. *Scandinavian Actuarial Journal*, 39-79.

DYBVIG, P.H., INGERSOLL, J.E. and ROSS, S.A. (1986) Do Interest Rates Converge? *Working paper*.

DYBVIG, P.H., INGERSOLL, J.E. and ROSS, S.A. (1996) Long Forward and zero-coupon rates can Never Fall. *Journal of Business* **69**, 1-25.

EL KAROUI, N., FRACHOT, N. and GEMAN, H. (1998) On the behavior of long zero coupon rates in a no arbitrage framework. *Journal of Derivatives Research* **1**, 351-369.

GIACOTTO, C. (1986) Stochastic Modelling of Interest Rates: Actuarial vs. Equilibrium Approach. *Journal of Risk and Insurance* **53 (3)**, 435-453.

GOOVAERTS, M. and TEUNEN, M. (1995) Evaluation of interest randomness for pension valuation. *Proceedings of the International Congress of Actuaries, Brussels*, 689-709.

GOOVAERTS, M., LABIE, E. and VANNESTE, M. (1994) The Distributions of Annuities. *Insurance: Mathematics and Economics* **15**, 37-48.

HICKS, J.R. (1937) Mr. Keynes and the 'Classics'; A suggested interpretation. *Econometrica* **5**, 147-159.

HOGAN, M. (1993) Problems in certain two-factor term structure models. *Annals of Applied Probability* **3**, 576-581.

MAGHSOODI, Y. (1996) Solution of the extended CIR term structure and bond option valuation. *Mathematical Finance* **6**, 89-109.

MILEVSKY, M. (1997) The present value of a stochastic perpetuity and the Gamma distribution. *Insurance: Mathematics and Economics* **20**, 243-250.

PARKER, G. (1992) An application of Stochastic Interest Rates Models in Life Assurance. Ph.D. thesis, *Heriot-Watt University*.

PARKER, G. (1993) Distribution of the Present Value of Future Cash Flows. *Proceedings of the 3rd AFIR International Colloquium, Rome*, 831-843.

PARKER, G. (1994) Moments of the present value of a portfolio of policies. *Scandinavian Actuarial Journal*, 53-67.

PARKER, G. (1997) Stochastic analysis of the interaction between investment and insurance risks. *North American Actuarial Journal* **1 (2)**, 55-84.

PEARSON, N.D. and SUN, T.-S. (1994) Exploiting the conditional density in estimating the term structure: an application of the Cox, Ingersoll and Ross model. *Journal of Finance* **49**, 1279-1304.

PITMAN, J. and YOR, M. (1982) A decomposition of Bessel Bridges. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **59**, 425-457.

REVUZ, D. and YOR, M. (1991) *Continuous Martingales and Brownian Motion.* Springer-Verlag Berlin, Heidelberg, New York.

TICE, J. and WEBBER, N. (1997) An nonlinear model of the term structure of interest rates. *Mathematical Finance* **7**, 177-209.

YAO, Y. (1998) Term structure models and asymptotic long rate. *Proceedings of the Second International Congress on Insurance: Mathematics and Economics.*

YOR, M. (1992) On some exponential functionals of Brownian motion. *Advances in Applied Probability* **24**, 509-531.

GRISELDA DEELSTRA
ENSAE, CREST, Timbre J120
3, Av. Pierre Larousse
92245 Malakoff CEDEX,
France

Department of Mathematics,
Vrije Universiteit Brussel,
Pleinlaan 2,
B-1050 Brussel,
Belgium

RECURSIVE FORMULAE FOR SOME BIVARIATE COUNTING DISTRIBUTIONS OBTAINED BY THE TRIVARIATE REDUCTION METHOD

BY

J.F. WALHIN*• AND J. PARIS*

* *Université Catholique de Louvain*
• *Secura Belgian Re*

ABSTRACT

In this paper we study some bivariate counting distributions that are obtained by the trivariate reduction method. We work with Poisson compound distributions and we use their good properties in order to derive recursive algorithms for the bivariate distribution and bivariate aggregate claims distribution. A data set is also fitted.

KEYWORDS

Bivariate counting distributions, Poisson compound distribution, mixed Poisson distribution, Hofmann distribution, recursive algorithms, fit.

1. INTRODUCTION

Ahmed (1961) and Papageorgiou and David (1995) discuss some bivariate counting distributions, namely, the joint distribution (N, M) where $N = N_0 + N_1$ and $M = N_0 + N_2$ with N_0, N_1 and N_2 independent random variables such that

$$\mathbb{P}(N_0 = n) = \int_0^\infty e^{-\lambda} \frac{\lambda^n}{n!} dU(\lambda) , \quad n \geq 0$$

$$\mathbb{P}(N_1 = n) = e^{-\lambda_1} \frac{\lambda_1^n}{n!} , \quad n \geq 0$$

$$\mathbb{P}(N_2 = n) = e^{-\lambda_2} \frac{\lambda_2^n}{n!} , \quad n \geq 0$$

i.e. N_1 and N_2 are Poisson distributed while N_0 is mixed Poisson distributed with mixing distribution whose cumulative density function (cdf) is $U(\lambda)$.

The joint probability function (pf) of (N, M) is given by

$$\mathbb{P}(N = n, M = m) = \sum_{k=0}^{\min(n,m)} \mathbb{P}(N_0 = k)\mathbb{P}(N_1 = n - k)\mathbb{P}(N_2 = m - k)$$

For some choices of the mixing distribution Λ, Papageorgiou and David (1995) give the density of (N, M) by using Stirling numbers of the second kind, C-numbers and modified Bessel functions of the third kind.

Using a general class of counting random variables that are simultaneously mixed and compound Poisson, it is possible to give simple expressions for the joint distribution of (N, M) which avoid these numbers. Moreover our methodology gives easily the joint pf of the random sums

$$(S_N,\ S_M) = (X_1 + ... + X_N,\ Y_1 + ... + Y_M)$$

where $X_1,\ X_2,\ ...$ (resp. $Y_1,\ Y_2,\ ...$) is a random sample of observations from X (resp. Y).

X and Y are independent nonnegative arithmetic random variables that are also independent of (N, M). The distribution of $(S_N,\ S_M)$ is of interest in insurance problems where it represents the aggregate claims distributions when X and Y are claim amounts.

We will also extend the model to $(N,\ M) = (N_0 + N_1,\ N_0 + N_2)$ where N_0, N_1 and N_2 are mixed Poisson distributions.

Finally a data set will be fitted.

We will use the following conventions: $\sum_{k=a}^{b} = 0$ when $b < a$ and $\mathbb{P}(N = n,\ M = m) = 0$ when $n < 0$ or $m < 0$.

In order to prove the algorithms leading to recursive formulae for some compound distributions, we will use extensively the concept of ordinary generating function (see Panjer and Willmot (1992) for a reference in actuarial sciences).

Let a sequence $\{a_n,\ n = 0,\ 1,\ 2,\ ...\}$ of real numbers.

The ordinary generating function of this sequence is defined as

$$T_{a_n}(z) = \sum_{n=0}^{\infty} a_n z^n$$

Of course z must be chosen such that the sum exists.

Ordinary generating functions have the following nice properties:

- There is a one-to-one correspondence between $\{a_n,\ n = 0,\ 1,\ 2,\ ...\}$ and $T_{a_n}(z)$
- $a_n = \frac{1}{n!}\frac{d^n T_{a_n}(z)}{dz^n}|_{z=0}$
- $c_n = \alpha a_n + \beta b_n \Leftrightarrow T_{c_n}(z) = \alpha T_{a_n}(z) + \beta T_{b_n}(z)$
- $c_n = \sum_{k=0}^{n} a_k b_{n-k} \Leftrightarrow T_{c_n}(z) = T_{a_n}(z) T_{b_n}(z)$
- $T_{na_n}(z) = z\frac{d}{dz}T_{a_n}(z)$

The philosophy for using ordinary generating functions is the following:
- we look for a relation between some sequences a_n, b_n, c_n, ...
- go in the z map where the calculations become easier (think of the convolution that becomes a product)
- go back to the initial map by inverting the expression in z thanks to the properties.

The notion of ordinary generating function and its properties trivially extend in a bivariate setting.
In this paper, the sequence a_n or $a_{(n,m)}$ in a bivariate setting will be probability functions. As a consequence we will not have problems of convergence for the ordinary generating functions: $|z| < \infty$.
In the present case, ordinary generating functions are just probability generating functions (pgf).
From now on we will only refer to pgf and we will use them extensively in the sense of ordinary generating functions.

2. A GENERAL FAMILY OF RANDOM VARIABLES THAT ARE SIMULTANEOUSLY MIXED AND COMPOUND POISSON

Walhin and Paris (2000b) review the characteristics of a general family of random variables that have the property of being mixed and compound Poisson distributions.

A mixed Poisson process is such that

$$\Pi(n,t) = \mathbb{P}(N(t) = n) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^n}{n!} dU(\lambda)$$

where $N(t)$ gives the number of occurrences in $(0, t]$.

By choosing

$$\Pi(0,t) = e^{-\theta(t)}$$
$$\theta(t) \geq 0$$
$$\theta(0) = 0$$
$$\frac{d}{dt}\theta(t) \text{ completely monotone}$$

Walhin and Paris (2000b) show that $N(t)$ can also be interpreted as a compound Poisson model:

$$N(t) = \sum_{i=1}^{L(t)} \xi_i$$

where $L(t)$ is Poisson distributed and independent of the ξ_i which are independent identically distributed (i.i.d.) random variables. We will use this property in the sequel. A good choice for the function $\theta'(t)$ is the choice made by Hofmann (1955) and studied in Walhin and Paris (2000b) and in Kestemont and Paris (1985):

$$\theta'(t) = \frac{p}{(1+ct)^a} \quad p > 0, \quad c > 0, \quad a \geq 0$$

By integration, one has

$$\theta(t) = \frac{p}{c(1-a)}\left[(1+ct)^{1-a} - 1\right]$$
$$\theta(t) = \frac{p}{c}\ln(1+ct) \quad \text{by continuity for } a = 1$$

Particular cases of interest are Poisson ($a = 0$), Poisson Inverse Gaussian ($a = 0.5$), Negative Binomial ($a = 1$), Polya-Aeppli ($a = 2$) and Neymann Type A ($a \to \infty$, $c \to 0$, $ac \to b$).

Some properties are

$$\psi_{N(t)}(z) = \sum_{n=0}^{\infty} \Pi(n,t)z^n = \Pi(0, t - tz) = e^{-\theta(t-tz)}$$
$$EN(t) = pt$$
$$VarN(t) = pt + pact^2$$
$$\psi_{N(t)}(z) = e^{-\theta(t)\left(1-\psi_\xi(z)\right)} \tag{1}$$

where $\psi_X(z) = \mathbb{E}[z^X]$ denotes the pgf of the random variable X.
The probability law of the ξ_i is deduced from

$$q(0) = 0$$
$$\psi_\xi(z) = \sum_{k=1}^{\infty} q(k)z^k = 1 - \frac{\theta(t-tz)}{\theta(t)} \tag{2}$$
$$\frac{q(n)}{q(n-1)} = r + \frac{s}{n}, \quad n > 1 \tag{3}$$
$$r = \frac{ct}{1+ct}$$
$$s = r(a-2)$$

One says that the ξ_i belong to the $(r, s, 1)$ class. The 1 in $(r, s, 1)$ is connected with the $n > 1$ in (3).

From now on we will refer to the Hofmann distribution ($Ho(p,\ c,\ a)$) with the convention that $t = 1$. In model 1, N_0 will be $Ho(p,\ c,\ a)$ while in model 2, N_i will be $Ho(p_i,\ c_i,\ a_i)$, $i = 0,\ 1,\ 2$. Note that in general, the $(r,\ s,\ 1)$ class is denoted as $(a,\ b,\ 1)$ class. We use the notation $(r,\ s,\ 1)$ class in order to avoid confusion with the a of the Hofmann distribution.

3. Model 1: bivariate counting variables

In this section we work with the model

$$(N,\ M) = (N_0 + N_1,\ N_0 + N_2)$$

where N_0 is $Ho(p,\ c,\ a)$ and N_1 and N_2 are respectively $Po(\lambda_1)$ and $Po(\lambda_2)$. The three random variables are assumed to be mutually independent.

Let

$$\phi(u,v) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p(n,m)u^n v^m$$

be the pgf of (N, M) where we use the notation

$$p(n,m) = \mathbb{P}[N = n,\ M = m]$$

Let ψ_0, ψ_1 and ψ_2 the pgf of N_0, N_1 and N_2 respectively.
We have

$$\begin{aligned}\phi(u,v) &= \mathbb{E}\left[u^N v^M\right] = \mathbb{E}\left[(uv)^{N_0} u^{N_1} v^{N_2}\right] \\ &= \psi_0(uv)\psi_1(u)\psi_2(v) \\ &= e^{-\theta(1)\left[1-\psi_\xi(uv)\right]} e^{-\lambda_1(1-u)} e^{-\lambda_2(1-v)}\end{aligned}$$

Differentiating with respect to u and multiplying by u gives

$$u\frac{\partial\phi(u,v)}{\partial u} = \theta(1)uv\frac{\partial\psi_\xi(uv)}{\partial u}\phi(u,v) + u\lambda_1\phi(u,v)$$

Inverting this expression gives

$$np(n,m) = \theta(1)\sum_{k=1}^{\min(n,m)} kq(k)p(n-k,\ m-k) + \lambda_1 p(n-1,m), \quad n > 0$$

Differentiating with respect to v gives a symmetric recursion.

We have proved

Theorem 1
For model 1 the probability function is given by the following recursion

$$p(0,0) = e^{-\theta(1)-\lambda_1-\lambda_2}$$

$$np(n,m) = \theta(1)\sum_{k=1}^{\min(n,m)} kq(k)p(n-k,\ m-k) + \lambda_1 p(n-1,m), \quad n > 0$$

$$mp(n,m) = \theta(1)\sum_{k=1}^{\min(n,m)} kq(k)p(n-k,\ m-k) + \lambda_2 p(n,m-1), \quad m > 0$$

□

Let us note that for the particular cases where N_0 is Negative Binomial or Poisson Inverse Gaussian, we have easier recursions.

The case Negative Binomial ($a = 1$) is given in Hesselager (1996) where the fact that the Negative Binomial belongs to the (r, s, 0) class is used:

$$\mathbb{P}(N_0 = 0) = (1+c)^{-\frac{p}{c}}$$

$$\frac{\mathbb{P}(N_0 = k)}{\mathbb{P}(N_0 = k-1)} = r + \frac{s}{k}, \quad k > 0$$

with

$$r = \frac{c}{1+c}$$

$$s = \frac{p-c}{1+c}$$

The Negative Binomial can be expressed in the following explicit notation:

$$\mathbb{P}(N_0 = k) = \frac{(p/c)(p/c+1)\ldots(p/c+k-1)}{k!}\left(\frac{c}{1+c}\right)^k\left(\frac{1}{1+c}\right)^{(p/c)}$$

In this case the recursion becomes (Hesselager (1996)):

$$\begin{aligned}
p(0,0) =& (1+c)^{-\frac{p}{c}} e^{-\lambda_1-\lambda_2} \\
p(n,m) =& (r+\frac{s}{n})p(n-1,\ m-1) + \frac{\lambda_1}{n}p(n-1,m) \\
& -\frac{\lambda_1 r}{n}p(n-2,\ m-1), \quad n > 0 \\
p(n,m) =& (r+\frac{s}{m})p(n-1,\ m-1) + \frac{\lambda_2}{m}p(n,m-1) \\
& -\frac{\lambda_2 r}{m}p(n-1,\ m-2), \quad m > 0
\end{aligned}$$

The case Poisson Inverse Gaussian ($a = 0.5$) is derived without using the distribution of the ξ_i (see equations (1) and (2)). The pgf of (N, M) has the following direct properties:

$$
\begin{aligned}
(4) \qquad \phi(u,v) &= e^{-\frac{2p}{c}\left((1+c(1-uv))^{\frac{1}{2}}-1\right)} e^{-\lambda_1(1-u)} e^{-\lambda_2(1-v)} \\
\frac{\partial}{\partial u}\phi(u,v) &= \left(\lambda_1 + \frac{pv}{(1+c(1-uv))^{\frac{1}{2}}}\right)\phi(u,v) \\
\frac{\partial^2}{\partial u^2}\phi(u,v) &= \lambda_1^2\phi(u,v) + \frac{1}{2}\frac{cpv^2}{(1+c(1-uv))^{\frac{3}{2}}}\phi(u,v) \\
&\quad + \frac{p^2v^2}{1+c(1-uv)}\phi(u,v) + \frac{2\lambda_1 pv}{(1+c(1-uv))^{\frac{1}{2}}}\phi(u,v)
\end{aligned}
$$

$$
\begin{aligned}
(5) \quad (1+c(1-uv))\frac{\partial^2}{\partial u^2}\phi(u,v) &= \phi(u,v)\left(-\lambda_1^2(1+c) + \lambda_1^2 cuv - \frac{1}{2}\lambda_1 cv + p^2v^2\right) \\
&\quad + \frac{\partial}{\partial u}\phi(u,v)\left(2\lambda_1(1+c) - 2\lambda_1 cuv + \frac{1}{2}cv\right)
\end{aligned}
$$

From (4) we easily find the initializing terms:

$$
\begin{aligned}
p(0,0) &= e^{-\lambda_1-\lambda_2-2\frac{p}{c}\left((1+c)^{0.5}-1\right)} \\
p(1,0) &= \lambda_1 p(0,0) \\
p(0,1) &= \lambda_2 p(0,0) \\
p(1,1) &= \left(\lambda_1\lambda_2 + \frac{p}{(1+c)^{\frac{1}{2}}}\right)p(0,0)
\end{aligned}
$$

Inverting (5) and using its similar expression in v gives

$$
\begin{aligned}
(1+c)n(n-1)p(n,m) &= c(n-1)(n-\tfrac{3}{2})p(n-1,m-1) - \lambda_1 c(2n-\tfrac{7}{2})p(n-2,m-1) \\
&\quad + 2\lambda_1(1+c)(n-1)p(n-1,m) - \lambda_1^2(1+c)p(n-2,m) \\
&\quad + \lambda_1^2 cp(n-3,m-1) + p^2 p(n-2,m-2), \qquad n \geq 2 \\
(1+c)m(m-1)p(n,m) &= c(m-1)(m-\tfrac{3}{2})p(n-1,m-1) - \lambda_2 c(2m-\tfrac{7}{2})p(n-1,m-2) \\
&\quad + 2\lambda_2(1+c)(m-1)p(n,m-1) - \lambda_2^2(1+c)p(n,m-2) \\
&\quad + \lambda_2^2 cp(n-1,m-3) + p^2 p(n-2,m-2), \qquad m \geq 2
\end{aligned}
$$

We recall that p is one of the parameters of the Hofmann distribution ($Ho(p,\ c,\ a)$) while $p(n,m)$ is the pf of (N, M).

4. Model 1: bivariate random sums

Now let us study the bivariate vector

$$(S_N,\ S_M) = (X_1 + \dots + X_{N_0+N_1},\ Y_1 + \dots + Y_{N_0+N_2})$$

whose pf is given by

$$\mathbb{P}[S_N = x,\ S_M = y] = g(x,y) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p(n,m) f_X^{*n}(x) f_Y^{*m}(y)$$

where $f_X(x)$ (resp. $f_Y(y)$) is the pf of X (resp. Y). Our aim is to give a recursive scheme in order to derive the pf $g(x,y)$.

Let $\Phi(u,v)$ be the pgf of $(S_N,\ S_M)$ and let $\psi_X(u)$ (resp. $\psi_Y(v)$) be the pgf of X (resp. Y).
We have

$$\begin{aligned}\Phi(u,v) &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} p(n,m)\psi_X^n(u)\psi_Y^m(v)\\ &= \phi(\psi_X(u),\ \psi_Y(v))\\ &= e^{-\theta(1)\left[1-\psi_\xi(\psi_X(u)\psi_Y(v))\right]} e^{-\lambda_1(1-\psi_X(u))} e^{-\lambda_2(1-\psi_Y(v))} \qquad (6)\end{aligned}$$

Differentiating with respect to u and multiplying by u gives

$$u\frac{\partial \Phi(u,v)}{\partial u} = \theta(1) u \frac{\partial \Psi_\xi(u,v)}{\partial u}\Phi(u,v) + \lambda_1 u \frac{\partial \psi_X(u)}{\partial u}\Phi(u,v) \qquad (7)$$

where $\Psi_\xi(u,v) = \psi_\xi(\psi_X(u)\psi_Y(v))$ is the probability generating function of the pair

$$(X_1 + \dots + X_\xi,\ Y_1 + \dots + Y_\xi)$$

whose pf will be denoted by

$$h(i,j) = \mathbb{P}(X_1 + \dots + X_\xi = i,\ Y_1 + \dots + Y_\xi = j)\ , \quad i \geq 0, \quad j \geq 0$$

Inverting (7) gives

$$xg(x,y) = \theta(1)\sum_{i=0}^{x}\sum_{j=0}^{y} ih(i,j)g(x-i,y-j) + \lambda_1 \sum_{i=0}^{x} if_X(i)g(x-i,y)\ , \quad x > 0 \qquad (8)$$

The following theorem is a trivial extension of the bivariate Panjer (1981) algorithm given in Walhin and Paris (2000a). The proof is given for illustration.

Theorem 2
The probability function $h(x,y)$ of

$$(X_1 + \dots + X_\xi,\ Y_1 + \dots + Y_\xi)$$

is given by

$$(9)\qquad h(0,0) = 1 - \frac{\theta(1 - f_X(0) f_Y(0))}{\theta(1)}$$

$$(10)\qquad h(x,y) = \frac{1}{1 - r f_X(0) f_Y(0)} \left(\sum_i^x \sum_j^y (r + s\frac{i}{x}) f_X(i) f_Y(j) h(x-i, y-j) + q(1) f_X(x) f_Y(y) \right), \quad x > 0$$

$$(11)\qquad h(x,y) = \frac{1}{1 - r f_X(0) f_Y(0)} \left(\sum_i^x \sum_j^y (r + s\frac{j}{y}) f_X(i) f_Y(j) h(x-i, y-j) + q(1) f_X(x) f_Y(y) \right), \quad y > 0$$

where we use the notation

$$\sum_i^x \sum_j^y w(i,j) = \sum_{i=0}^x \sum_{j=0}^y w(i,j) - w(0,0)$$

Proof
Equation (9) follows immediately from equation (2).
Now we prove equation (10). We have

$$(12)\qquad kq(k) = r(k-1)q(k-1) + (r+s)q(k-1)\,, \quad k > 1$$

Multiplying each side of (12) by $\psi_X^{k-1}(u) \frac{d}{du} \psi_X(u) \psi_Y^k(v)$ and summing from $k = 1$ to $k = \infty$ we find

$$\begin{aligned} \frac{\partial}{\partial u} \Psi_\xi(u,v) - q(1) \frac{d}{du} \psi_X(u) \psi_Y(v) &= r \frac{\partial}{\partial u} \Psi_\xi(u,v) \psi_X(u) \psi_Y(v) \\ &\quad + (r+s) \Psi_\xi(u,v) \frac{d}{du} \psi_X(u) \psi_Y(v) \\ &\quad - (r+s) q(0) \frac{d}{du} \psi_X(u) \psi_Y(v) \end{aligned}$$

Multiplying by u and inverting gives

$$\begin{aligned} xh(x,y) &= r\sum_{i=0}^{x}\sum_{j=0}^{y}(x-i)f_X(i)f_Y(j)h(x-i,\ y-j) \\ &+ (r+s)\sum_{i=0}^{x}\sum_{j=0}^{y} if_X(i)f_Y(j)h(x-i,\ y-j) \\ &+ q(1)xf_X(x)f_Y(y) \end{aligned}$$

Rearranging gives (10).
(11) follows similarly. □

Of course more general results can be derived if (12) becomes

$$kq(k) = r(k-1)q(k-1) + (r+s)q(k-1)\ , \quad k > m$$

for a general m. In this case, one says that the ξ_i belong to the $(r,\ s,\ m)$ class.
With the symmetric expression of (8) we have the following result:

Theorem 3
For the model 1, the probability function $g(x,y)$ of the compound distribution is given by the following recursion:

$$g(0,0) = e^{-\theta(1-f_X(0)f_Y(0))}e^{-\lambda_1(1-f_X(0))}e^{-\lambda_2(1-f_Y(0))} \tag{13}$$

$$g(x,y) = \theta(1)\sum_{i=1}^{x}\sum_{j=0}^{y}\frac{i}{x}h(i,j)g(x-i,\ y-j) + \lambda_1\sum_{i=1}^{x}\frac{i}{x}f_X(i)g(x-i,y)\ , \qquad x > 0 \tag{14}$$

$$g(x,y) = \theta(1)\sum_{i=0}^{x}\sum_{j=1}^{y}\frac{j}{y}h(i,j)g(x-i,\ y-j) + \lambda_2\sum_{j=1}^{y}\frac{j}{y}f_Y(j)g(x,y-j)\ , \qquad y > 0 \tag{15}$$

where $h(i,j)$ is given by theorem 2.

Proof
Equation (13) is immediately derived from equations (6) and (2).
Equation (14) follows immediately from equation (8) while its similar expression valid for $y > 0$ gives (15). □

5. Model 2

In this section we consider the following model

$$(N,\ M) = (N_0 + N_1,\ N_0 + N_2)$$

where N_i are independent $Ho(p_i,\ c_i,\ a_i)$. The corresponding pf of the ξ_i are denoted by q_i.
Then we have the following results. The proofs are similar to those given in sections 3 and 4. So we omit them.

Theorem 4
For model 2, the probability function $p(n,m)$ is given by the following recursion:

$$p(0,0) = e^{-\theta_0(1)-\theta_1(1)-\theta_2(1)}$$

$$p(n,m) = \sum_{i=1}^{\min(n,m)} \frac{i}{n}\theta_0(1)q_0(i)p(n-i,\ m-i) + \sum_{i=1}^{n}\frac{i}{n}\theta_1(1)q_1(i)p(n-i,m),\quad n>0$$

$$p(n,m) = \sum_{i=1}^{\min(n,m)} \frac{i}{m}\theta_0(1)q_0(i)p(n-i,\ m-i) + \sum_{i=1}^{m}\frac{i}{m}\theta_2(1)q_2(i)p(n,m-i),\quad m>0$$

□

Theorem 5
For model 2, the probability function $g(x,y)$ is given by the following recursion:

$$g(0,0) = e^{-\theta_0(1-f_X(0)f_Y(0))}e^{-\theta_1(1-f_X(0))}e^{-\theta_2(1-f_Y(0))}$$

$$g(x,y) = \theta_0(1)\sum_{i=1}^{x}\sum_{j=0}^{y}\frac{i}{x}h(i,j)g(x-i,\ y-j) + \theta_1(1)\sum_{i=1}^{x}\frac{i}{x}b_X(i)g(x-i,y),\quad x>0$$

$$g(x,y) = \theta_0(1)\sum_{i=0}^{x}\sum_{j=1}^{y}\frac{j}{y}h(i,j)g(x-i,\ y-j) + \theta_2(1)\sum_{j=1}^{y}\frac{j}{y}b_Y(j)g(x,y-j),\quad y>0$$

where $h(i,j)$ is given by theorem 2 and $b_X(k)$ is given by

$$b_X(0) = 1 - \frac{\theta_1(1-f_X(0))}{\theta_1(1)}$$

$$b_X(k) = \frac{1}{1-r_1 f_X(0)}\left(\sum_{i=1}^{k}(r+s\frac{i}{k})f_X(i)b_X(k-i) + q_1(1)f_X(k)\right),\quad k>0$$

and $b_Y(k)$ is defined similarly. □

Remark: for the case where the N_i belong to the $(r, s, 0)$ class, Hesselager (1996) gives an easier algorithm. However, for the case Negative Binomial which is a member of the $(r, s, 0)$ class, numerical examples show that this algorithm is not stable while the combination of theorems 5 and 2 give stable recursions.

6. A fit

We use in this section a set of accident data used in Papageorgiou and David (1995) for illustration.

We fit model 1 for the following choices of N_0:

- $a = 0$: Poisson
- $a = 1$: Negative Binomial
- $a = 0.5$: Poisson Inverse Gaussian
- a free: general Hofmann distribution.

The fits are proceeded by maximum likelihood. N and M are accident observations.
It can be shown that

$$\hat{p} + \hat{\lambda}_1 = \bar{N}$$
$$\hat{p} + \hat{\lambda}_2 = \bar{M}$$

where $\bar{N}$ (resp. $\bar{M}$) is the empirical mean of N (resp. M).
This reduces the number of estimates to be found by numerical techniques. For the Hofmann fit we need to maximize numerically the loglikelihood subject to three variables.
We find the following estimates:

TABLE 1

MAXIMUM LIKELIHOOD ADJUSTMENT

	λ_1	λ_2	p	c	a	*loglikelihood*
Poisson	1.0319	1.2724	0.6388		0	−112.7380
PIG	1.0893	1.3298	0.5815	0.8432	0.5	−112.3577
NB	1.0939	1.3344	0.5769	0.4092	1	−112.3802
Hofmann	1.0796	1.3201	0.5912	1.6697	0.2546	−112.3467

Based on the comparison of the loglikelihoods, it does not seem necessary to work with a more complicated model than the one obtained with all N_i, $i = 0, 1, 2$ being Poisson distributed. A likelihood ratio test would not reject this hypothesis.

The original data with fitted values are in the next table:

TABLE 2

OBSERVED AND FITTED DISTRIBUTIONS FOR THE NUMBER OF ACCIDENTS

N	*M*	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
0	*o.bs*	5	6	4	1	1	0	0	0
	$a = 0$	4.16	5.30	3.37	1.43	0.45	0.12	0.02	0.00
	$a = 0.5$	4.29	5.71	3.80	1.68	0.56	0.15	0.03	0.01
	$a = 1$	4.29	5.73	3.82	1.70	0.57	0.15	0.03	0.01
	$a = 0.2546$	4.29	5.67	3.74	1.65	0.54	0.14	0.03	0.01
1	*obs*	4	9	3	4	3	0	0	0
	$a = 0$	4.30	8.13	6.86	3.63	1.38	0.41	0.10	0.02
	$a = 0.5$	4.68	8.06	6.58	3.46	1.33	0.40	0.10	0.02
	$a = 1$	4.70	8.03	6.53	3.43	1.32	0.40	0.10	0.02
	$a = 0.2546$	4.63	8.10	6.65	3.50	1.34	0.41	0.10	0.02
2	*obs*	2	5	5	4	2	0	0	0
	$a = 0$	2.22	5.56	6.14	4.06	1.87	0.65	0.18	0.04
	$a = 0.5$	2.55	5.39	5.52	3.57	1.65	0.59	0.17	0.04
	$a = 1$	2.57	5.35	5.47	3.55	1.65	0.59	0.17	0.04
	$a = 0.2546$	2.50	5.44	5.61	3.63	1.67	0.59	0.17	0.04
3	*obs*	1	6	4	1	1	2	1	0
	$a = 0$	0.76	2.39	3.30	2.70	1.51	0.62	0.20	0.05
	$a = 0.5$	0.92	2.32	2.93	2.40	1.39	0.60	0.21	0.06
	$a = 1$	0.94	2.30	2.91	2.41	1.41	0.62	0.21	0.06
	$a = 0.2546$	0.90	2.34	2.97	2.41	1.37	0.59	0.20	0.06
4	*obs*	0	0	0	2	0	0	0	0
	$a = 0$	0.20	0.74	1.23	1.22	0.82	0.40	0.15	0.05
	$a = 0.5$	0.25	0.73	1.11	1.14	0.85	0.47	0.20	0.07
	$a = 1$	0.26	0.73	1.11	1.16	0.87	0.49	0.21	0.07
	$a = 0.2546$	0.24	0.74	1.12	1.12	0.82	0.45	0.19	0.06
5	*obs*	0	0	1	0	0	0	0	0
	$a = 0$	0.04	0.18	0.35	0.41	0.33	0.19	0.08	0.03
	$a = 0.5$	0.05	0.18	0.32	0.41	0.39	0.29	0.16	0.07
	$a = 1$	0.06	0.18	0.32	0.42	0.40	0.29	0.16	0.07
	$a = 0.2546$	0.05	0.18	0.32	0.39	0.37	0.27	0.15	0.06
6	*obs*	0	0	0	0	1	0	0	1
	$a = 0$	0.01	0.03	0.08	0.11	0.10	0.07	0.03	0.01
	$a = 0.5$	0.01	0.04	0.08	0.11	0.14	0.13	0.10	0.05
	$a = 1$	0.01	0.04	0.08	0.12	0.14	0.13	0.09	0.05
	$a = 0.2546$	0.01	0.04	0.07	0.11	0.13	0.12	0.09	0.05

We have also conducted a χ^2 test in order to judge the goodness of fit. As usual it is important to be extremely cautious with the results of the χ^2 test. The grouping rule we have adopted may lead to conclusions that are not matched by another grouping rules. Moreover the χ^2 test is an asymptotic test. However we have only 79 observations in our data set.

The grouping rule we have adopted is the rule A in Lemaire (1995), i.e. all the theoretical values > 1 and 80% of the theoretical values > 5.

The cells have been grouped as follows: (0,0), (0,1), $(0, \geq 2)$, (1,0), (1,1), (1,2), $(1, \geq 3)$, (2,1), (2,2), (2,3), $(2, \geq 4)$, $(3, \geq 0)$, $(\geq 4, \geq 0)$. The χ^2 values as well as the associated p-values are given in the following table.

TABLE 3

GOODNESS OF FIT TEST

	χ^2	*df*	*p-value*
Poisson	5.36	9	0.80
PIG	6.48	8	0.59
NB	6.42	8	0.60
Hofmann	6.42	7	0.49

Based on this figures, all the fits are acceptable but the Poisson fit wins. This is coherent with the conclusions drawn after analysing the loglikelihoods.

ACKNOWLEDGEMENTS

We would like to thank both anonymous referees for their constructive remarks that led to a better presentation of the paper.

REFERENCES

AHMED, M.S. (1961) On a Locally Most Powerful Boundary Randomized Similar Test for the Independence of two Poisson Variables. *Ann. Math. Statist.* **32**, 809-827.

HESSELAGER, O. (1996) Recursions for Certain Bivariate Counting Distributions and their Compound Distributions. *Astin Bulletin* **26**, 35-52.

HOFMANN, M. (1955) Uber zusammengesetzte Poisson-Prozesse und ihre Anwendungen in der Unfall versicherung. *Bulletin of the Swiss Actuaries* **55**, 499-575.

KESTEMONT, R.M. and PARIS, J. (1985) Sur l'Ajustement du Nombre de Sinistres. *Bulletin of the Swiss Actuaries* **85**, 157-164.

LEMAIRE, J. (1995) *Bonus-Malus Systems in Automobile Insurance*. Kluwer Academic Publishers, Boston.

PANJER, H.H. (1981) Recursive Evaluation of a Family of Compound Distributions. *Astin Bulletin* **12**, 22-26.

PANJER, H.H. and WILLMOT, G.E. (1992) *Insurance Risk Models*. Society of Actuaries.

PAPAGEORGIOU, H. and DAVID, K.M. (1995) On a Class of Bivariate Compounded Poisson Distributions. *Statistics & Probability Letters* **23**, 93-104.

WALHIN, J.F. and PARIS, J. (2000a) Excess-of-Loss Reinsurance with Reinstatements and Ruin of the Cedent. *submitted*.

WALHIN, J.F. and PARIS, J. (2000b) A Large Family of Discrete and Overdispersed Probability Laws. *forthcoming*.

JEAN-FRANÇOIS WALHIN
SECURA Belgian Re
Rue Montoyer 12
1000 Bruxelles

ECONOMIC ASPECTS OF SECURITIZATION OF RISK

BY

SAMUEL H. COX, JOSEPH R. FAIRCHILD AND HAL W. PEDERSEN [1]

ABSTRACT

This paper explains securitization of insurance risk by describing its essential components and its economic rationale. We use examples and describe recent securitization transactions. We explore the key ideas without abstract mathematics. Insurance-based securitizations improve opportunities for all investors. Relative to traditional reinsurance, securitizations provide larger amounts of coverage and more innovative contract terms.

KEYWORDS

Securitization, catastrophe risk bonds, reinsurance, retention, incomplete markets.

1. INTRODUCTION

This paper explains securitization of risk with an emphasis on risks that are usually considered insurable risks. We discuss the economic rationale for securitization of assets and liabilities and we provide examples of each type of securitization. We also provide economic axguments for continued future insurance-risk securitization activity. An appendix indicates some of the issues involved in pricing insurance risk securitizations. We do not develop specific pricing results. Pricing techniques are complicated by the fact that, in general, insurance-risk based securities do not have unique prices based on axbitrage-free pricing considerations alone. The technical reason for this is that the most interesting insurance risk securitizations reside in *incomplete* markets.

A market is said to be *complete* if every pattern of cash flows can be replicated by some portfolio of securities that are traded in the market. The payoffs from insurance-based securities, whose cash flows may depend on

---

[1] Please address all correspondence to Hal Pedersen.

hurricanes, earthquakes and so on, cannot be closely approximated by a portfolio of the traditional assets that are already traded in the market such as stocks and bonds. This is because there are states of the world reflected in insurance-based securities that are not reflected by the existing traditional securities. In a complete market a new security can always be priced relative to existing securities by finding a replicating portfolio and pricing it. The no-arbitrage property implies that the new security and the replicating portfolio must have the same price because they have the same payoffs. However, if the market is incomplete a replicating portfolio may not exist and arbitrage considerations alone may not determine a unique price. The appendix describes a method for dealing with incompleteness. In the main body of the paper we do not discuss arbitrage based pricing theory further but the reader who will ultimately be involved in the pricing of these products should bear in mind that there are fundamental practical differences between products to be valued in complete markets and products that are valued in incomplete markets. Incompleteness is one of the unusual characteristics of insurance-based securities relative to many other securitizations. It is very interesting to note that the fundamental reason insurance risk securitizations tend to reside in incomplete markets – namely that states of the world reflected in insurance based securities are not reflected by the existing traditional securities, is also the fundamental reason why these securities provide diversification of investment risk and thereby make these attractive investments for many portfolio managers. Although we will not explicitly use the notion of incompleteness in the main body of the paper because the focus of this paper is not on technical valuation, an actuary involved in these securitization deals must be aware of these fundamental pricing issues.

Two actuarial principles, diversification and contractual risk transfer, play important roles in most securitizations, yet relatively few actuaries work in the securitization business. It seems that the opportunities for actuaries in securitization will increase and we may see more actuaries working in this field in the future.

We begin this paper with an idealized catastrophe property risk securitization. This example illustrates the key ideas without abstract mathematical or financial theories. We hasten to emphasize that although the key ideas of securitization can be illustrated without these theories, the practical implementation of a securitization deal requires financial theory for pricing and risk measurement. As a broad definition, securitization means "the bundling or repackaging of rights to future cash flows for sale in capital markets." In all the cases we mention here, and more generally in all of the deals we know of, the repackaging provides a more efficient allocation of risk. This process can be costly, but evidently the reallocation is valuable enough to make it worthwhile.

After describing this simple example, we turn to the common features of securitization and then review some recent catastrophe risk securitizations. We compare catastrophe risk securitizations with the asset securitizations: bond strips, mortgage-backed securities, life insurance policyholder loans, and life insurance premium loadings. The Chicago Board of lyade offers options based on property insurance loss ratios. We mention them only to contrast them with catastrophe risk securitizations. We discuss some possible future uses of securitization of insurance risks. The paper ends with a discussion of the economics of securitization. We offer a discussion of the reasons for these transactions and attempt to answer the questions:

- Why do investors buy insurance-based securities?
- Why do insurers use securitizations to cover insurable risks?

2. Securitization of Catastrophe Risk

We will give a simple illustrative idealization of catastrophe risk bonds – customarily referred to as cat bonds. During 1997 and 1998 there were successful catastrophe risk bond issues by USAA, Swiss Re, Winterthur, St. Paul Re, and others. Later we will provide an economic rationale for the supply (why to insurers sell cat bonds?) and demand (why do investors buy cat bonds?). For now we focus on the mechanics of these transactions.

We illustrate the model with two examples, first a single-period model and second a two-period model. In each example catastrophe risk has a binomial structure. There is no interest rate risk in either example. The market interest rate on risk-free securities is a constant 8% per year. The probability of a catastrophe that triggers a "default" is a constant 3% per year [1]. These values are merely to illustrate the mechanics of the transactions. In practice we would use the prevailing interest rate term structure and a model for insurance losses to determine the probabilities. Embrechts and Meister take this approach to develop a valuation model for exchange-traded insurance options [12].

Example 1. The first example is similar to the USAA bonds. The face amount is 100 and the annual coupon rate is 12%. Coupon and principle are at risk. This means that the principal and coupon are paid only if no catastrophe occurs during the period [0, 1]. The total principal and coupon 112 is paid at time 1 only if no catastrophe occurs during the period [0, 1]. The catastrophe states and probabilities, along with the corresponding cat bond cash flows are shown in Figure 1. The positive cash flow is paid to the bondholders; the negative cash flow is the price the bondholders pay to obtain the rights to the future cash flow.

---

[1] Failure to pay a coupon or to repay the principal because a catastrophe occurs is not a default in the legal sense. The catastrophic event is well-defined in the bond indenture and buyers and sellers understand the circumstances under which coupons and principal will not be paid. Nevertheless, it is convenient to refer to this event as a default.

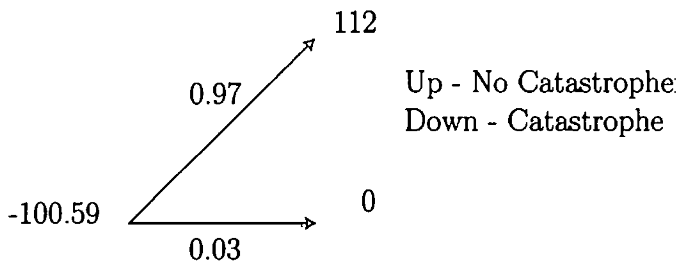


FIGURE 1: One-Period Catastrophe Risk Bond Cash Flow.

The expected bondholder payments, averaged over the catastrophe distribution, are $\bar{c}(1) = 112(0.97) + (0)(0.03) = 108.64$. The discounted expected value, using the constant 8%, is the price of the cat bond:

$$\frac{1}{1.08}[108.64] = 100.59$$

Consider a bond that has the same prospective cash flow (*i.e.*, 12% coupon), but no possibility of default. This is called a *straight bond*. The price of the straight bond at the time the cat bond is issued is found by discounting the cash flow:

$$\frac{1}{1.08}[112] = 103.70$$

The cash flows of the straight bond are shown for comparison to the cat bond in Figure 2.

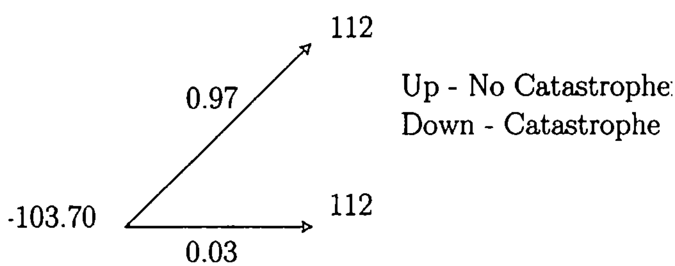


FIGURE 2: One-Period Straight Bond Cash Flow.

Suppose an insurer (like USAA) issues the cat bond and simultaneously buys the straight bond. The straight bond is more expensive. The trades cost the insurer 3.11 per 100 of face value (ignoring transactions costs). What does the insurer get in return? If there is no catastrophe, the insurer's net cash flow is zero because it receives the straight bond coupon and pays the cat bond coupon. However, if there is a catastrophe, it still receives the straight bond coupon and principal (112), but does not pay the corresponding cat bond cash flow. In effect, the insurer has purchased a one-year catastrophe reinsurance contract which pays 12 in case a catastrophe occurs during the period. This increases the insurers capacity to sell insurance for one year (just as a traditional reinsurance does) by 112 at a cost of 3.11 per 100 of bond face value. The rate on line [1] for this "synthetic" reinsurance is $100 \times 3.11/112 = 2.78$ per 100 of coverage per year. The net cash flow is shown in Figure 3.

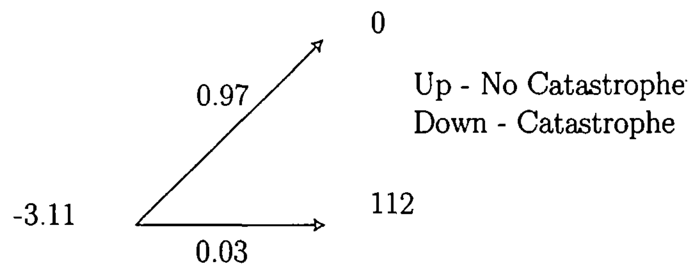


FIGURE 3: One-Period Net Cash Flow: Long Straight Bond and Short Cat Bond.

There are several multiple period cat bonds. The majority are essentially extensions of the concept illustrated in Example 1 in that the bond "defaults" as soon as a catastrophe occurs, regardless of when the catastrophe occurs. The bond indenture may specify that future coupon and principal payments to bondholders are forfeited as soon as a catastrophe occurs. Alternatively it may specify that coupons only are at risk or that coupons and a fraction of the principal is at risk. USAA actually issued one series with coupon only at risk and another with principal and coupon at risk. The Swiss Re [20] and Yasuda Marine [29] bonds have a single limit applicable over several years. The Winterthur bonds take yet another form allowing the limit to be reset each year. Our second example is like the bond Winterthur issued in 1997[2].

Example 2. Coupons only are at risk. This means that the principal of 100 is paid to the bondholder at $k = 2$ with probability one. A coupon of 12 is paid at $k = 1, 2$ provided no catastrophe occurs during the period $[k - 1, k]$. The catastrophe states and probabilities, along with the corresponding cat bond cash flows are shown in Figure 4. The positive cash flows are paid to the bondholders, the negative cash flow is the price the bondholders pay to obtain the rights to future cash flows.

[1] For a one-year policy, *rate on line* is the ratio of premium to coverage layer, usually multiplied by 100. The concept is not usually applied to multiple year policies.

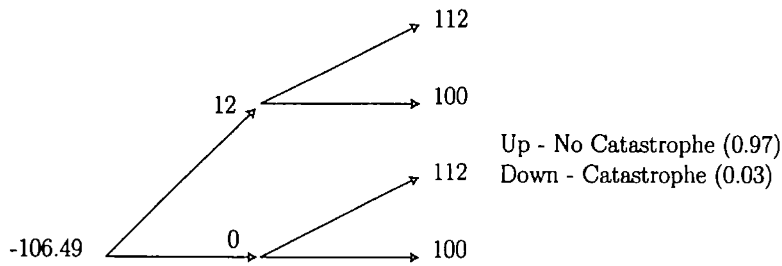


FIGURE 4: Two-Period Catastrophe Risk Bond Cash Flow.

As in the first example, the expected bondholder payments are $\bar{c}(1) = 12(0.97) = 11.64$ and $\bar{c}(2) = 100 + 11.64 = 111.64$. The discounted expected value is the price of the cat bond:

$$\frac{1}{1.08}\left[11.64 + 111.64\left(\frac{1}{1.08}\right)\right] = 106.49$$

Consider a bond that has the same prospective cash flow (i.e., 12% coupons), but no possibility of default. This is called a *straight bond.* The price of the straight bond at the time the cat bond is issued is found by discounting the cash flows:

$$\frac{1}{1.08}\left[12 + 112\left(\frac{1}{1.08}\right)\right] = 107.13$$

The cash flows of the straight bond are shown for comparison to the cat bond in Figure 5.

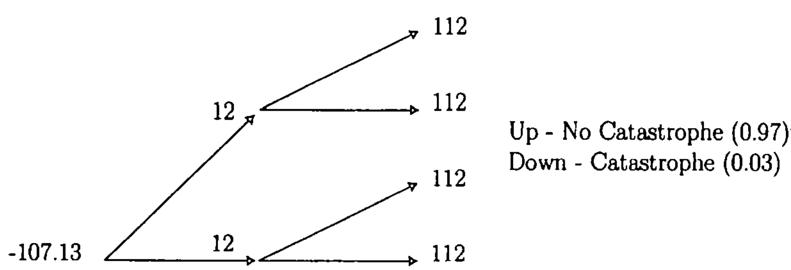


FIGURE 5: Two-Period Straight Bond Cash Flow.

As before, suppose an insurer (like Winterthur or Swiss Re) issues the cat bond and simultaneously buys the straight bond. The trades cost the insurer 0.64 per 100 of bond face value and provide 12 units of coverage per period. The "rate on line" is $100 \times 0.64/12 = 5.33$, but one must keep in mind that this is the rate paid once at the beginning of the policy period for a two year cover. If we must compare this to a one year policy, we should divide by two: $5.33/2 = 2.66$. In each of the two future periods, if there is no catastrophe, the insurer's net cash flow is zero because it receives the straight bond coupon and pays the cat bond coupon. However, if there is a catastrophe in either period, it still receives the straight bond coupon (12), but does not pay the cat bond coupon. In effect, the insurer has purchased a two year catastrophe reinsurance contract which pays 12 in case a catastrophe occurs during either period. This increases the insurers capacity to sell insurance for each of the next two years by 12 at cost of 0.64 per 100 of face value (or 5.33 single premium per 100 of coverage for a two year cover). The net cash flow is shown in Figure 6.

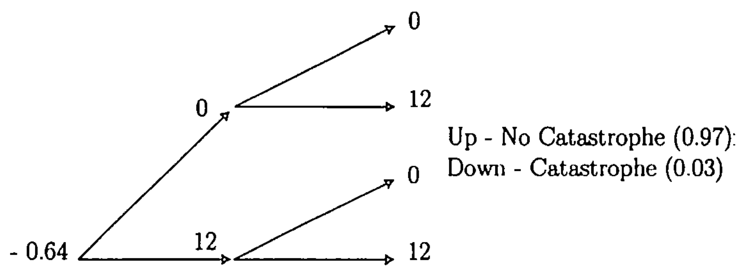


FIGURE 6: Net Cash Flow: Long Straight Bond and Short Cat Bond.

The actual deals we have described all increase the bond issuer's capacity. The technology required to issue cat securities is being developed and refined and thus the transactions costs of these deals will probably decrease in the future. Moreover, investors are becoming more familiar with the product which will have a further tendency to render future deals relatively less costly. Lastly, as others have pointed out [13], the insurance industry would be strained by a $50 billion hurricane loss, but the capital markets could withstand it with relative calm. Catastrophe bonds may become a routine method of transferring catastrophe risk. Practical considerations and economic theory would both predict this outcome.

It should be emphasized that the line of insurance is immaterial to the capital market – it does not have to be catastrophe risk. We will show later that investors will demand these bonds because their returns have low correlation with stock returns. There may be many kinds of insurance risks that have low covariance with the stock market. At the 1997 Swiss Actuarial Summer School held at the University of Lausanne we heard from Winterthur actuaries of a proposal to issue bonds which would transfer mortality risk to bondholders [1]. It seems intuitively clear to us that mortality risk has low covariance with the stock market and thus we expect these bonds would be attractive to investors. As we understand it, Winterthur has long term annuity liabilities and as a result faces the risk of unexpected improvement in beneficiary mortality. A security with bondholder cash flows

[1] In late 1999 we learned that three large international insurers are considering securitization of mortality risk.

tied to a mortality index would provide Winterthur with very long term coverage that is not available in the traditional reinsurance market. In the United States some companies offer very attractive term life insurance rates on selected lives in a very competitive market. There is little experience to indicate what the ultimate mortality will be for these select lives. Securitization would allow very long term coverage of the risk that ultimate mortality will diverge greatly from projected mortality for the selected lives.

3. Structure of Securitization

The securitization technology applies to many kinds of risk, not merely catastrophe risk. In asset and liability securitizations the common structure typically involves four entities: retail customers, a retail contract issuer, a special purpose company, and investors. In the case of catastrophe risk bonds, the four entities are as follows:

(1) Homeowners who buy policies from an insurer.
(2) The insurance company that issues the homeowners policies (*i.e.*, the retail contracts) and buys reinsurance from a special purpose reinsurer (*i.e.*, the special purpose company).
(3) The special purpose reinsurer that issues the reinsurance and sells bonds.
(4) Investors who buy the bonds.

Figure 7 illustrates the direction and timing of cash flows to and from each entity involved in or related to a securitization.

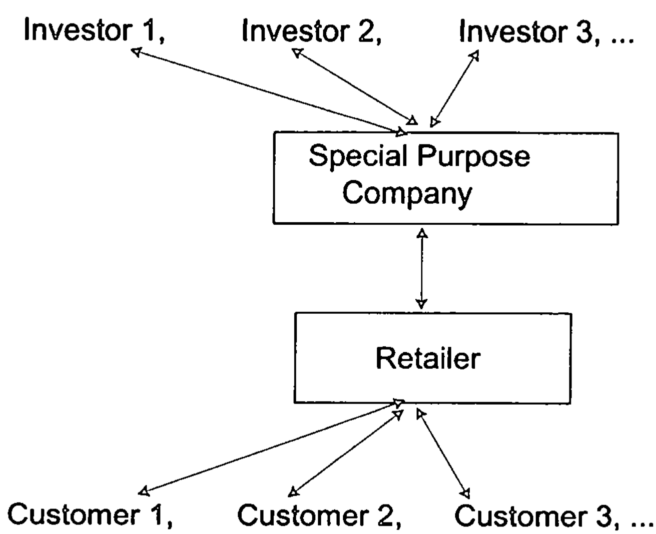


FIGURE 7: Securitization Components.

Each of the arrows denotes an exchange of cash corresponding to a contract. The timing varies with the application. For example, in the case of homeowners insurance, the customers pay a cash premium to the insurer and get a contract (the homeowners policy) in exchange. Later the cash flows the other way for those customers who suffer losses and obtain insurance benefits. The insurer pays a premium initially to the special purpose reinsurer and gets a reinsurance policy in exchange. Later, the cash may flow the other way if the catastrophic event or events occur. The investors initially pay cash to the special purpose company and get bonds in exchange. Later they receive coupons and principal, provided no catastrophes occur. The special purpose company invests the combined premiums and proceeds from the sale of the bonds in default free securities.

These transactions provide a structure for which the price of the bonds (paid by the investors), the reinsurance premium (paid by the retailer) and investment income are adequate to cover the catastrophe loss with certainty. Tilley [25, 26] refers to this as a fully collateralized transaction since the special purpose insurer cannot default on the reinsurance contract. By collateralizing the transaction the risk of default, called *counter-party risk*, is eliminated [1]. The ability to eliminate counter-party risk is an advantage of securitization relative to traditional reinsurance.

Insurance risk securitizations present a moral hazard problem that has to be addressed. The insurer has an incentive to apply the coverage to a loss so it will not have pay a coupon, so the investors will want to see that the terms of the coverage are applied properly. We are aware of two methods for resolving the problem that have been used in practice.

Method (1)
The security can be written in terms of an independently determined loss ratio. This takes determination of the security's coverage out of the hands of the insurer, solving the problem, but introducing *basis-risk* – the contract covers industry losses, not the insurer's own losses.

Method (2)
An independent firm is hired to provide claims services.

We now turn our attention to some recent catastrophe risk bond deals.

USAA hurricane bonds. USAA is a personal lines insurer based in San Antonio. It provides financial management products to current or former US military officers. *Business Insurance* [27] in reporting on the USAA deal, described USAA as "over exposed" to hurricane risk due to its personal

[1] Counter-party risk is the risk that the other party will fail to pay as required by the contract. This can be a significant risk in a reinsurance contract, but it is nil in securitizations as we have described them.

automobile and homeowners business along the US Gulf and Atlantic coasts. In June 1997, USAA arranged for its captive Cayman Islands reinsurer, Residential Re, to issue $477 million face amount of one-year bonds with coupon and/or principal exposed to the risk of property damage incurred by USAA policyholders due to Gulf or East coast hurricanes. Residential Re issued reinsurance to USAA based on the capital provided by the bond sale. USAA sold $450 million of similar bonds again in 1998 according to an article in the *Financial Times* [1].

The 1997 bonds were issued in two series (also called tranches), according to an article in *The Wall Street Journal* [22]. In the first series only the coupons are exposed to hurricane risk – the principal is guaranteed. The return of principal will be at the end of the first year if there is no loss (described below), but the return will be at the end of ten years if a loss occurs. For the second series both coupons and principal are at risk. The risk is defined as damage to USAA customers on the Gulf or East coast during the year beginning in June due to a Class-3 or stronger hurricane. The coupons and/or principal will not be paid to investors if these losses exceed one billion dollars. That is, the risk begins to reduce coupons at $1 billion and at $1.5 billion the coupons in the first series are completely gone (and the principal repayment delayed nine years) and in the second series the coupons and principal are lost. The coupon-only tranche has a coupon rate of LIBOR plus 2.73%. The principal and coupon tranche has a coupon rate of LIBOR + 5.76%. The press reported that the issue was "oversubscribed," meaning there were more buyers than bonds, *i.e.*, demand exceeded supply. The press reports indicated that the buyers were life insurance companies, pension funds, mutual funds, money managers, and, to a very small extent, reinsurers. As a point of reference for the risk involved, we note that industry losses due to hurricane Andrew in 1992 amounted to $16.5 billion and USAA's Andrew losses amounted to $555 million. Niedzielski reported in the *National Underwriter* that the cost of the coverage was about 6% rate on line plus expenses [1]. According to Niedzielski's (unspecified) sources the comparable reinsurance coverage is available for about 7% rate on line. The difference, however, may or may not be completely offset by the expenses related to establishing Residential Re and the fees to the investment bank for issuing the bonds. One might argue that the higher cost of securitization is justified by lower *counter-party* risk. The rate on line refers only to the cost of the reinsurance. The reports did not give the sale price of the bonds, but the investment bank probably set the coupon so that they sold at face value.

---

[1] As we noted earlier, rate on line is the ratio of premium to coverage layer. The reinsurance agreement provides USAA with 80 percent of $500 million in excess of $1 billion. The denominator of the rate on line is (0.80)(500) = 400 million, so this implies USAA paid Residential Re a premium of about (0.06)(400) = 24 million.

As successful as this issue turned out, it was a long time coming. Despite advice of highly regarded advocates such as Morton Lane and Aaron Stern [13, 14, 19], catastrophe bonds have developed more slowly than many experts expected. According to press reports, USAA has obtained 80% of the coverage of its losses in the \$1.0 to \$1.5 billion layer with this deal. On the other hand, we have to wonder why these are one year deals. Perhaps it is a matter of getting the technology in place, forcing reinsurers to lower prices on future deals, related US tax code issues, *etc.* The off-shore reinsurer is re-usable and the next time USAA goes to the capital market investors will be familiar with these exposures. If the traditional catastrophe reinsurance market gets tight, they will have a capital market alternative. The cost of this issue is offset somewhat by the gain in access to alternative sources of reinsurance. The 1998 issue was more favorable to USAA; it a reported a yield to the bondholders of LIBOR + 4% [1].

Winterthur Windstorm Bonds. Winterthur is a large insurance company based in Switzerland. In February 1997, Winterthur issued three year annual coupon bonds with a face amount of 4700 Swiss francs. The coupon rate is 2.25%, subject to risk of windstorm (most likely hail) damage during a specified exposure period each year to Winterthur motor insurance customers. The deal was described in the trade press and Schmock has written an article in which he values the coupon cash flow [21]. The deal has been mentioned in US and European publications (for example, *Investment Dealers Digest* [18] and *Euroweek* [2]). If the number of motor vehicle (automobile and motorcycle) windstorm claims during the annual observation period exceeds 6000, the coupon for the corresponding year is not paid. The bond has an additional financial wrinkle. It is convertible at maturity; each face amount of CHF 4700 plus the last coupon is convertible to five shares of Winterthur common stock at maturity. Furthermore, due to the merger of Winterthur Insurance and Crédit Suisse Group on December 15, 1997, investors can now convert into 35.5 Crédit Suisse Group registered shares at maturity of the WinCat bond [1].

Swiss Re California Earthquake Bonds. The Swiss Re deal is similar to the USAA deal in that the bonds were issued by a Cayman Islands reinsurer, evidently created for issuing catastrophe risk bonds, according to an article in *Business Insurance* [28]. However, unlike USAA's deal, the underlying California earthquake risk is measured by an industry-wide index rather than Swiss Re's own portfolio of risks. The index was developed by Property Claims Services. The bond contract is written on the same (or similar) California index underlying the Chicago Board of Trade (CBOT) Catastrophe Options. The CBOT options have been the subject of numerous scholarly and trade press articles [8, 10, 11, 12]. As described above, in a

[1] We learned of this from one of the ASTIN referees.

securitization of insurance risk there is a moral hazard problem that has to be addressed in the contract defining the contingent events covered by the security. The investors demand that the losses be reported accurately and in accordance with the contract. The Swiss Re bonds are written in terms of the PSC index, neatly solving the moral hazard problem, although it introduces basis-risk. In this case, basis-risk is the risk that the actual Swiss Re losses differ from industry losses, evidently acceptably small.

Zolkos reported details on the Swiss Re bonds in *Business Insurance*. There were earlier reports that Swiss Re was looking for a ten year deal. This deal is not it and perhaps they are still looking for such a ten year deal. According to Zolkos, SR Earthquake Fund (a company Swiss Re set up for this purpose) issued Swiss Re $122.2 million in California reinsurance coverage based on funds provided by the bond sale.

Yasuda Fire and Marine Bonds. Finally we note that recently Aon Capital Markets structured and marketed a catastrophe bond providing windstorm coverage to Yasuda Fire and Marine Insurance Company [29]. Munich Re "validated the transaction from the perspective of investors" and will provide claims services. Evidently, the moral hazard problem we mentioned earlier is resolved in this case by using Munich Re's claims services. No investment banks were mentioned in the reports because Aon Capital Markets acted as and is registered as its own investment bank. This is an example of how brokers and reinsurers have reacted to securitization – they are acquiring the skills needed to enter the business and marketing services explicitly. The coverage is long term, provides Yasuda with dual "trigger" options (we discuss these in detail in another paper [9]), and makes use of the reputation and administrative services of an established reinsurer. In the next section we review securitization of assets.

4. Asset Securitizations

We are going to describe five examples: stripping coupons, mutual funds, mortgage-backed securities, life insurance policyholder loans, and life insurance premium loadings.

Stripping Coupons. Merrill Lynch and other investment banks create default free zero coupon bonds by means of an asset securitization. This is an example of securitization of securities – repackaging and reselling securities. The resulting securities are called T-bond-backed securities. The bank buys U.S. Treasury bonds. It issues its own zero coupon bonds based on the cash flow from its pool of coupon bearing bonds. In this case, the "customers" are all the same entity: the U.S. government. The retailer and the special purpose company are the same, the bank. The investors buy the zero coupon bonds from the bank. The zero coupon bonds are issued by a private corporation but the bond covenant conveys the pooled Treasury bond cash flow to the zero coupon bondholders. Therefore, the bank's bonds are default-free. The popularity of zero coupon bonds led the U.S. and Canadian governments to assign registration numbers to coupons of some bonds when they are issued. This allows the coupons to be traded directly without securitization. Nevertheless, securitization is still used to create zero coupon bonds. The actuarial textbook [4, page 73] has a simple numerical illustration and the investments textbook [3, page 414] describes some of the marketing aspects of this securitization.

T-bond securitization is a simple asset securitization example, but it illustrates the essential components and principles of these deals. The reason for this securitization is that the demand for default-free zero coupon bonds exceeds the supply provided by the government. A 30-year coupon bearing bond exposes its owner to changes in interest rates corresponding to maturities over the 30-year term of the bond. A zero coupon bond is sensitive only to the interest rate corresponding to its only payment. Therefore, this securitization divides the pooled cash flow into pieces that better meet the needs of some investors and provide a preferable (or more efficient) allocation of interest rate risk. It is an illustration of the use of contracts to transfer and reallocate risk.

Mutual Funds. Pooling also underlies mutual funds and mortgage backed securities (MBS). A mutual fund purchases assets, such as stocks or bonds. The fund sells securities (or shares) that provide the owner a proportional share of the market value of the pool. In this way, an investor receives the average return of the pooled assets without buying shares in each individual asset. Fund managers issue shares in the mutual fund to the investors in exchange for cash and the fund managers have a contractual obligation to buy individual stocks. Owners are entitled to a proportionate share of the fund, less operating fees and commissions. Why would investors prefer to buy a mutual fund rather than the individual shares? Under "perfect market" assumptions, the absence of transactions costs, perfect divisibility of shares, *etc.*, investors would *not* buy mutual funds as they could do for themselves exactly what the mutual fund does for them. However, the real world is not perfect and mutual funds exist because of market "imperfections." First consider transactions costs. Trading stocks is costly because stock brokers charge commissions, but the commission rates are less for those who make large trades on a regular basis. Therefore, a mutual fund has an advantage relative to individual investors because it will have lower transactions costs. A second imperfection is lack of divisibility. An individual may want to buy a stock with a high price per share. Berkshire Hathaway is trading for about $52,000 per share (January 2000). Some investors might want to have some Berkshire Hathaway shares, but buying as few as 10 shares might be impossible. On the other hand, the same individual may have shares in a mutual fund that can easily own 100 or more shares, providing the individual a fraction of Berkshire Hathaway's value. A third consideration is the cost of information acquisition. Under the conditions of "perfect markets" all investors have access to the same information – an assumption that is clearly violated in the real world. Information acquisition is expensive, but a mutual fund applies the same information on behalf of all of its owners, providing an economy of scale.

Finally we consider the diversification of risk. We begin with a brief discussion of the Markowitz [16] risk-return model in order to illustrate diversification. Later we will use the same model to determine the effect of adding insurance-based securities to a portfolio. We will follow the Luenberger's exposition [15]. A different but equivalent approach appears in [4, Chapter 8]. Luenberger shows how to use the model, with some additional assumptions, to describe the effect of diversification. This is a one period market model, focused on the first two moments of the joint distribution of return random variables R_1, R_2, ..., R_n, namely

- the expected returns $\mu_i = E[R_i]$ and
- the covariance matrix $\Sigma = [\sigma_{ij}]$ where $\sigma_{ij} = \mathrm{Cov}(R_i, R_j)$.

These moments can be estimated by observing return outcomes over several time periods, assuming stationarity. Statistics derived from the observations estimate the risk versus return relation in the future for portfolios of assets.

Following Luenberger's discussion of diversification [15, page 200], let us assume that we can write the return of each asset in terms of a single factor: $R_i = a_i + b_i F + \varepsilon_i$ where a_i and b_i are constants, F is a random variable (the single factor), and the ε_i are random error terms. Assume that the following relations hold:

$$E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = 0 \text{ for } i \neq j, \quad \mathrm{Cov}(F, \varepsilon_i) = 0,$$

and their variances have a common bound $\mathrm{Var}(\varepsilon_i) \leq s^2$.

A portfolio is constructed from the n given assets by specifying the percentage of the value of the portfolio which is invested in each asset. Under the assumptions commonly used, the scale of investment does not affect the percentages in the sense that investors with the same risk-return preferences will select the same portfolios regardless of the size of their investments. Hence in specifying a portfolio, we need only specify the percentage invested in each security. We let w_i denote the percentage invested in the i-th asset; it is called the *weight* of asset i in the portfolio.

For a "well diversified" portfolio, we can assume that each weight is about $1/n$. The portfolio return is $R_w = \sum_{i=1}^{n} w_i r_i = a + bF + \varepsilon$ where

$$a = \sum_{i=1}^{n} w_i a_i, \quad b = \sum_{i=1}^{n} w_i b_i \quad \text{and} \quad \varepsilon = \sum_{i=1}^{n} w_i \varepsilon_i.$$

Under the assumptions we made above, the variables ε and F are uncorrelated so $\mathrm{Var}(R) = b^2 \mathrm{Var}(F) + \mathrm{Var}(\varepsilon)$. Since the errors ε_i are uncorrelated, $\mathrm{Var}(\varepsilon) = s^2/n$ and as n increases this term tends to zero. Diversification eliminates this component. The other component does not tend to zero because b is the average of the b_i. This term represents *non-diversifiable risk*. The diversification principle is familiar to actuaries from its application to pools of insurance policies.

In summary, mutual funds exist because they provide greater efficiency, overcome some of the effects of market imperfections, and provide diversification of risks more efficiently than individual investors can achieve on their own.

Mortgage-Backed Securities (MBS). A mortgage is a loan requiring periodic payments of principal and interest with real estate as collateral [1]. The mortgage may be for a residence or for commercial real estate. We limit our discussion to US residential mortgages [2]. They are commonly issued with a fixed interest rate for a period of 15 to 30 years and require level monthly payments of interest and principal. Fixed-rate mortgages carry substantial interest-rate risk for the lender, especially in volatile economic times. For example, when interest rates fall, borrowers may re-finance their mortgages, returning the principal to the lender at a time when interest rates are lower than the rate at which the mortgage was issued. There are costs to re-financing, but when rates fall enough, borrowers have financial incentives to refinance. Mortgage securitization shifts the interest rate risk to investors through the securities market.

For mortgage-backed securities the components of the securitization are easy to identify: The customers are the mortgage borrowers. Initially the borrowers obtain cash and in exchange provide the lenders with a contractual obligation to repay the loan. The lenders convey their rights to a trust in exchange for cash. The trust issues securities based on the pooled mortgage contracts. The securities can take a variety of different forms.

One purpose of mortgage securitization (re-packaging) is to allow for a more efficient allocation of interest rate risk. Primary mortgage lenders (*e.g.*, banks and thrifts) usually have short-term demand deposits as liabilities, so for most of them mortgage assets are not well matched to their liabilities. On the other hand, life insurers, with long term liabilities, may desire to have mortgage-backed securities in their asset portfolios. We discuss two mortgage-backed securities: pass-through securities and stripped mortgage-backed securities. Several other forms exist, but these illustrate the basic ideas.

First we discuss *pass-through* mortgage-backed securities. With pass-through securities, mortgage borrowers make their monthly payments to the pool administrator. The pool collects the cash, deducts administrative fees,

---

[1] This section relies on [24, Chapter 6].

[2] Mortgage terms and lending practices are different in other countries. For example, in Canada, mortgages are typically written for 5 or 10 years with a balloon payment (which is often refinanced) and no prepayment option. The Canadian practices put the interest rate risk on the borrower, lenders bear none, and there is no need for reallocating the lender's interest rate risk – and no mortgage-backed securities.

and passes the remaining cash to the security owners on a pro-rata basis. Thus, if a pool issues ten securities, each security owner receives one-tenth of the aggregate monthly cash flow, less fees. If a mortgage is repaid during the month, the repaid principal is paid to the security owners along with the monthly cash flow. Thus, the security owners bear the prepayment risk. Valuation of a pass-through security requires knowing the rates and maturities of the pooled mortgages. This and other information is provided to potential purchasers. An actuarial approach would involve modeling the "life" of a mortgage and considering the cash flow to be a cash-refund annuity. The difficulty, and the distinction from mortality-dependent cash flow, is that the mortgage life depends of the interest rate environment. All mortgage-backed securities present these same valuation problems.

A *stripped* mortgage-backed security divides the payments from pooled mortgages into classes with each class's security holder receiving income only from its portfolio, instead of distributing it on a pro-rata basis. For example, consider a stripped security with two classes: interest only and principal only. The interest-only class receives the interest paid on the pooled mortgages each month. The principal-only class receives each month's principal payments. Suppose that a representative mortgage in the pool carries an outstanding principal of $90,000, an interest rate of 6 percent, and a level monthly payment of $600. Ignoring fees, the interest-only class would be allocated $450 ($\$90,000 \times 0.06/12$) this month from this mortgage. The principal-only class receives the principal paid with respect to the illustrative pool mortgage; that is, $150 ($600 – $450) if the mortgage is not repaid during the month. If the illustrative mortgage loan is repaid during the month, the principal-only class receives $90,000. The two classes receive similar payments from each mortgage with an outstanding balance at the beginning of the month.

The stripped pass-through security owners bear the interest rate risk of the pool, but it is allocated differently than it is for straight pass-throughs. The interest-only class receives interest until all the mortgages are repaid. Refinancing activity increases with falling interest rates, so the downside for interest-only security owners arises with declining interest rates. The principal-only class benefits from a decline in interest rates because refinancing means principal-only security owners receive their principal sooner. Thus, the stripped pass-through divides the cash flow pool into segments that give a pure reflection of the result of an increase or decrease in interest rates. This is more flexible than a straight pass-through mortgage-backed security and will appeal to many investors. After all, an investor who wants a straight pass-through could simply buy shares of both interest-only and principal-only classes.

Securitization of the mortgage industry has allowed investors to enter the mortgage market without having to be (or own) a mortgage originator. Insurance companies and pension funds have become substantial investors in MBS. Thus, securitization has allowed for a better allocation of interest rate risk and provided a more efficient way for capital to enter the home financing industry. The securitization technique is important for actuaries because the resulting products are used by insurance companies, the technique can be applied to other asset classes, and, perhaps most important, the expertise required to design and value these securities is fundamentally actuarial in nature. Let us illustrate this claim with the following idealized model.

Suppose that we are interested in a pass-through MBS for which the contractually specified monthly payments for the mortgage borrowers (in the absence of additional cash flows due to prepayment) per \$1 of face amount of the mortgage is denoted by c. Let the contractually specified effective monthly interest rate on the mortgage be denoted by r. In the absence of prepayment risk, level monthly payments are made over the entire term of the mortgage and the present value of these payments is equal to the face amount of the mortgage pool. In practice, mortgage borrowers will prepay with varying intensity and this rate of prepayment could depend on a variety of economic variables. For the sake of this illustration, let us assume that the rate of prepayment depends on the time since the issue of the mortgage (this makes an allowance for the average time a home is owned) and an annualized key interest rate level (for example, the 10-year yield rate on US treasury bonds, which makes an allowance for the cost of refinancing) denoted i. Providing the actuary has access to sufficient data, he would then estimate a two-dimensional table of prepayment rates. Let $q(t, i)$ denote the amount prepaid over month t to month $t+1$ per dollar of principal remaining when the key rate is equal to i. Let ℓ_t denote the amount of principal remaining in the mortgage pool at the end of the t-th month after the mortgage is issued. The total cash flow to the mortgage pool over month t to month $t+1$ is

$$\ell_t c + (\ell_t - [\ell_t c - \ell_t r])q(t, i).$$

In words, this monthly cash flow is the ordinary payment of interest and principal – namely $\ell_t c$, plus the amount of the remaining principal that is prepaid [1] – namely $(\ell_t - [\ell_t c - \ell_t r])q(t, i)$. This is a stochastic cash flow that depends on the key rate history. The evolution of the remaining principal in the mortgage pool can be determined recursively through the equation

$$\begin{aligned}\ell_{t+1} &= \ell_t - [\ell_t c - \ell_t r] - (\ell_t - [\ell_t c - \ell_t r])q(t, i)\\ &= [1 - q(t, i)](\ell_t - [\ell_t c - \ell_t r]).\end{aligned}$$

This is a stochastic equation for the evolution of the outstanding principal in the mortgage pool. The actuary can then value the MBS using stochastic cash flow valuation techniques from financial economics. The MBS market is a complete market and the valuation will be done in the context of a

---

[1] Note that $\ell_t c - \ell_t r$ is the amount of the regular payment that is applied to principal reduction that month prior to the prepayment amount being applied.

complete term structure model. Although the estimation of the prepayment rates and the definition of the MBS cash flows are fundamentally actuarial, the actuary must also be able to use tools from modern financial economics to complete his calculation of the value of the MBS and to assess the risks in the MBS.

Policy-Loan-Backed Securities. The laws of the United States and some other countries require certain life insurance policies to have cash values (savings). In still other countries, cash values are not legally required, but are commonly provided. In general, cash values emerge when the expected value of future benefits promised under a policy exceed the expected value of future (adjusted) premiums. In lay terms, cash values emerge when policyholders prepay future mortality costs. Cash values can be thought of as a type of savings within a life insurance policy that is available when a policyholder terminates (surrenders) his or her policy.

Economically, cash values are policyholder assets in the custody of the insurance company. Rather than surrendering their policies to obtain funds, policyholders may elect to borrow an amount not greater than the cash value from the insurance company on the security of their cash values. In the United States and some other countries, cash value policies are required to allow such borrowing privileges. Of course, the policyholder pays interest on the loan. Traditionally, U.S. insurers offered fixed-rate policy loans, but as interest rate volatility increased in the 1970s and 1980s, most companies began issuing policies with an indexed loan interest rate. When the interest rate is fixed, the policy loan provision is an interest rate call option. The value of the option increases with the volatility of interest rates.

Policy loans are carried on insurers' financial statements as assets. Securitization of a portfolio of policy loans allows the company to sell them. One reason for doing this is to reduce the cash strain induced by policy loan activity. Also, there may be a tax advantage when the loans are sold at a loss relative to their statement value. These reasons led to a large securitization of policy loans by the Prudential Insurance Company of America in 1987.

Policy loan interest and principal payments formed the cash flow to support the securities that were sold to investors as private placement policy-loan-backed securities. A special purpose corporation (SPC) was formed to issue the securities and simultaneously purchase the loan cash flow from Prudential, similar to collateralized mortgage obligations, as discussed in the preceding section. While the Prudential securitization borrowed concepts from the securitization of mortgage loans, it also employed new features. Since policy loan securitization was new, security buyers had no experience with loan repayment rates. To reduce the repayment risk to security owners, the securities provided for a minimum and maximum repayment schedule. If actual repayments fell behind the minimum schedule, Prudential promised to advance the needed cash to meet the required payments to security owners. (Cash flow simulations indicated that this was highly unlikely.) If repayments proved more rapid than the maximum, the SPC would invest the excess cash flow in a guaranteed investment contract (GIC). The SPC bought a 54-year GIC from a AAA-rated Swiss bank to provide security owners with evidence that the SPC would be able to perform on these promises.

The circumstances surrounding this transaction may be comparatively rare. Perhaps high interest rates might make them attractive again someday. On the other hand, as the value of the loan option in newer U.S. policies is nil, the magnitude of the problem created by increased exercise activity is steadily decreasing. Also, the costs of a policy loan securitization are substantial. Therefore, it may be a long time before we see another policy loan securitization in the United States.

Loadings in Premiums. In January 1997 the US life insurer American Skandia Life Assurance Corporation securitized mortality and expense risk fees that it will collect in the future from a portfolio of its variable annuity (VA) policies [6].

When a company issues a VA it pays a commission to an agent or financial advisor. Profits develop later. Thus issuing a VA requires cash, in contrast to other types of business that merely require setting up a reserve that may be financed with a non-cash asset or reinsurance. The faster the company grows the greater the need for cash. Skandia's fast growth led it to supplement traditional methods of financing growth (retained earnings, surplus notes, bank loans and reinsurance) with a securitization of the future fees Skandia will collect from a block of policies.

According to Connolly, the mortality and expense risk fees were taken from a block of approximately 33,000 American Skandia variable annuities, net of reinsurance, issued during the period between January 1, 1994 and June 30, 1996. The rights to the fees for a specified period of time were sold to American Skandia Investment Holdings, American Skandia Life Assurance Corporation's immediate parent, which transferred them to a trust, collateralized them and sold them to two investors, TIAA-CREF and Prudential Insurance Company. A total of eight insurers expressed interest in the offering.

Skandia managers think the costs of securitization will decrease as the process becomes more efficient and, ultimately, it should be cheaper than financing growth with reinsurance.

In this example, the customers are the variable annuity policyholders. Skandia is the retailer and the trust is the special purpose company. The investors are TIAA-CREF and Prudential. The actuarial modeling developed for traditional purposes (designing, pricing, cash flow testing, *etc.*) can be used in the securitization process. Since the buyers are also life insurers, they should have the expertise to evaluate the future fee cash flows. In general, a securitization of insurance risks would probably require independent consulting actuaries to resolve this moral hazard problem. In general, investors are not likely to have the expertise and they are not likely to accept the retailer's analysis without independent corroboration.

We have not seen an increase in life insurance securitizations. Mutual companies have more difficulty than stock companies in raising capital and cash. In the US many of them are electing to demutualize but at least one UK company is using securitization as an alternative.

5. The demand for insurance-based securities

Why do investors buy catastrophe risk bonds? The demand for securities based on insurance risk can be justified by the Markowitz mean-variance model. As we mentioned earlier, this is a one period market model. The assets returns over the period are random variables R_1, R_2, ..., R_n with means and covariances assumed to be know and denoted by $\mu_1 = E(R_i)$ and $\Sigma = [\sigma_{ij}]$ where $\sigma_{ij} = \mathrm{Cov}(R_i, R_j)$.

The n by n matrix $\Sigma = [\sigma_{i,j}]$, called the covariance matrix, is symmetric and the diagonal elements are simply the variances. We assume it is invertible.

A portfolio is constructed from the n given assets by specifying the percentage of the value of the portfolio which is invested in each asset. As stated earlier, we assume that the scale of investment does not affect the percentages in the sense that investors with the same risk-return preferences will select the same portfolios regardless of the size of their investments. Hence in specifying a portfolio, we need only specify the percentage invested in each security. We let w_i denote the percentage invested in the i-th asset; it is called the weight of asset i in the portfolio.

The return on the portfolio specified by the vector

$$w^{\mathrm{T}} = [w_1,\ w_2,\ \ldots,\ w_n]$$

is denoted by

$$R_w = \sum_{i=1}^{n} w_i R_i.$$

The portfolio return is the weighted average of the individual security returns. Thus the expected portfolio return, $\mu_w = E[R_w]$, and variance, $\sigma_w^2 = \mathrm{Var}[R_w]$, can be calculated in terms of the weights and the statistics of the individual securities as follows:

$$\mu_w = \sum_{i=1}^{n} w_i E[R_i] = \mu^{\mathrm{T}} w$$

$$\begin{aligned} \sigma_w^2 &= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j Cov(R_i, R_j) \\ &= w^{\mathrm{T}} \Sigma w \end{aligned}$$

The portfolio variance is a function of the vector of weights $w^T = [w_1, w_2, ..., w_n]$ and the covariance matrix $\Sigma = [\sigma_{i,j}]$.

An efficient portfolio is defined to be one which is not dominated by another portfolio. It is a portfolio for which there is none other with lower variance [1] and an equal or higher expected return. Figure 8 illustrates the concept of efficiency and the associated notion of portfolio dominance. Note that portfolio B dominates portfolio A since it offers the same variance but has a higher expected return. Similarly, portfolio B dominates portfolio C since it offers the same expected return but a lower variance. The basic portfolio problem is to find the maximum portfolio return for a given portfolio variance or the minimum portfolio variance for a given portfolio return. These optimal portfolios are said to be mean-variance efficient portfolios.

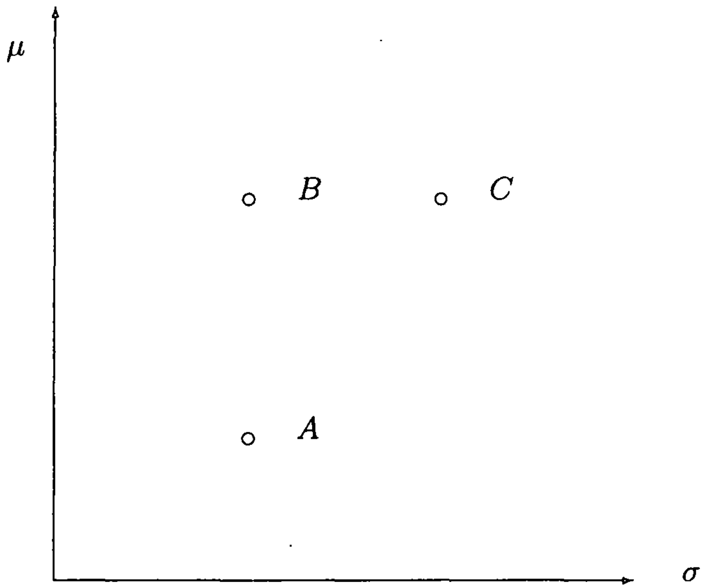


FIGURE 8: Risk and Return Relations.

There are a number of variants of the general portfolio problem. The following formulation of the standard version comes from [4]: Given the investor's required portfolio expected return $r > 0$ and a set of n securities

[1] Either portfolio variance σ_w^2 or standard deviation σ_w can be used to measure risk. In the graphs we follow the usual practice of plotting expected return μ_w on the vertical axis and risk represented by standard deviation σ_w on the horizontal axis.

with expected returns vector $\mu^{\mathsf{T}} = [\mu_1, \mu_2, \ldots, \mu_n]$ and covariance matrix Σ, determine the portfolio weights w in order to minimize the variance $\sigma_w^2 = w^{\mathsf{T}}\Sigma w$ subject to two constraints:

$$\sum_{i=1}^{n} w_i = 1$$

and

$$\mu_w = \mu^{\mathsf{T}} w = r.$$

The first constraint simply requires that the portfolio be 100% invested in the n risky securities being considered for inclusion in the optimal portfolio. It is convenient to introduce the n-vector $e^{\mathsf{T}} = [1, 1, \ldots, 1]$. The first constraint can be written compactly as $w^{\mathsf{T}}e = 1$. The second constraint selects the portfolio return to meet the investor's requirement. Of course there is a potentially different efficient portfolio for each target return we might select. In fact, we can graph an entire set of efficient portfolios, plotting the points (σ_w, r) by solving the portfolio problem for different values of σ_w corresponding to a range of values of target expected returns r. This graph is called the *efficient frontier* for the given n assets. The efficient frontier can be completely defined in terms of two efficient portfolios. This is the "two fund theorem," described by Luenberger [15, page 163] as follows.

The objective function, augmented with Lagrange terms corresponding to the constraints, is

$$\frac{1}{2} w^{\mathsf{T}}\Sigma w + \lambda(w^{\mathsf{T}}\mu - r) + \nu(w^{\mathsf{T}}e - 1).$$

The factor $\frac{1}{2}$ in the variance term is for convenience only. The objective is quadratic in the unknown weights w and linear is the Lagrange multipliers λ, ν, so the first order conditions for a minimum form a system of $n+2$ linear equations:

$$\sum_{j=1}^{n} \sigma_{i,j} w_j + \lambda\mu_i + \nu = 0 \quad \text{for } 0 \le i \le n$$

$$w^{\mathsf{T}}\mu = r$$

$$w^{\mathsf{T}}e = 1$$

Write this as a single matrix equation

$$\Sigma^{\mathrm{aug}}[w, \lambda, \nu]^{\mathsf{T}} = [0, \ldots, 0, r, 1]^{\mathsf{T}} \tag{1}$$

where Σ^{aug} is the result of augmenting the matrix Σ with two rows and columns:

$$\Sigma^{\text{aug}} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,n} & \mu_1 & 1 \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \sigma_{2,n} & \mu_2 & 1 \\ & & \dots & & & \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_{n,n} & \mu_n & 1 \\ \mu_1 & \mu_2 & \dots & \mu_n & 0 & 0 \\ 1 & 1 & \dots & 1 & 0 & 0 \end{bmatrix}$$

In addition to assuming that the covariance matrix Σ is invertible, we also assume the expected return vector μ is not a multiple of e. This just means that the last two columns of the augmented matrix Σ^{aug} are linearly independent. Clearly, each of the first n columns of Σ^{aug} is linearly independent of each of the last two. Because Σ is invertible, the first n columns of Σ^{aug} are linearly independent. Because of the independence of μ and e, the last two columns of Σ^{aug} are linearly independent. Therefore, the columns of Σ^{aug} are linearly independent, it is invertible, and there is a unique solution for the weights w and the multipliers λ, ν.

Let $(\sigma_B,\ r_B)$ denote the risk and expected return of a minimum variance portfolio. By this we mean an efficient portfolio with minimum variance among all efficient portfolios for various values of r. In general, the minimum variance could be zero corresponding to a market with a risk free security. However we assume that at this point we are considering only risky assets and $\sigma_B > 0$. We might think of this as a portfolio of corporate bonds; they are risky but not so risky as equity securities. Let w_B, λ_B, ν_B denote the corresponding weights and multipliers. Of course $\sigma_B = \left((w_B)^{\mathsf{T}}\Sigma w_B\right)^{1/2}$ and $r_B = \mu^{\mathsf{T}} W_B$.

Select any other efficient portfolio with weights w_S, multipliers λ_S, ν_S, return $r_S = \mu^{\mathsf{T}} w_S$ and risk $\sigma_S = \left((w_S)^{\mathsf{T}}\Sigma w_S\right)^{1/2}$ with $r_S > r_B$ and $\sigma_S > \sigma_B$. While the lower risk fund $(\sigma_B,\ r_B)$ intuitively represents a bond fund, the more risky fund $(\sigma_S,\ r_S)$ represents an equity portfolio.

Given any point (σ, r) on the efficient frontier, form the portfolio with weights w and multipliers λ, ν satisfying

$$[w,\ \lambda,\ \nu]^{\mathsf{T}} = (1-a)[w_B,\ \lambda_B,\ \nu_B]^{\mathsf{T}} + a[w_S,\ \lambda_S,\ \nu_S]^{\mathsf{T}}$$

where $a = (r - r_B)/(r_S - r_B)$. Now since $\Sigma^{\text{aug}}[w_B,\ \lambda_B,\ \nu_B] = [0,\ 0,\ \dots,\ r_B,\ 1]^{\mathsf{T}}$ and $[w_S,\ \lambda_S,\ \nu_S] = [0,\ 0,\ \dots,\ r_S,\ 1]^{\mathsf{T}}$, then $\Sigma^{\text{aug}}[w,\ \lambda_1,\ \lambda_2] = [0,\ 0,\ \dots,\ r,\ 1]$.

The solution is unique so we have

$$r = \mu w = (1-a)r_B + a r_S$$

and

$$\begin{aligned}\sigma^2 = \sigma_w^2 &= \mathrm{Var}[(1-a)R_B + aR_S] \\ &= (1-a)^2\sigma_B^2 + 2a(1-a)\rho\sigma_B\sigma_S + a^2\sigma_S^2\end{aligned}$$

where we have abbreviated the notation with $R_B = \mu^{\mathrm{T}} W_B$ and $R_S = \mu^{\mathrm{T}} W_S$. Also we wrote the covariance term as

$$\mathrm{Cov}(R_B,\ R_S) = \rho\sigma_B\sigma_S,$$

where the correlation coefficient is ρ. In effect every point on the efficient frontier can be obtained as a weighted average of the two fixed portfolios W_B and W_S. This is what Luenberger calls the two fund theorem. Figure 9 illustrates the two fund theorem, showing two frontiers that differ only in that the solid frontier has a greater value of ρ than the dashed frontier. This illustrates that if nothing changes except the correlation is reduced, then the frontier pushes out to the left for those points between B and S.

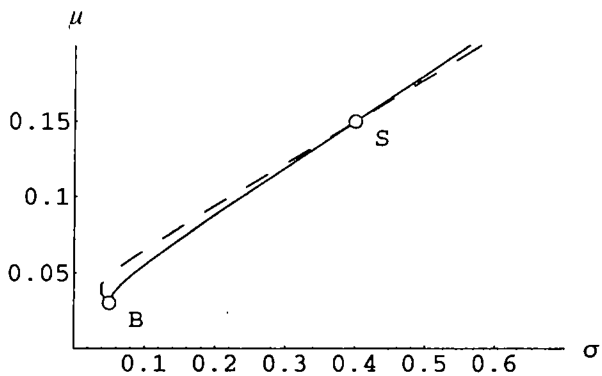


FIGURE 9: Efficient Frontiers – Two Fund Theorem.

Now we add to the investment opportunity set two new securities. The first is a risk-free bond. It has return r_f, zero variance, and zero covariance with every other security. Every investor is better off (or no worse off) as a result of this expanded opportunity set. This is illustrated by the "one fund theorem" described by Luenberger [15, page 168]. There is an efficient portfolio M of risky assets with weights $w_M^{\mathrm{T}} = [w_1,\ \ldots,\ w_n]$ such that any efficient portfolio can be constructed as a combination of M and the risk-free bond. This is represented graphically in Figure 10. The equation of the line is

$$r = r_f + \frac{r_M - r_f}{\sigma_M}\sigma.$$

This is called the capital market line (CML).

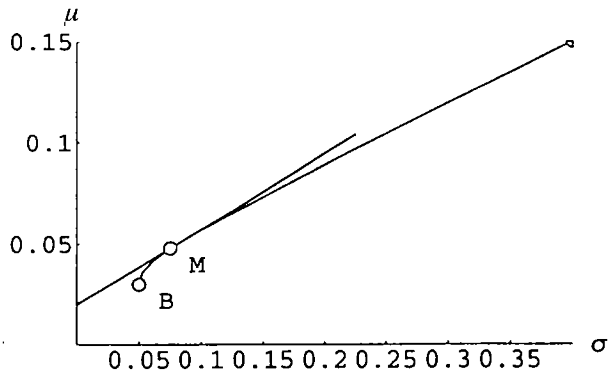


FIGURE 10: Efficient Frontiers.

As before, the efficient frontier before introducing the risk-free bond is the curved line. Any point (σ, r) on the CML can be obtained by investing the proportion $a = (\sigma_M - \sigma)\sigma_M^{-1}$ in the risk-free bond and $1 - a$ in the fund M. The capital market line lies above the original efficient frontier, except at M where they are equal. All investors hold a portfolio of the form $ar_f + (1 - a)r_M$ for some a, given this opportunity set. That is, all mean-variance optimizing investors will demand a portfolio on the capital market line [1]. Luenberger shows how to solve for the weights defining the portfolio M, which we will refer to as the *market portfolio*.

Now we introduce an insurance-based security C with high expected return, correspondingly high variance, but relatively low correlation with other risky assets. C could be a cat bond. At least for the case that the underlying insurance risk is catastrophe property loss, there is evidence that the return has zero correlation with the market portfolio M [5, 13]. The new asset has risk and return parameters σ_C and r_C and its correlation with the market $\rho_{C,M} = \dfrac{\mathrm{Cov}(R_C,\ R_M)}{\sigma_C \sigma_M}$ is relatively small.

The portfolio returns obtained as linear combinations of R_M and R_C

$$R_a = aR_M + (1 - a)R_C$$

correspond to points $(\sigma_a,\ r_a)$ where

$$r_a = a\mathrm{E}[R_M] + (1 - a)\mathrm{E}[R_C] = ar_M + (1 - a)r_C$$

and

$$\sigma_a^2 = a^2\sigma_M^2 + 2a(1 - a)\rho_{C,M}\sigma_C\sigma_M + (1 - a)^2\sigma_C^2.$$

---

[1] The one-fund and two-fund theorems are valid whether markets are complete or not. Individual investor risk preferences are reflected in the choice of the factor a, but they nevertheless choose positions on the CML.

The dashed curve in Figure 11 illustrates the graph of the parametric equations for the points $\{(\sigma_a, r_a)|0 \leq a \leq 1\}$ for the case that $\sigma_C > \sigma_F$ and $\mu_C > \mu_F$. The following argument shows that so long as $\rho_{C,M} < \sigma_M/\sigma_C$, the curve joining C and M has a negative slope at M (where $a = 1$) and so it punches through the CML. As a consequence the new CML, determined after investors take into account the new security must have a greater slope than the original. This means all investors are better off.

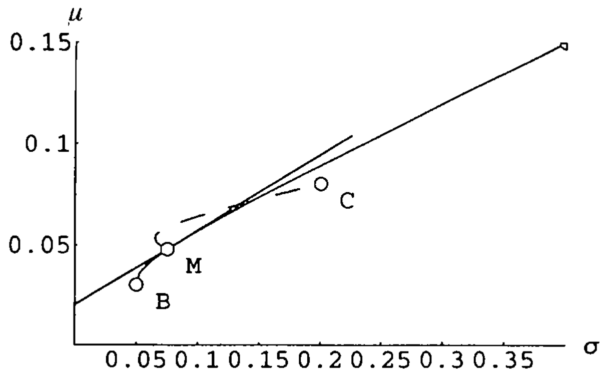


FIGURE 11: Efficient Frontiers.

In order for the curve to push up and to the left relative to the CML, it is sufficient that the slope of the curve at M be negative. Calculation of the slope goes like this:

$$\begin{aligned} 2\sigma_a \frac{\partial\sigma}{\partial a} &= \frac{\partial\sigma^2}{\partial a} \\ &= 2a\sigma_M^2 + 2(1-2a)\rho_{C,M}\sigma_M\sigma_C - 2(1-a)\sigma_C^2 \end{aligned}$$

For $a = 1$, we find that

$$\left.\frac{\partial\sigma}{\partial a}\right|_{a=1} = \sigma_M - \rho_{C,M}\sigma_C.$$

The slope of the curve at M, therefore, is

$$\frac{\dfrac{\partial r_a}{\partial a}}{\dfrac{\partial\sigma}{\partial a}} = \frac{r_M - r_C}{\sigma_M - \rho_{C,M}\sigma_C}.$$

In the case illustrated in Figure 11, $r_C > r_M$ and the slope is negative provided only that $\rho_{C,M} < \sigma_M/\sigma_C$. We think that this describes the recently observed market for cat bonds. The correlation does not have to be zero. All investors are better off when catastrophe-based securities are introduced.

In the case that $r_C < r_M$, adding C also expands the efficient frontier provided that the slope of the (C, M)-curve is positive at M. This leads to the same condition, $\rho_{C,M} < \sigma_M/\sigma_C$, on the correlation coefficient, as illustrated in Figure 12. Our conclusion is that adding a security with nonnegative relatively low correlation (or negative correlation of any magnitude) with the market results in a new market equilibrium in which all investors have improved opportunities.

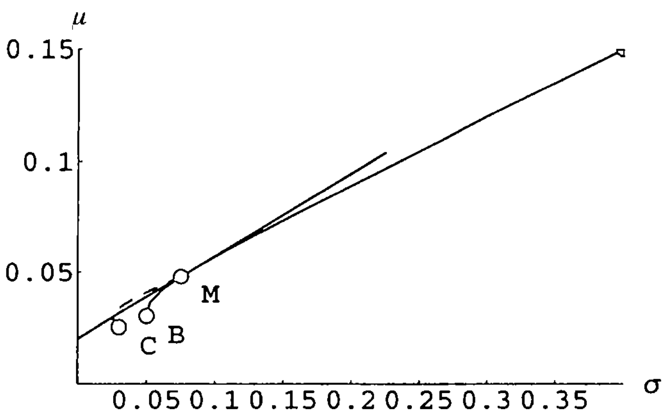


FIGURE 12: Efficient Frontiers.

We have shown that for the investment opportunities to improve it is sufficient that the covariance of the new security's returns with existing assets is relatively small in absolute value or negative. For example, it seems likely that long term bonds with coupons based on a mortality index would also improve investment opportunities, even if the risk and return were below equity levels. Thus the mean-variance model provides a rational for the demand for new insurance based securities. That is, all investors will now demand portfolios on the new capital market line [1]. The insurance press reports that investors (so far) like cat bonds. Some issues have been described as over-subscribed. This behavior seems to be consistent with the model.

In our construction we assumed the original opportunity set of n risky assets had an invertible covariance matrix, which means that no single asset is a linear combination of the other $n - 1$ assets. We assumed also that transaction costs were zero, all available information was revealed to all investors instantaneously, and other market imperfections such as taxes were not present. We call these the perfect market assumptions.

[1] This holds regardless of individual investor risk preferences.

In the usual construction, the original n risky assets contain firm specific risks that we have now assumed to be engineered into cat bonds. This risk is usually assumed to be costlessly diversified away. Given the assumptions on transactions costs, information, *etc.*, the equilibrium that obtains would not be altered by the introduction of cat bonds or other such securities. The introduction of such securities does not change the CML.

Our argument is that in actual, imperfect markets, the introduction of such securities results in the market being more efficient. Their introduction allows investors to construct portfolios consistent with their preferences for less costs. The more efficient distribution of capital over risks results in a new equilibrium in which all investors are better off.

Our construction was designed to show that by adding such securities the market is pushed closer to the idealized perfect markets equilibrium. This is done by increasing the present value of profits of the firm via this activity in ways investors cannot on their own account (increasing efficiency of the firm) and by packaging the risks (that are assumed to be in the original n risky securities) and issuing them to the market in such a way that investors can distribute capital over these risks more efficiently than they could when they were contained in the original n risky securities (due to increased efficiencies such as lower bid-ask spreads, information acquisition costs, and so on).

We argue that this is the economic justification for this activity and, correspondingly, it should continue to be observed in the capital markets as long as securitization improves efficiency.

6. The supply of insurance-based securities

Why do insurers and reinsurers securitize insurance risks? Capacity to handle very large losses is frequently mentioned as a motive for catastrophe risk securities [7, 13]. We note also that many of the catastrophe risk deals provide long term coverage, in contrast to traditional reinsurance which is normally issued for a one year term. What about other insurance risks? As we described earlier, there have been few securitizations of mortality risk. This makes sense economically. Securitization brings more capital to cover risks that would not be covered otherwise. There seems to be a need for even more capital as economies develop and more property is insured. Securitization of insurance risk is expensive compared to an asset securitization such as a T-bond securitization or traditional reinsurance. Some of the additional cost is due to costs of measuring the risks and explaining them to investors – resolving the moral hazard problem. However, we expect these costs will decline as investors become more familiar with the risks. Perhaps securitization will always be more expensive than reinsurance, but we expect it will continue to be used for these reasons:

Securitization often provides innovative contract terms such as larger amounts of coverage (catastrophe property risk), coverage of risks not provided by traditional reinsurance (long term mortality risks), or unusual risks.

Counter-party risk is eliminated with securitization.

Securitizations may provide more favorable tax treatment. The special purpose reinsurer is usually located in a jurisdiction which allows favorable tax treatment of reserves.

The question (why do insurers buy reinsurance?) is interesting because in an ideal world – one with no taxes, transactions costs, or other "imperfections" the insurance company shareholders would not compensate managers for managing a risk they can diversify on their own behalf in the capital markets.

For example consider the risk of fire damage to the corporation's property. A shareholder with X dollars invested in the corporation will suffer a loss if the property burns. However, the investor can find a second company and invest $X/2$ in each company. This diversifies the shareholder's fire risk. Further diversification reduces the fire (and other risks) even more. This does not cost shareholders anything, so they will direct managers to retain diversifiable risks, rather than insuring them. This suggests corporations should not buy insurance, yet they buy a lot. Mayers and Smith [17] offer answers that can be summarized as follows: real-world imperfections make insurance a rational corporate purchase.

The rationale for reinsurance purchases and securitizations of insurance risk is analogous. The demand for securitizations will persist as long as it has an advantage in addressing the imperfections we described earlier.

7. The Role of Actuaries in Securitization

So far actuaries have been on the sidelines with a few exceptions. Of course, actuaries were involved in Prudential's securitization of policy loans and Skandia's securitization of premium loadings. In addition James Tilley developed the concept of a catastrophe risk bond [25] in connection with Morgan Stanley's effort to help fund the California Earthquake Authority. Prakash Shimpi is leading a Swiss Re subsidiary dedicated to trading insurance risks [23]. These are important developments, but we should see many more actuaries working in the field. The role of the actuary should go well beyond modeling loss distributions. The actuary has the skills to see the big picture as well as the technical details. There is an opportunity to contribute to contract design, security valuation, investor communications, etc.

8. Summary

There are four components of securitization. A retailer bundles customer risks and passes them as a group to a special purpose company. The special purpose company issues securities based on the pool. The process can be used to reallocate risk or rearrange cash flows to better suit the needs of investors. When applied to insurance risk, the process is costly but costs may decline to some extent but will likely remain more expensive than traditional reinsurance. The additional cost may be the price to be paid to overcome counter-party risk. The securitization business will grow since it provides access to large amounts of capital and it allows for innovative contracting, relative to traditional reinsurance. Introducing insurance-based securities into the capital market improves opportunities for all investors provided the underlying insurance risk is not correlated with existing market risk. This provides a rationale for the demand for such securities.

Appendix – On the Role of Incompleteness in Insurance Risk Securitization

In this appendix we attempt to give the reader a "feel" for the notions of completeness and incompleteness and why they are relevant and so important in insurance risk securitization. We begin by providing some general intuition on these concepts.

Some Intuition. It can be difficult to differentiate between insurance markets which are complete and those which are incomplete. Table 1 provides some examples of insurance products each with its embedded insurance risk and the type of market (complete or incomplete) which the product resides in.

TABLE 1

SOME EXAMPLES OF INSURANCE PRODUCTS AND THE TYPE OF MARKET THEY RESIDE IN

Insurance Product	*Nature of Risk*	*Market Type*
Variable Annuities	Mortality Risk	Complete
Catastrophe Risk Bonds	Catastrophe Risk	Incomplete
Mortality Risk Bonds	Mortality Risk	Incomplete
Equity-Indexed Annuities	Market Risk	Complete

We offer a brief rationale why each of these products resides in a complete or incomplete market.

Variable Annuity: We shall assume that the primary risk in issuing variable annuities is that a contract holder dies during the contract period and the insurance company must honor the minimum investment return guarantee. Since the investments are made in standard securities such as S&P 500 index funds, providing the mortality of the contract pool follows a deterministic mortality pattern, this investment risk can be fully hedged. Investors who purchase variable annuities are generally not purchasing portfolio diversification. Instead, they are seeking tax advantages and protection of principal. These products offer little additional diversification of risk as the assets they are invested in are already available in the market.

Catastrophe Risk Bonds: The primary risk in cat bonds is the occurrence of a catastrophe that triggers the loss of principle. Since there are no securities, other than cat bonds, whose payoffs are contingent on the occurrence of catastrophes, cat bonds cannot be priced in terms of a portfolio of the assets that are already traded and priced in the market. Therefore, cat bonds reside in an incomplete market. Furthermore, cat bonds provide investors diversification of risk because the payoffs from these bonds are contingent on states that are not picked up by existing securities.

Mortality Risk Bonds: When mortality is assumed to follow a deterministic life table, the payments from traditional insurance products follow a fixed and known pattern. When deterministic mortality is assumed, even life insurance products whose benefits are contingent on the stock market or interest rates reside in a complete market. However, if there are substantial fluctuations in mortality experience across all policies then this risk cannot be hedged using existing securities because there are no existing securities whose payoffs are contingent on the mortality fluctuations. Consequently, mortality risk bonds reside in an incomplete market.

Equity-Indexed Annuities: Equity-indexed annuities have characteristics similar to variable annuities. For this reason, they too reside in a complete market.

A Simple One-Period Example. We now consider the concepts of complete and incomplete markets in the context of a simple model involving ordinary default-free bonds.

Let us consider a single-period model in which two bonds are available for trading, one of which is a one-period bond and the other a two-period bond. For convenience we shall assume that both bonds are zero coupon bonds. We further assume that the financial markets will evolve to one of two states at the end of the period, "interest rates go up" or "interest rates go down" and that the price of each bond will assume to behave according to the binomial model depicted in Figure 13.

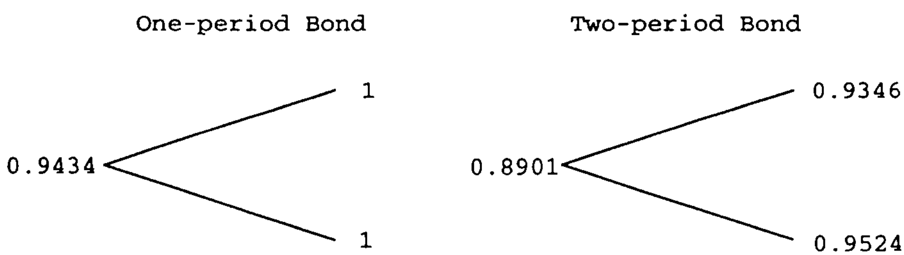


FIGURE 13: Payoffs from one-period and two-period bonds.

The bond prices for this model could be derived from a specification of risk-neutral probabilities and one-period rates but for simplicity we merely display these prices and note that we obtained them from an arbitrage-free model.

Suppose that we select a portfolio of the one-period and two-period bonds. Let us denote the number of one-period bonds held in this portfolio by n_1 and the number of two-period bonds held in this portfolio by n_2. This portfolio will have a value in each of the two states at time 1. Let us represent the state dependent price of each bond at time 1 using a column vector. Then we may represent the value of our portfolio at time 1 by the following matrix equation.

$$\begin{bmatrix} 1 & 0.9346 \\ 1 & 0.9524 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{2}$$

The cost of this portfolio is given by

$$0.9434n_1 + 0.8901n_2. \tag{3}$$

The 2×2 matrix of bond prices at time 1 appearing in equation (2) is nonsingular. Therefore, any vector of cash flows at time 1 may be generated by forming the appropriate portfolio of these two bonds. For instance, if we want the vector of cash flows at time 1 given by the column vector,

$$\begin{bmatrix} c^u \\ c^d \end{bmatrix} \tag{4}$$

then we form the portfolio

$$\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.9346 \\ 1 & 0.9524 \end{bmatrix}^{-1} \begin{bmatrix} c^u \\ c^d \end{bmatrix}$$

at a cost of $0.9434n_1 + 0.8901n_2$. Carrying out the arithmetic, one finds that the price of each cash flow of the form (4) is given by [1] the expression

$$0.4717c^u + 0.4717c^d. \tag{5}$$

Since every such set of cash flows at time 1 can be obtained and priced in the model we say that the one-period model is *complete*. The notion of pricing in this complete model is justified by the fact that the price we assign to each uncertain cash flow stream is exactly equal to the price of the portfolio of one-period and two-period bonds that generates the value of the cash flow stream at time 1.

Let us see how the model changes when catastrophe risk exposure is incorporated as part of the information structure. Suppose that we have the

---

[1] With the rounding errors introduced in the calculations the reader who performs this calculation will probably obtain the expression $0.4716c^u + 0.4718c^d$ instead.

framework of the previous model with the addition of catastrophe risk. Furthermore, let us suppose that the catastrophic event occurs independently of the underlying financial market variables. Therefore, there will be four states in the model which we may identify as follows.

$$\begin{aligned}
&\{\text{interest rate goes up, catastrophe occurs}\} &\equiv\ &\{u,+\}\\
&\{\text{interest rate goes up, no catastrophe occurs}\} &\equiv\ &\{u,-\}\\
&\{\text{interest rate goes down, catastrophe occurs}\} &\equiv\ &\{d,+\}\\
&\{\text{interest rate goes down, no catastrophe occurs}\} &\equiv\ &\{d,-\}
\end{aligned} \tag{6}$$

The reader will note that the symbol $\{u,+\}$ is shorthand for "interest rates go up" and "catastrophe occurs" and so forth. This information structure is represented on a single-period tree with four branches such as is shown in Figure 14.

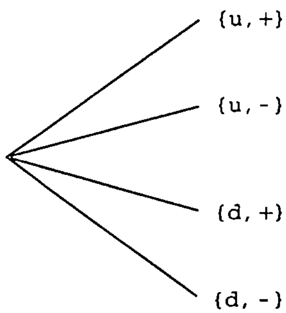


FIGURE 14: Revised states of the world with the introduction of catastrophe risk.

The values at time 1 of the one-period bond and the two-period bond are not linked to the occurrence or nonoccurrence of the catastrophic event and therefore do not depend on the catastrophic risk variable. We may represent the prices of the one-period and two-period bond in the extended model as shown in Figure 15.

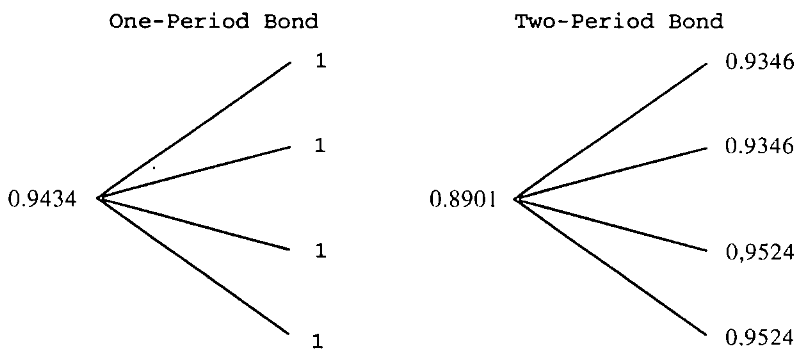


FIGURE 15: Payoffs from one-period and two-period bonds when states involving catastrophe risk are introduced into the model.

In contrast to equation (2), the state contingent payoffs at time 1 of a portfolio of the one-period and two-period bonds is now given by the following matrix equation.

$$\begin{bmatrix} 1 & 0.9346 \\ 1 & 0.9346 \\ 1 & 0.9524 \\ 1 & 0.9524 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \tag{7}$$

The cost of this portfolio is still given by $0.9434n_1 + 0.8901n_2$. The most general vector of cash flows at time 1 in this model is of the following form:

$$\begin{bmatrix} c^{u,+} \\ c^{u,-} \\ c^{d,+} \\ c^{d,-} \end{bmatrix} \tag{8}$$

On reviewing equation (7) we see that the span of the assets available for trading in the model [*i.e.* the one-period and two-period bonds] are not sufficient to span all cash flows of the form (8). Consequently, we cannot derive a pricing relation such as (5) that is valid for all cash flow vectors of the form (8). The best we can do is to obtain bounds on the price of a general cash flow vector so that its price is consistent with the absence of arbitrage. This can be done using state price vectors and these calculations and some examples may be found in [4, Chapter 5].

Pricing in Incomplete Markets. We will offer only the briefest of indications on how one can price insurance risk securitizations when working in incomplete markets. Details may be found in [4, Chapter 4] and references cited therein.

The benchmark financial economics technique used to price uncertain cash flow streams in an incomplete markets setting is the *representative agent*. The representative agent technique consists of an assumed representative utility function and an aggregate consumption process. Let us suppose that we are in a T-period economy in which agents can make choices and consume each period. The agent makes choices about his future consumption, represented by the stochastic process $\{c(k) : k = 0, 1, ..., T\}$. The aggregate consumption process may be thought of as the total consumption available in the economy (for all agents) at each point in time and in each state of the world. Let us denote the aggregate consumption stochastic process by $\{C^*(k)|k = 0, 1, ..., T\}$. Only the first choice is known with certainty at time $k = 0$. The other choices at future times are random and depend on the random state prevailing when each time point is reached. In "simple" applications it is customarily assumed that the representative agent's utility is time-additive and separable as well as differentiable. Time-additive and separable means that there are utility functions u_0, u_1, ..., u_T such that the agent's expected utility for a generic consumption process $\{c(k)|k = 0, 1, ..., T\}$ is given by

$$\mathrm{E}\left[\sum_{k=0}^{T} u_k(c(k))\right]. \tag{9}$$

It follows from the theory of the representative agent that the price, which we will denote $V(c)$, of a generic future cash flow process $\{c(k)|k = 1, ..., T\}$ at time 0 is given by the expectation

$$V(c) = \mathrm{E}\left[\sum_{k=1}^{T} \frac{u_k'(C^*(k))}{u_0'(C^*(0))} c(k)\right]. \tag{10}$$

If the aggregate consumption process in (10) is known (or can be determined for the model that is being used) then this equation gives a linear pricing relation for all uncertain consumption (or cash flow streams) for the model. This is very much like the risk-neutral expectation that occurs in complete markets valuation but here the model has been "closed" with an explicit assumption on utility. Evidently, different choices of utility functions will generally result in different pricing relations. Note that the aggregate consumption process plays a role in the pricing relation. In many implementations of this pricing relation the aggregate consumption process is assumed to evolve according to an exogenous process.

Acknowledgement

The authors are grateful for very helpful reviews from two referees.

References

1. Adams, C. and Authers, J. (1998) Fresh sale of catastrophe securities. *Financial Times* **1**.
2. Anonymous (1997). CSFB, Winterthur storm Swiss CB sector with first catastrophe linked bonds. *Euroweek*, no. January 17, 5.
3. Bodie, Z., Kane, A. and Marcus, A.J., (eds.) (1996) *Investments*, third ed., Richard D. Irwin, Chicago.
4. Boyle, P.P., Cox, S.H., Dufresne, D., Gerber, H.U., Mueller, H.H., Panjer, H.H., (Editor), Pedersen, H.W., Pliska, S.R., Sherris, M., Shiu, E.S. and Tan, K.S. (1998) *Financial economics with applications to investments, insurance and pensions*. The Actuarial Foundation, Schaumburg, Il 60173-2226.
5. Cantor, M.S., Cole, J.B. and Sandor, R.L. (1997) A new asset class for the capital markets and a new hedging tool for the insurance industry. *Journal of Applied Corporate Finance* **10**, no. 3, 69-83.
6. Connolly, J. (1997) Insurers eye American Skandia's VA securitization. *National Underwriter* January 13, p. 1.
7. Coppack, L. (1998) Capital markets set to take on bidder role. *Financial Times* **1**.
8. Cox S.H. and Schwebach, R. (1992) Insurance futures and hedging insurance price risk. *Journal of Risk and Insurance LIX* **4**, 628-644.

9. Cox, S.H., Fairchild, J.R. and Pedersen, H.W. (1998) Financial economics of securitization and alternative risk transfers, Working paper, Georgia State University, Center for Risk Management and Insurance Research submitted to *Insurance: Mathematics and Insurance.*
10. D'Arcy, S.P. and France, V.G. (1992) Catastrophe futures: A better hedge for insurers. *Journal of Risk and Insurance LIX* 4, 565-601.
11. D'Arcy, S.P. and France, V.G. (1993) Catastrophe insurance futures. *Journal of the Society of CPCU*, 202-213.
12. Embrechts, P. and Meister, S. (1995) *Pricing insurance derivatives, the case of cat futures.* Securitization of Insurance Risk (Schaumburg, Illinois) Society of Actuaries Monograph Series, May 1995, Bowles Symposium, Georgia State University.
13. Froot, K.A., Murphy, B.S., Stern, A.B. and Usher, S. (1995) *The emerging asset class: insurance risk.* Securitization of Insurance Risk (Schaumburg, Illinois) Society of Actuaries Monograph Series, May 1995, Bowles Symposium, Georgia State University.
14. Lane, M. (1995) *The perfume of the premium ... or pricing insurance derivatives.* Securitization of Insurance Risk (Schaumburg, Illinois), Society of Actuaries Monograph Series, May 1995, Bowles Symposium, Georgia State University.
15. Luenberger, D.G. (1998) *Investment science.* Oxford University Press, Oxford.
16. Markowitz, H. (1959) *Portfolio selection: Efficient diversification of investments.* John Wiley & Sons, New York.
17. Mayers, D. and Smith, C.W. Jr. (1982) On the corporate demand for insurance. *Journal of Business* **55**, no. 2, 281-296.
18. Monroe, A. (1997) Natural disaster bonds don't pay off – yet. *Investment Dealers Digest* **63**, no. 11, 28.
19. Niedzielski, J. (1997) Securitization continues to evolve. *National Underwriter* **101**, no. 28, 19, 28.
20. Niedzielski, J. (1997) Swiss Re in cat bond mkt for Calif. *National Underwriter* **101**, no. 11, 6.
21. Schmock, U. (1999) Estimating the value of the wincat coupons of the winterthur insurance convertible bond. *Astin Bulletin* **29**, no. 1, 101-163.
22. Scism, L. (1997) *Investors in USAA "disaster bonds" could get the idnd knocked out of them if storm strikes.* The Wall Street Journal, June 18, p. C21.
23. Shimpi, P. (1997) The context for trading insurance risk. *The Geneva Papers on Risk and Insurance* **22**, no. 82, 17-25.
24. Skipper, H. (1998) *International risk management*, Irwin, New York.
25. Tilley, J.A. (1995) *The latest in financial engineering: Structuring catastrophe reinsurance as a high-yield bond.* Tech. report, Morgan Stanley, New York, October 1995.
26. Tilley, J.A. (1997) *The securitization of catastrophic property risks.* ASTIN Proceedings, vol. XXVIII, ASTIN-AFIR Joint Day Proceeding, 13 August 1997, Cairns, Australia, pp. 499-549.
27. Zolkos, R. (1997) *Hurricane bond issue takes market by storm.* Business Insurance, June 17, 1997, p. 1.
28. Zolkos, R. (1997) *Swiss Re issues cat bond.* Business Insurance, July 28, 1997, pp. 2, 21.
29. Zolkos, R. (1998) *Aon storms markets with typhoon cat bond.* Business Insurance, June 29, pp. 3, 7.

Samuel H. Cox and Hal W. Pedersen
Actuarial Science Program,
Department of Risk Management and Insurance,
Georgia State University,
Atlanta, Georgia, 30302 USA
E-mail address, Samuel H. Cox: samcox@gsu.edu
E-mail address, Hal W. Pedersen: inshwp@panther.gsu.edu

JOSEPH R. FAIRCHILD
Zürich Centre Group,
One Chase Manhattan Plaza, 39th Floor,
New York, NY 10005
E-mail address: joe.fairchild@zurich.com

PORTFOLIO OPTIMIZATION

By

René Schnieper

Zurich Insurance Company, Reinsurance

Keywords

Reinsurance, retentions, non linear optimization, insurance risk, financial risk, Markowitz's portfolio selection method, CAPM.

Abstract

Based on the profit and loss account of an insurance company we derive a probabilistic model for the financial result of the company, thereby both assets and liabilities are marked to market. We thus focus on the economic value of the company.

We first analyse the underwriting risk of the company. The maximization of the risk return ratio of the company is derived as optimality criterion. It is shown how the risk return ratio of heterogeneous portfolios or of catastrophe exposed portfolios can be dramatically improved through reinsurance. The improvement of the risk return ratio through portfolio diversification is also analysed.

In section 3 of the paper we analyse the loss reserve risk of the company. It is shown that this risk consists of a loss reserve development risk and of a yield curve risk which stems from the discounting of the loss reserves. This latter risk can be fully hedged through asset liability matching.

In section 4 we derive our general model. The portfolio of the company consists of a portfolio of insurance risks and of a portfolio of financial risks. Our model allows for a simultaneous optimization of both portfolios of risks. A theorem is derived which gives the optimal retention policy of the company together with its optimal asset allocation.

Some of the material presented in this paper is taken from Schnieper, 1997. It has been repeated here in order to make this article self contained.

TABLE OF CONTENTS

1. INTRODUCTION

The **profit and loss account** of an insurance company typically details the following income items:
- earned premiums (net of premiums for outwards reinsurance),
- investment income,
- realized capital gains,

and the following expenditure positions:
- incurred claims (net of reinsurance recoveries),
- expenses,
- dividends to policyholders,
- dividends to shareholders.

We assume that the accounts of the company are on an accident year basis. Any other commonly used basis (e.g. underwriting year) can be dealt with after some minimal changes. We shall some times refer to the financial year which is the period covered by the company's accounts.

We split the premium into its different components;
- pure risk premium,
- loading for expenses,
- loading for profit.

We split incurred claims into the following two components:
- incurred claims pertaining to the current accident year
- changes in claim amounts in respect of claims pertaining to previous accident years.

We also take unrealized capital gains into account as an income item.

We make the following simplifying assumptions:
- expenses and loading for expenses are identical and therefore cancel out;
- dividends to policyholders are accounted for as claims,
- we are interested in the change in value of the surplus of the company before dividend to shareholders. We therefore ignore this item,
- the period under consideration is the financial year of the company. This is an arbitrary assumption. We could take any other period e.g. a quarter or a multi year period corresponding to the planing horizon of the company,
- payments pertaining to a given period are made at the end of the period,
- the premium written in a given period is earned in that period, i.e. the company has no unearned premium reserves. (This assumption can be dropped at the cost of a slight increase in the model complexity. The interest rate risk pertaining to the unearned premium reserves would be treated in a similar way as the interest rate risk pertaining to the loss reserves. Since the former is much less material than the latter, we have chosen to ignore it.)

We make the following model **assumptions**:
1. All random variables appearing in the model have finite second order moments.
2. The pure risk premium is the present value of the expected loss payments.
3. The loss reserves are equal to the present values of expected future loss payments.
4. The discount factors used to assess the pure risk premium and the loss reserves are based on the yield curve as defined by the bond market.
5. The assets of the company are valued at market value.

We introduce the following notation, where random variables are denoted by a tilde:

$\tilde{S}$	total claims amount pertaining to the current accident year
$E(\tilde{S})$	the mathematical expectation of the above random variable; this is the pure risk premium
ℓ	the profit loading for assuming the underwriting risk $\tilde{S}$
$\tilde{\Delta L}$	increase in claim amounts in respect of claims pertaining to previous accident years
$\tilde{\Delta A}$	investment income plus realized capital gains plus unrealized capital gains
u	capital (economic value) of the company at the beginning of the financial year
$\tilde{\Delta u}$	increase in capital (in economic value) during the financial year, return of the company during the financial year.

The following relation holds true

$$\tilde{\Delta u} = E(\tilde{S}) + \ell - \tilde{S} - \tilde{\Delta L} + \tilde{\Delta A}$$

$\tilde{S} - E(\tilde{S})$ is referred to as the underwriting risk, $\tilde{\Delta L} - E(\tilde{\Delta L})$ as the loss reserve risk, $\tilde{\Delta A} - E(\tilde{\Delta A})$ as the asset risk and $\tilde{\Delta u} - E(\tilde{\Delta u})$ as to the total risk of the company.

2. Underwriting Risk

2.1. Simplified Model

We split the assets of the company between a liability fund and a capital fund $A = A_L + A_U$. This means that some of the assets (A_L) are earmarked to cover the liabilities of the company and the rest of the assets (A_U) match the equity of the company. Since in this section we focus on the underwriting risk, we assume that there is no loss reserve risk and no asset risk. To be more specific, we make the following

Assumptions

- There is no loss reserve risk, i.e. amount and time of payment in respect of outstanding losses are perfectly known to the company.
- The liability fund, i.e. those assets which cover the liabilities, perfectly match the amounts and maturities of the liabilities. The liabilities are discounted with the discount factors corresponding to the liability fund. As a consequence any change in the yield curve will have a perfectly offsetting effect on $\tilde{\Delta}L$ and $-\tilde{\Delta}A_L$.
- The capital fund is invested in the risk free rate of return: $\tilde{\Delta}A_U = \rho_0 u$.

The total return of the company now is

$$\tilde{\Delta}u = E(\tilde{S}) + \ell - \tilde{S} - \tilde{\Delta}L + \tilde{\Delta}A_L + \tilde{\Delta}A_U = E(\tilde{S}) + \ell - \tilde{S} + \rho_0 u$$

2.2. Optimality Criterion

The objective of the present article is to provide a method to optimize the portfolio of the company. We first define and discuss the optimality criterion. The owners of the company are interested in the excess return on equity provided by the insurance portfolio

$$\tilde{\delta}(u) = \frac{\tilde{\Delta}u - \rho_0 u}{u} .$$

Let

$$E(\tilde{S}) - \ell - \tilde{S} = \sum_{i=1}^{m} E(\tilde{X}_i) + \ell_i - \tilde{X}_i$$

be a breakdown of the portfolio into m individual risks (policies, lines of business, customer segments, etc.). The company manages its portfolio by defining for each risk $\tilde{X}_i - E(\tilde{X}_i)$ the share $\alpha_i \in [0, 1]$ it wants to retain and by ceding $(1 - \alpha_i)(\tilde{X}_i - E(\tilde{X}_i))$ to its reinsurers. It is assumed that the company also cedes a proportional share of the corresponding profit $(1 - \alpha_i)\ell_i$ to its reinsurers. The return of the net retained portfolio is thus

$$\tilde{\Delta}u_{net} = \sum_{i=1}^{m} \alpha_i \big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big) + \rho_0 u$$

and the corresponding excess return on equity is

$$\tilde{\delta}_\alpha(u) = \frac{\tilde{\Delta}u_{net} - \rho_0 u}{u} = \sum_{i=1}^{m} \alpha_i \frac{E(\tilde{X}_i) + \ell_i - \tilde{X}_i}{u}$$

We introduce the following notation

$$\mu_\alpha(u) = E(\tilde{\delta}_\alpha(u)) \qquad \sigma^2_\alpha(u) = Var(\tilde{\delta}_\alpha(u)).$$

We have now to define the criterion according to which the company optimizes its portfolio. The approach is the same as Markowitz's mean variance method. (See H. Panjer et al., 1998.) It is assumed that the owners of the company have two objectives:

- maximization of the expected value $\mu_\alpha(u)$ of the company return on equity
- minimization of the risk as measured by $\sigma^2_\alpha(u)$.

According to their preferences, the owners put weights on these conflicting objectives and maximize

$$2\tau\mu_\alpha(u) - \sigma^2_\alpha(u), \qquad \text{with } \tau \geq 0.$$

The parameter τ is called the **risk tolerance**.

Note that the total investment constraint of the Markowitz Model $\left(\sum_{i=1}^{m} \alpha_i = 1\right)$ is meaningless in the present framework and has been dropped.

We first assume that the amount of equity of the company, u is given. The set of all points in the (μ, σ) diagram, which correspond to efficient portfolios is called the **efficient frontier**. The efficient frontier is convex, and piecewise hyperbolic. Because there exists a riskless investment, the first piece of the efficient frontier is linear. (See H. Panjer et al., 1998.)

Example

We assume that there are two uncorrelated risks with expected profit ℓ_1 and ℓ_2 respectively and standard deviation σ_1 and σ_2 respectively. We introduce the following notation

$$\lambda_i = \frac{\ell_i}{u} \quad \text{and} \quad \tau_i = \frac{\sigma_i}{u} \quad i = 1,\ 2.$$

We have

$$\mu_\alpha(u) = \alpha_1\lambda_1 + \alpha_2\lambda_2$$
$$\sigma^2_\alpha(u) = \alpha_1^2\tau_1^2 + \alpha_2^2\tau_2^2$$

The objective is

$$2\tau\mu_\alpha(u) - \sigma^2_\alpha(u) = \max_{\underline{\alpha}\in\beta}! \quad \text{with} \quad \beta = \{\underline{\alpha} | \alpha_1, \alpha_2 \in [0, 1]\}$$

which leads to the following unconstrained optimum

$$\alpha_i = \tau\frac{\lambda_i}{\tau_i^2} \qquad i = 1,\ 2.$$

Without any loss of generality we assume

$$\frac{\lambda_1}{\tau_1^2} \geq \frac{\lambda_2}{\tau_2^2}$$

and we make the following case distinction:

1. $\tau \leq \frac{\tau_1^2}{\lambda_1}$

   In that case α_1 and α_2 are as above and

   $$\mu_\alpha = \tau\left(\frac{\lambda_1}{\tau_1^2} + \frac{\lambda_2}{\tau_2^2}\right), \quad \sigma_\alpha^2 = \tau^2\left(\frac{\lambda_1^2}{\tau_1^2} + \frac{\lambda_2^2}{\tau_2^2}\right).$$

   Hence $(\mu_\alpha, \sigma_\alpha)$ describes a straight line as τ varies.

2. $\tau \in \left[\frac{\tau_1^2}{\lambda_1}, \frac{\tau_2^2}{\lambda_2}\right]$

   In that case $\alpha_1 = 1$ and $\alpha_2 = \tau\frac{\lambda_2}{\tau_2^2}$ and $\mu_\alpha = \lambda_1 + \tau\frac{\lambda_2^2}{\tau_2^2}$, $\sigma_\alpha^2 = \sigma_1^2 + \tau^2\frac{\lambda_2^2}{\tau_2^2}$

and $(\mu_\alpha, \sigma_\alpha)$ describes a hyperbole as τ varies.

3. $\tau \geq \frac{\tau_2^2}{\lambda_1}$

   In that case $\alpha_1 = \alpha_2 = 1$ and $\mu_\alpha = \lambda_1 + \lambda_2$, $\sigma_\alpha^2 = \sigma_1^2 + \sigma_2^2$ and this segment of the efficient frontier degenerates to a single point.

We now let the amount of equity of the company, u vary. We have

$$\mu_\alpha(u) = E\big(\tilde{\delta}_\alpha(u)\big) = \frac{\sum_{i=1}^{m} \alpha_i \ell_i}{u} = \frac{R(\underline{\alpha})}{u}$$

$$\sigma_\alpha^2(u) = Var\big(\tilde{\delta}_\alpha(u)\big) = \frac{\sum_{i,j} \alpha_i \alpha_j \sigma_{ij}^2}{u^2} = \frac{V(\underline{\alpha})}{u^2}$$

where $\sigma_{i,j}^2 = Cov(\tilde{X}_i, \tilde{X}_j)$. Hence $\mu_\alpha(u) = \frac{R(\underline{\alpha})}{u}$, $\sigma_\alpha(u) = \frac{V(\underline{\alpha})^{\frac{1}{2}}}{u}$.

Thus if P is a point on the efficient frontier as defined above – i.e. on the basis of a fixed amount of equity – any point on the straight line $0P$ can be reached through a proper choice of the amount of equity u. It is therefore natural to start the optimization process with the following requirement

1. $\frac{R(\underline{\alpha})}{(V(\underline{\alpha}))^{\frac{1}{2}}} = \max\limits_{\underline{\alpha}\in\beta}!$ with $\beta = \{\underline{\alpha} | \alpha_i \in [0, 1] \text{ all } i\}$

   the above requirement amounts to a maximization of the risk return ratio or, in the terminology of financial economics, of Sharpe's ratio. In general, the above ratio is maximized for a whole set of admissible values of α. Let β_1 denote the set of those values. It is reasonable to make the following additional requirement

$$\sum_{i=1}^{m} \alpha_i \ell_i = \max_{\underline{\alpha}\in\beta_1}!$$

This amounts to maximizing the net expected profit.
Let $\underline{\alpha}_M$ denote the net retentions for which the above requirement is satisfied. Let

$$R = R(\underline{\alpha}_M) \quad and \quad V = V(\underline{\alpha}_M).$$

The optimal amount of equity is now defined by the following requirement

2. $2\tau \dfrac{R}{u} - \dfrac{V}{u^2} = \max_u!$

which leads to the following optimal amount of equity

$$u = \tau^{-1} \frac{V}{R}$$

Remarks

1. Whilst the present optimization method is based on the same objective function as Markowitz's mean variance method, there are however major differences between the two methods. First, the portfolio to be optimized consists of a set of insurance risks rather than financial assets. (Later we shall optimize a combined portfolio of insurance risks and financial assets.) This leads to a different set of constraints. In particular the total investment constraint ($\sum \alpha_i = 1$) is meaningless and has been dropped. Second, in addition to optimizing the composition of the portfolio, the company can also decide on the amount of equity it needs to support the business. This additional degree of freedom leads to a different efficient frontier than in the Markowitz framework.
2. One of the drawbacks of the above method is that it only takes into account the first two moments of the distribution of the risks in the portfolio. In the case of insurance risks which are typically skewed and leptocurtic, this is a serious limitation. In the remainder of this section we shall nevertheless analyze a few insurance optimization problems with the help of the above method. It is felt that this parallel between insurance and finance is of interest in spite of the above mentioned limitations. Within the framework of our general model (introduced in section 4) we optimize a combined portfolio of insurance risks and of risky financial assets. Since the insurance risks entering into the portfolio are net of reinsurance, it is not unreasonable to assume that the distribution of returns is close to multivariate normal.

We now turn to the problem of allocating capital to individual risks. Let $\tilde{\Delta}u = \sum_{i=1}^{n} \tilde{X}_i$ be any split of the total risk of the company into individual risks. The capital is proportional to

$$Var(\tilde{\Delta}u) = \sum_{i=1}^{n} Cov(\tilde{X}_i, \tilde{\Delta}u)$$

It is thus fair to allocate to each risk $\tilde{X}_i$ an amount of capital u_i, which is proportional to the contribution of that risk to the overall volatility of the result of the company: $u_i = k \cdot Cov(\tilde{X}_i, \tilde{\Delta}u)$. Since $u = \sum_{i=1}^{n} u_i$ we obtain

$$u_i = u \cdot \frac{Cov(\tilde{X}_i, \tilde{\Delta}u)}{Var(\tilde{\Delta}u)}.$$

The excess return which the company expects to achieve for assuming the risk $\sigma(\tilde{\Delta}u)$ is equal to $(\rho - \rho_0)u$, where ρ_0 denotes the risk free rate of return. It is fair to split the excess return proportionally to the capital.

Definition
The **fair loading** of risk $\tilde{X}_i$ is

$$(\rho - \rho_0)u_i = (\rho - \rho_0) \cdot u \cdot \frac{Cov(\tilde{X}_i, \tilde{\Delta}u)}{Var(\tilde{\Delta}u)}$$

It is equal to the cost of the capital needed for assuming risks $\tilde{X}_i$.

We assume that the company is a price taker, the fair loading is thus not a way to compute prices but a way to define benchmarks. In general there will be cross-subsidies. Certain risks well have a higher expected profit than the fair loading, others will have a lower expected profit. Later we show that if the portfolio of risks is optimized in an unconstrained way, the actual loading of each risk is equal to the fair loading. This is a further justification for our way of allocating capital to individual risks.

We now turn to the problem of maximizing the underwriting risk return ratio. Assuming that the loadings of individual risks are given there are two main possibilities to increase the above ratio: combining risks in a portfolio and buying reinsurance. We illustrate the impact of reinsurance and the portfolio effect on the risk return ratio.

2.3. Portfolio Heterogeneity

Let $\tilde{X}_1$, $\tilde{X}_2$, ..., $\tilde{X}_n$ be the uncorrelated risks of a portfolio $\tilde{S} = \sum_{i=1}^{n} \tilde{X}_i$. Let ℓ_i denote the loading of risk i and σ_i^2 its variance. We have thus

$$\ell = \sum \ell_i \quad \text{and} \quad \sigma(\tilde{S}) = \left(\sum \sigma_i^2\right)^{\frac{1}{2}}.$$

Let us assume that for each individual risk i the company keeps a share α_i for its own account and cedes a share $(1-\alpha_i)$ to its reinsurers.

Theorem
Under the above assumptions, the choice of $\alpha_1, \ldots, \alpha_n$ which maximizes the net underwriting risk return ratio

$$r_{net} = \frac{\sum \alpha_i \ell_i}{(\sum \alpha_i^2 \sigma_i^2)^{\frac{1}{2}}}$$

is

$$\alpha_i = c \frac{\ell_i}{\sigma_i^2}$$

where c is some norming constant which must be chosen in such a way that

$$0 \leq \alpha_i \leq 1$$

for all i. With the so defined set of retentions we have

$$r_{net} = \left(\sum_i \frac{\ell_i^2}{\sigma_i^2} \right)^{\frac{1}{2}}$$

Proof
Deriving r_{net} with respect to α_j and setting the derivative equal to 0 we obtain

$$\frac{\ell_j(\sum \alpha_i^2\sigma_i^2)^{\frac{1}{2}} - (\sum \alpha_i \ell_i)(\sum \alpha_i^2 \sigma_i^2)^{-\frac{1}{2}} \alpha_j \sigma_j^2}{\sum \alpha_i^2 \sigma_i^2} = 0$$

$$\ell_j \left(\sum \alpha_i^2 \sigma_i^2 \right) = \left(\sum \alpha_i \ell_i \right) \alpha_j \sigma_j^2$$

$$\alpha_j = \frac{\ell_j}{\sigma_j^2} \frac{\sum \alpha_i^2 \sigma_i^2}{\sum \alpha_i \ell_i} = c \frac{\ell_j}{\sigma_j^2}$$

and the value of the optimal r_{net} is obtained by plugging the above value of α_j into the expression defining r_{net}. □

Special case
Let

$$\tilde{X}_i = \begin{cases} L_i & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p \end{cases}$$

and

$$\ell_i = E(\tilde{X}_i)\lambda = pL_i\lambda$$

we now have

$$Var(\tilde{X}_i) = p(1-p)L_i^2 \simeq pL_i^2 \quad \text{for} \quad p << 1$$

and the optimal retention becomes

$$\alpha_i = c\frac{\ell_i}{\sigma_i^2} \simeq c\frac{pL_i\lambda}{pL_i^2} = \frac{1}{L_i}c\lambda$$
$$\Rightarrow \alpha_i L_i = c\lambda$$

and the retention of each risk is such that the net monetary amount retained is the same for all risks i.e. the reinsurance arrangement which maximizes the underwriting risk return ratio is a surplus treaty, where the retention is equal to the smallest sum insured.

On a gross basis the risk return ratio is

$$r = \frac{\sum_{i=1}^{n} L_i p\lambda}{\left(\sum_{i=1}^{n} L_i^2 p\right)^{\frac{1}{2}}} = \lambda\sqrt{p}\frac{\sum_{i=1}^{n} L_i}{\left(\sum_{i=1}^{n} L_i^2\right)^{\frac{1}{2}}}$$

and on a net basis

$$r_{net} = \left(\sum_i \frac{\ell_i^2}{\sigma_i^2}\right)^{\frac{1}{2}} = \lambda\sqrt{pn}$$

It is seen that $r_{net} \geq r$. The inequality is strict unless all L_i's are equal.

Numerical Example

Let us assume that there are two types of risks

$$\tilde{X}_1 = \begin{cases} 1 & \text{with probability} \quad 10^{-3} \\ 0 & \text{with probability} \quad 0.999 \end{cases}$$

and

$$\tilde{X}_2 = \begin{cases} 100 & \text{with probability} \quad 10^{-3} \\ 0 & \text{with probability} \quad 0.999 \end{cases}$$

There are $n = 10^5$ risks of the first type, and $n = 10^3$ risks of the second type. The profit loading is $\lambda = 3\%$ of the pure risk premium. We have

$$\sigma(\tilde{S}) \simeq \sqrt{10^{-3}(10^5 + 10^7)} = 100.5, \quad \ell = 6.0, \quad r = 0.060$$

According to the above theorem, the reinsurance arrangement which maximizes the underwriting risk return ratio is a surplus treaty with a retention of 1. On a net basis we have

$$\sigma(\tilde{S}_{net}) \simeq \sqrt{10^{-3} \cdot (10^5 + 10^3)} = 10.05, \qquad \ell = 3.03, \qquad r = 0.301$$

The net underwriting risk return ratio is much higher than the gross.

2.4. Catastrophe Exposure

Let $\tilde{S} = \sum_{i=1}^{n} \tilde{X}_i$ be a portfolio of individual risks where each risk is the sum of an ordinary risk and of a catastrophe risk:

$$\tilde{X}_i = {}_o\tilde{X}_i + {}_c\tilde{X}_i.$$

We have thus

$$\tilde{S} = \sum_{i=1}^{n} {}_o\tilde{X}_i + \sum_{i=1}^{n} {}_c\tilde{X}_i.$$

It is further assumed that

$$Cov({}_o\tilde{X}_i, {}_o\tilde{X}_j) = \delta_{ij}\sigma_0^2 \quad \text{for all } i, j, \text{ where } \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

and that

$$Cov({}_c\tilde{X}_i, {}_c\tilde{X}_j) = \sigma_c^2 \quad \text{for all } i, j$$

i.e. ordinary risks are uncorrelated and catastrophe risks are perfectly correlated. It is further assumed that

$$Cov({}_o\tilde{X}_i, {}_c\tilde{X}_j) = 0 \quad \text{for all } i, j.$$

It follows that

$$Cov(\tilde{X}_i, \tilde{X}_j) = Cov({}_o\tilde{X}_i + {}_c\tilde{X}_i,\ {}_o\tilde{X}_j + {}_c\tilde{X}_j) = \delta_{ij}\sigma_0^2 + \sigma_c^2$$

and

$$Var(\tilde{S}) = n\sigma_0^2 + n^2\sigma_c^2.$$

Let us now assume that the catastrophe exposure is reinsured through a per event excess of loss reinsurance with retention x

$$S_{net} = \sum_{i=1}^{n} {}_0\tilde{X}_i + \left(\sum_{i=1}^{n} {}_c\tilde{X}_i\right) \wedge x$$

where $x \wedge y$ denotes the minimum of x and y.

To compute the value of

$$\left(\sum_{i=1}^{n} {}_c\tilde{X}_i\right) \wedge x$$

as a function of x we would need to make distributional assumptions on the catastrophe risk. We make the extreme assumption that the catastrophe risk is fully reinsured, i.e. $x = 0$.

As a consequence we have

$$Var(\tilde{S}_{net}) = n\sigma_0^2.$$

Let μ_o and μ_c denote the pure risk premium of an ordinary risk and of a catastrophe risk respectively. Let λ_o and λ_c denote the premium loading of an ordinary risk and of a catastrophe risk respectively. We have

$$r = \frac{\ell}{\sigma(\tilde{S})} = \frac{n(\mu_0\lambda_0 + \mu_c\lambda_c)}{(n\sigma_0^2 + n^2\sigma_c^2)^{\frac{1}{2}}} = \frac{\mu_0\lambda_0 + \mu_c\lambda_c}{\left(\frac{\sigma_0^2}{n} + \sigma_c\right)^{\frac{1}{2}}}$$

Assuming that the loading of the reinsurance premium for the catastrophe risk is the same loading as for the original catastrophe risk, we obtain

$$r_{net} = \sqrt{n}\frac{\mu_0\lambda_0}{\sigma_0}$$

which is usually much larger than r.

Numerical Example

$${}_0\tilde{X}_i = \begin{cases} 100 & \text{with probability} \quad 10^{-3} \\ 0 & \text{with probability} \quad 0.999 \end{cases}$$

$${}_c\tilde{X}_i = \begin{cases} 5 & \text{with probability} \quad 10^{-3} \\ 0 & \text{with probability} \quad 0.99 \end{cases}$$

${}_0\tilde{X}_i$ could be a fire claim and ${}_c\tilde{X}_i$ an earthquake claims from a given fire policy.

We have

$$\mu_0 = 0.1,\ \mu_c = 0.05,\ \sigma_0 \simeq 10^{-\frac{3}{2}} \cdot 100 = 3.16,\ \sigma_c \simeq 10^{-1} \cdot 5 = 0.5$$

Let us assume that

$$\lambda_0 = 5\%,\ \lambda_c = 20\% \text{ and } n = 10^5.$$

We obtain

$$\sigma(\tilde{S}) = 50'010 \qquad \ell = 1'500 \qquad r = 0.030$$
$$\sigma(\tilde{S}_{net}) = 1'000 \qquad \ell_{net} = 500 \qquad r_{net} = 0.500$$

The net underwriting risk return ratio is much higher than the gross. Assuming $\tau = 0.25$ we obtain the following amount of required equity

$$u = \tau^{-1}\frac{\sigma^2(\tilde{S}_{net})}{\ell_{net}} = 8'000$$

which leads to the following optimal risk, excess return pair

$$\mu = \frac{\ell_{net}}{u} = 6.25\%, \quad \sigma = \frac{\sigma(\tilde{S}_{net})}{u} = 12.5\%.$$

2.5. Portfolio Diversification

Let $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n$ denote the different insurance portfolios of our company (e.g. homeowners, private automobile, commercial multiperil, commercial automobile, assumed reinsurance business, etc.).

Let

$$\pi(\tilde{X}_i) = E(\tilde{X}_i) + \ell_i$$

denote the premium of portfolio $\tilde{X}_i$, ℓ_i is thus the corresponding loading.

We use the following notation

$$\sigma_{ij} = Cov(\tilde{X}_i, \tilde{X}_j) \quad \sum = (\sigma_{ij})$$

We assume that the company keeps a share α_i of portfolio $\tilde{X}_i$ for own account and cedes a share $(1 - \alpha_i)$ to its reinsurers.

The combined net portfolio of the company is thus

$$\tilde{S}_{net} = \alpha_1\tilde{X}_1 + \alpha_2\tilde{X}_2 + ... + \alpha_n\tilde{X}_n$$

and its combined net profit loading is

$$\ell_{net} = \alpha_1\ell_1 + \alpha_2\ell_2 + ... + \alpha_n\ell_n$$

Theorem

We assume that $\sum^{-1}$ exists.

1. The vector $\underline{\alpha}' = (\alpha_1, \alpha_2, ..., \alpha_n)$ which maximizes the net underwriting risk return ratio

$$r_{net} = \frac{\ell_{net}}{\sigma(\tilde{S}_{net})}$$

is given by

$$\underline{\alpha} = c \cdot \Sigma^{-1} \cdot \underline{\ell}$$

where $\underline{\ell}' = (\ell_1, \ell_2, \ldots, \ell_n)$ and c is a scalar which is chosen in such a way that $\max\limits_{i=1,\ldots,n} \alpha_i = 1$.

The optimal risk return ratio is equal to

$$r_{net} = (\underline{\ell}'\Sigma^{-1}\underline{\ell})^{\frac{1}{2}}$$

2. $\underline{\alpha}$ maximizes the risk return ratio if and only if the net loadings $(\alpha_i\ell_i \quad i = 1, \ldots, n)$ are equal to the fair loadings.

Remark

The solution $\underline{\alpha}$ provided by the theorem is only meaningful if $\alpha_i \geq 0$ for all i. It is indeed unrealistic to assume that the company can take a short position in any of the insurance portfolios $\tilde{X}_i$. To find a solution $\underline{\alpha}$ which always satisfies the condition $\underline{\alpha} \geq 0$ is a convex optimization problem with restrictions. It is a standard problem in finance theory, see for instance W.F. Sharpe (1970).

Proof

1. We have to maximize the following expression

$$r = \frac{\alpha_1\ell_1 + \alpha_2\ell_2 + \ldots + \alpha_n\ell_n}{\left(\sum\limits_{i,j} \alpha_i\alpha_j\sigma_{ij}\right)^{\frac{1}{2}}}$$

deriving with respect to $\alpha_1, \alpha_2, \ldots, \alpha_n$ and equating the expression to 0, we obtain

$$\frac{\delta r}{\delta \alpha_1} = \frac{\ell_1\sigma(\tilde{S}_{net}) - \ell_{net}\frac{1}{2}\sigma(\tilde{S}_{net})^{-1}\left(2\sum\limits_{j=1}^{n}\alpha_j\sigma_{1j}\right)}{\sigma^2(\tilde{S}_{net})} = 0$$

$$\vdots$$

$$\vdots$$

$$\frac{\delta r}{\delta \alpha_n} = \frac{\ell_n\sigma(\tilde{S}_{net}) - \ell_{net}\frac{1}{2}\sigma(\tilde{S}_{net})^{-1}\left(2\sum\limits_{j=1}^{n}\alpha_j\sigma_{nj}\right)}{\sigma^2(\tilde{S}_{net})} = 0$$

and after some straightforward rearrangement of terms

$$\ell_1 \sigma^2(\tilde{S}_{net}) = \ell_{net_j} \sum_{j=1}^{n} \alpha_j \sigma_{1j}$$

$$\vdots$$

$$\ell_n \sigma^2(\tilde{S}_{net}) = \ell_{net_j} \sum_{j=1}^{n} \alpha_j \sigma_{nj}$$

or in matrix notation

$$\underline{\ell} \frac{\sigma^2(\tilde{S}_{net})}{\ell_{net}} = \Sigma \underline{\alpha}$$

$$\underline{\alpha} = c \Sigma^{-1} \underline{\ell}$$

This proves the first part of the theorem. (Note that by definition $\underline{\alpha}$ is only defined up to a norming constant c.)
We now prove the statement about r_{net}.

$$Var(\tilde{S}) = \underline{\alpha}' \Sigma \underline{\alpha} = c^2 \underline{\ell}' \Sigma^{-1} \Sigma \Sigma^{-1} \underline{\ell} = (c\underline{\ell}')(c\Sigma^{-1}\underline{\ell}) = c\underline{\ell}'\underline{\alpha}$$

$$r_{net} = \frac{\underline{\alpha}'\underline{\ell}}{\sqrt{c}(\underline{\alpha}'\underline{\ell})^{\frac{1}{2}}} = \frac{1}{\sqrt{c}}(\underline{\alpha}'\underline{\ell})^{\frac{1}{2}} = \frac{\sqrt{c}}{\sqrt{c}}(\underline{\ell}'\Sigma^{-1}\underline{\ell})^{\frac{1}{2}}$$

$$r_{net} = (\underline{\ell}'\Sigma^{-1}\underline{\ell})^{\frac{1}{2}}$$

2. $\alpha_i \ell_i \quad i = 1, \ldots, n$ are the fair loadings if and only if $\alpha_i \ell_i = c \cdot Cov(\alpha_i \tilde{X}_i, \tilde{S}_{net}) \quad i = 1, \ldots, n$ for some constant c. This in turn is equivalent with the following system of equations

$$\alpha_i \ell_i = c \cdot \sum_{j=1}^{n} \alpha_i \alpha_j \sigma_{ij} \quad i = 1, 2, \ldots, n$$

$$\ell_i = c \cdot \sum_{j=1}^{n} \sigma_{ij} \alpha_j \quad i = 1, 2, \ldots, n$$

$$\underline{\ell} = c \Sigma \underline{\alpha}$$

$$\underline{\alpha} = c^{-1} \Sigma^{-1} \underline{\ell}$$

which is equivalent with $\underline{\alpha}$ maximizing the risk return ratio.

q.e.d.

Numerical Example

There are three portfolios with

$$\sigma_{11} = 1 \qquad \ell_1 = 0.2 \qquad \Rightarrow \frac{\ell_1}{\sqrt{\sigma_{11}}} = 20\%$$

$$\sigma_{22} = 4 \qquad \ell_2 = 0.2 \qquad \Rightarrow \frac{\ell_2}{\sqrt{\sigma_{22}}} = 30\%$$

We think of $\tilde{X}_1$ and $\tilde{X}_2$ as of a motor portfolio and a homeowners portfolio respectively. We assume that both portfolios are exposed to the same natural peril (e.g. storm), which is only reinsured in excess of a substantial retention. The correlation between the two portfolios is therefore positive. Let us assume that it is equal to 0.20.

The third class of business consists of industrial risks with

$$\sigma_{33} = 9 \cdot (1.5)^2 = 20.25 \quad \ell_3 = 1.8 \quad \Rightarrow \frac{\ell_3}{\sqrt{\sigma_{33}}} = 40\%$$

The interpretation is that for the same premium income as the homeowners portfolio, the industrial portfolio has a standard deviation of 3, instead of 2 for the homeowners portfolio. The industrial portfolio has 50% more volume than the homeowners portfolio. It is assumed that the industrial portfolio and each of the personal lines portfolio are uncorrelated. We have thus

$$\sum = \begin{pmatrix} 1 & 0.4 & 0 \\ 0.4 & 4 & 0 \\ 0 & 0 & 20.25 \end{pmatrix}, \quad \underline{\ell} = \begin{pmatrix} 0.2 \\ 0.6 \\ 1.8 \end{pmatrix}$$

From our theorem we obtain that the optimal retentions are

$$\underline{\alpha}' = (1, \quad 0.93, \quad 0.61)$$

yielding

$$\sigma(\tilde{S}_{net}) = 3.57 \quad \underline{\ell}_{net} = 1.85 \quad r_{net} = 0.518$$

Thus the optimal risk return ratio is much higher than each of the risk return ratios of the individual classes.

Let $\tilde{S}$ be the gross combined portfolio $\tilde{S} = \tilde{X}_1 + \tilde{X}_2 + \tilde{X}_3$ we have

$$\sigma(\tilde{S}) = 5.10 \quad \ell = 2.60 \quad r = \frac{\ell_1 + \ell_2 + \ell_3}{\left(\sum_{i,j} \sigma_{ij}\right)^{\frac{1}{2}}} = \frac{2.6}{(26.05)^{\frac{1}{2}}} = 0.509$$

which is nearly as high as the optimal risk return ratio. To achieve the optimal ratio the company must cede 7% of its homeowners business and 39% of its industrial business. It must thus forgo an expected profit of 0.75 out of a total expected profit of 2.6. It is questionable whether in this case the slight improvement in the risk return ratio is worth this sacrifice.

Let us assume that for given $R = E(\Delta \tilde{u})$ and $V = Var(\Delta \tilde{u})$ the company chooses the amount of equity u in such a way as to maximize

$$2\tau \frac{R}{u} - \frac{V}{u^2}.$$

This is tantamount to utilizing Markowitz's objective function to determine the optimal amount of capital for a given risk and return. The optimal amount of equity is

$$u = \tau^{-1} \frac{V}{R} \, .$$

For $\tau = 0.25$ and utilizing the notation

$$\mu = E(\tilde{\delta}(u)), \quad \sigma = Var^{\frac{1}{2}}(\tilde{\delta}(u))$$

we obtain

Portfolio number	r	σ	μ	u
1	0.200	5%	1%	20.0
2	0.300	7.5%	2.25%	26.67
3	0.400	10%	4%	45.0
4	0.509	12.75%	6.5%	40.0
5	0.518	12.95%	6.71%	27.56

where portfolio number 4 is the combined portfolio and portfolio number 5 is the optimal portfolio.

This example illustrates that combining portfolios results in substantial capital savings and improvements of the risk return ratio. This example also illustrates the fact that, when we combine portfolios in a non optimal way, there is a cross subsidization between portfolios: Let $\tilde{S}$ denote the gross combined portfolio. The fair loadings are

$$\ell_i = \mu \cdot u \cdot \frac{Cov(\tilde{X}_i, \tilde{S})}{Var(\tilde{S})}$$

thus

$$\ell_1 = 6.5\% \cdot 40.0 \cdot \frac{1.4}{26.05} = 0.14 \quad \ell_2 = 0.44 \quad \ell_3 = 2.02$$

whereas the actual loadings are

$$\ell_1 = 0.20 \quad \ell_2 = 0.60 \quad \ell_3 = 1.80$$

There is a subsidization of $\tilde{X}_3$ from $\tilde{X}_1$ and $\tilde{X}_2$.

3. Loss Reserve Risk

3.1. Individual Accident Year

Since we only consider one accident year, we can assume that the development year t of risk $\tilde{X}$ is also the financial year t of the company. This amounts to a renumbering of the financial years. We first analyze the problem on an undiscounted basis. Later we introduce discounting.

Let $\tilde{X}$ denote a risk, or a portfolio of risks pertaining to a given accident year. Let $\pi(\tilde{X})$ and ℓ denote respectively the premium and the loading of risk $\tilde{X}$. We have

$$\pi(\tilde{X}) = E(\tilde{X}) + \ell.$$

As with all other random variables we assume that $E(\tilde{X}^2)$ is finite. Let us assume that $\tilde{X}$ is paid out over ω development years.

$$\tilde{X} = \sum_{t=1}^{\omega} \tilde{P}_t.$$

$\tilde{P}_t$ denotes the payment made in development year t in respect of risk $\tilde{X}$. Let $\mathcal{H}_t$ denote the information of the company on risk $\tilde{X}$ in development year t. $\mathcal{H}_0$ is the information on the risk prior to underwriting it and we have thus $E(\tilde{X}) = E(\tilde{X}|\mathcal{H}_0)$.

We further introduce the following notation

$$\tilde{X}_t = E(\tilde{X}|\mathcal{H}_t)$$

$\tilde{X}_t$ is the company's estimate of risk $\tilde{X}$ in development year t.

We assume that $\mathcal{H}_0, \mathcal{H}_1, ..., \mathcal{H}_t, ...$ is an increasing sequence of σ-algebras. It is easily seen that $\tilde{X}_t$ is a martingale. Let

$$\tilde{L}_t = E(\tilde{P}_{t+1} + \tilde{P}_{t+2} + ...|\mathcal{H}_t)$$

be the loss reserve of the company at the end of development year t in respect of risk $\tilde{X}$.

Based on the pure risk premium $E(\tilde{X})$, the contribution to results produced by risk $\tilde{X}$ in the successive development years are as follows

$$\tilde{R}_t = \tilde{L}_{t-1} - \tilde{P}_t - \tilde{L}_t \qquad t = 1, 2, ...$$

and the following relation holds true

$$\tilde{R}_t = E(\tilde{X}|\mathcal{H}_{t-1}) - E(\tilde{X}|\mathcal{H}_t) \qquad t = 1, 2, ...$$

$\tilde{R}_t$ is the difference process of a martingale (i.e. of $E(-\tilde{X}|\mathcal{H}_t)$).

Note that according to our terminology, $\tilde{R}_t$ is the underwriting risk and $\tilde{R}_2 + ... + \tilde{R}_\omega$ is the loss reserve risk.

Since $-\tilde{X}_t = -E(\tilde{X}|\mathcal{H}_t)$ is a martingale and $\tilde{R}_t$ is the corresponding difference process, the following holds true

$$E(\tilde{R}_t) = 0 \qquad t = 1,\, 2,\, ..., \qquad Cov(\tilde{R}_t, \tilde{R}_s) = 0 \qquad t \neq s,$$

$$Var(\tilde{X}) = \sum_{t=1}^{\omega} Var(\tilde{R}_t).$$

Let ℓ denote the loading for profit pertaining to risk $\tilde{X}$. We make the assumption that ℓ is earned over the whole development period of risk $\tilde{X}$. The amount earned during development year t is

$$\ell_t = \ell \cdot \frac{Var(\tilde{R}_t)}{Var(\tilde{X})}.$$

The above ensures that $\sum_{t=1}^{\omega} \ell_t = \ell$.

We now introduce discounting. Let $\tilde{\delta}(u)$, a random variable, denote the interest rate intensity at time u. The present value at time s of one monetary unit paid at time t is then

$$\tilde{v}(s, t) = e^{-\int_s^t \tilde{\delta}(u)du}$$

Let $\mathcal{G}_t$ denote the cumulative information on the interest rate intensity up to the end of financial year t (which is also development year t of risk $\tilde{X}$). It is assumed that $\mathcal{G}_0$, $\mathcal{G}_1$, ..., $\mathcal{G}_t$, ... is an increasing sequence of σ-algebras.

We have now

$$\tilde{X} = \tilde{v}(0,\ 1) \cdot \tilde{P}_1 + \tilde{v}(0,\ 2) \cdot \tilde{P}_2 + ... + \tilde{v}(0, \omega) \cdot \tilde{P}_\omega$$

Let

$$\tilde{L}_t = E\left(\sum_{s=1}^{\omega - t} \tilde{v}(t, t+s) \tilde{P}_{t+s} | \mathcal{H}_t, \mathcal{G}_t \right)$$

be the loss reserve of the company in respect of risk $\tilde{X}$ at the end of development year t. As a special case we have $L_0 = E(\tilde{X})$.

The loss development risk in development year t is

$$
\begin{aligned}
\tilde{R}_t &= \tilde{L}_{t-1} - \tilde{P}_t - \tilde{L}_t \\
\tilde{R}_t &= E\left(\sum_{s=1}^{\omega-t+1} \tilde{v}(t-1, t-1+s) \tilde{P}_{t-1+s} | \mathcal{H}_{t-1}, \mathcal{G}_{t-1} \right) - \tilde{P}_t \\
&\quad - E\left(\sum_{s=1}^{\omega-t} \tilde{v}(t, t+s) \tilde{P}_{t+s} | \mathcal{H}_t, \mathcal{G}_t \right) \\
&= \left[E\left(\sum_{s=0}^{\omega-t} \tilde{v}(t-1, t+s) \tilde{P}_{t+s} | \mathcal{H}_{t-1}, \mathcal{G}_{t-1} \right) \right. \\
&\quad \left. - E\left(\sum_{s=0}^{\omega-t} \tilde{v}(t-1, t+s) \tilde{P}_{t+s} | \mathcal{H}_t, \mathcal{G}_{t-1} \right) \right] \\
&\quad + \left[E\left(\sum_{s=0}^{\omega-t} \tilde{v}(t-1, t+s) \tilde{P}_{t+s} | \mathcal{H}_t, \mathcal{G}_{t-1} \right) \right. \\
&\quad \left. - E\left(\sum_{s=0}^{\omega-t} \tilde{v}(t, t+s) \tilde{P}_{t+s} | \mathcal{H}_t, \mathcal{G}_t \right) \right] \\
\tilde{R}_t &= {}_1\tilde{R}_t + {}_2\tilde{R}_t
\end{aligned}
$$

Assumption 6
The interest rate process and the claims process are stochastically independent.

Under the above assumption we obtain

$$
{}_1\tilde{R}_t = \sum_{s=0}^{\omega-t} E(\tilde{v}(t-1, t+s) | \mathcal{G}_{t-1}) \cdot \left(E(\tilde{P}_{t+s} | \mathcal{H}_{t-1}) - E(\tilde{P}_{t+s} | \mathcal{H}_t) \right)
$$

${}_1\tilde{R}_t$ is the loss reserve development risk. It is seen at once that $E({}_1\tilde{R}_t) = 0$. In addition the company will earn a profit loading ℓ_t as defined above, for assuming the risk ${}_1\tilde{R}_t$.

We also have

$$_2\tilde{R}_t = \sum_{s=0}^{\omega-t} E(\tilde{P}_{t+s}|\mathcal{H}_t) \cdot (E(\tilde{v}(t-1,t+s)|\mathcal{G}_{t-1}) - E(\tilde{v}(t,t+s)|\mathcal{G}_t))$$

$$_2\tilde{R}_t = \sum_{s=0}^{\omega-t} E(\tilde{P}_{t+s}|\mathcal{H}_t) \cdot (E(\tilde{v}(t-1,t+s)|\mathcal{G}_{t-1}) - E(\tilde{v}(t-1,t+s)|\mathcal{G}_t)$$
$$+ E(\tilde{v}(t-1,t+s)|\mathcal{G}_t) - E(\tilde{v}(t,t+s)|\mathcal{G}_t))$$

$$_2\tilde{R}_t = \sum_{s=0}^{\omega-t} E(\tilde{P}_{t+s}|\mathcal{H}_t) \cdot (E(\tilde{v}(t-1,t+s)|\mathcal{G}_{t-1}) - E(\tilde{v}(t-1,t+s)|\mathcal{G}_t))$$
$$+ \sum_{s=0}^{\omega-t} E(\tilde{P}_{t+s}|\mathcal{H}_t) \cdot E(\tilde{v}(t-1,t+s)|\mathcal{G}_t) \cdot \left(1 - \tilde{v}^{-1}(t-1,t)\right)$$

and it is seen that the first term is the yield curve risk stemming from the discounting of the loss reserves and the second term is the unwinding of the discount.

$-_2\tilde{R}_t$ can be viewed as the yield in financial year t of a bond portfolio with the amounts $E(\tilde{P}_t|\mathcal{H}_t)$, $E(\tilde{P}_{t+1}|\mathcal{H}_t)$, ..., $E(\tilde{P}_\omega|\mathcal{H}_t)$ maturing at time t, $t+1$, ..., ω respectively. The risk $_2\tilde{R}_t$ can therefore be perfectly hedged through asset liability matching.

3.2. Different Accident Years

Let $\tilde{X}_1$, $\tilde{X}_2$, ..., $\tilde{X}_\omega$ denote a risk or a portfolio of risks pertaining to accident years 1, 2, ..., ω. Let $\tilde{P}_{t,s}$ denote the claims payment made in respect of accident year t, in development year s. It is assumed that each $\tilde{X}_t$ is paid over ω development years. We have

$$\tilde{X}_t = \sum_{s=1}^{\omega-t+1} \tilde{v}(t-1,t-1+s)P_{t,s}$$

where $v(s,t)$ is defined as the preceding subsection. $\mathcal{H}_{t,s}(s=1,2,...,\omega)$ is the σ-algebra generated by $\{\tilde{P}_{t,1}, \tilde{P}_{t,2}, ..., \tilde{P}_{t,s}\}$. $\mathcal{G}_t$ is the σ-algebra generated by $\{\tilde{\delta}(u)|u \le t\}$. The loss reserve held by the company in respect of accident year t at the beginning of financial year ω is

$$L_{t,\omega-t} = E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega-1,s+t-1)\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t},\mathcal{G}_{\omega-1}\right).$$

At the end of financial year ω it pays $P_{t,\omega-t+1}$ and puts up a reserve

$$L_{t,\omega-t+1} = E\left(\sum_{s=\omega-t+2}^{\omega} \tilde{v}(\omega,s+t-1)\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t+1},\mathcal{G}_\omega\right).$$

The risk materializing during financial year ω in respect of accident year t is

$$\tilde{R}_{t,\omega-t+1} = L_{t,\omega-t} - P_{t,\omega-t+1} - L_{t,\omega-t+1}.$$

And the overall loss reserve risk is thus

$$\tilde{\Delta} L = -\sum_{t=1}^{\omega-1} R_{t,\omega-t+1}.$$

Note that $\tilde{R}_{\omega,1}$ is the underwriting risk in respect of accident year ω and is therefore not part of the loss reserve risk.

Upon rearranging terms, we obtain

$$\begin{aligned}
\tilde{R}_{t,\omega-t+1} =& E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega-1, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t}, \mathcal{G}_{\omega-1} \right) \\
&- E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t+1}, \mathcal{G}_{\omega} \right) \\
=& \left[E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega-1, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t}, \mathcal{G}_{\omega-1} \right) \right. \\
&\left. - E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega-1, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t+1}, \mathcal{G}_{\omega-1} \right) \right] \\
&+ \left[E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega-1, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t+1}, \mathcal{G}_{\omega-1} \right) \right. \\
&\left. - E\left(\sum_{s=\omega-t+1}^{\omega} \tilde{v}(\omega, s+t-1)\tilde{P}_{t,s} | \mathcal{H}_{t,\omega-t+1}, \mathcal{G}_{\omega} \right) \right] \\
R_{t,\omega-t+1} =& {}_1R_{t,\omega-t+1} + {}_2R_{t,\omega-t+1}
\end{aligned}$$

Using assumption 6 we obtain

$$\begin{aligned}
{}_1R_{t,\omega-t+1} =& \sum_{s=\omega-t+1}^{\omega} E(\tilde{v}(\omega-1, s+t-1)|\mathcal{G}_{\omega-1}) \\
&\cdot \left(E(\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t}) - E(\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t+1}) \right) \\
{}_2R_{t,\omega-t+1} =& \sum_{s=\omega-t+1}^{\omega} E(\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t+1}) \\
&\cdot (E(\tilde{v}(\omega-1, s+t-1)|\mathcal{G}_{\omega-1}) - E(\tilde{v}(\omega, s+t-1)|\mathcal{G}_{\omega}))
\end{aligned}$$

Let $\tilde{\Delta} L = \tilde{\Delta} L_1 + \tilde{\Delta} L_2$ with $\tilde{\Delta} L_i = \sum_{t=1}^{\omega-1} {}_iR_{t,\omega-t+1} \quad i = 1,\ 2.$

$\tilde{\Delta}L_1$ is the loss reserve development risk and $\tilde{\Delta}L_2$ is the yield curve risk combined with the unwinding of the discount.

It is easily seen that $E(\tilde{\Delta}L_1) = 0$. In return for the assumption of the risk $\tilde{\Delta}L_1$ the company earns a profit loading

$$\ell_1 = \sum_{t=1}^{\omega-1} \ell_{t,\omega-t+1}$$

where $\ell_{t,\omega-t+1}$ is the profit loading pertaining to accident year t in development year $\omega - t + 1$ (see section 3.1).

Upon rearranging terms we obtain

$$\tilde{\Delta}L_2 = -\sum_{t=1}^{\omega-1} R_{t,\omega-t+1}$$

$$\tilde{\Delta}L_2 = \sum_{t=1}^{\omega-1} \sum_{s=\omega-t+1}^{\omega} E(\tilde{P}_{t,s}|\mathcal{H}_{t,\omega-t+1}) \cdot (E(\tilde{v}(\omega, s+t-1))|\mathcal{G}_\omega)$$
$$- E(\tilde{v}(\omega-1, s+1-1))|\mathcal{G}_{\omega-1}))$$

$$\tilde{\Delta}L_2 = \sum_{s=0}^{\omega-2} k_s(E(\tilde{v}(\omega, \omega+s)|\mathcal{G}_\omega) - E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_{\omega-1}))$$

with

$$k_s = \sum_{t=s+1}^{\omega} E(\tilde{P}_{t,\omega+s+1-t}|\mathcal{H}_{t,\omega-t+1}) \ .$$

Thus

$$\tilde{\Delta}L_2 = \sum_{s=0}^{\omega-2} k_s(E(\tilde{v}(\omega, \omega+s)|\mathcal{G}_\omega) - E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_\omega)$$
$$+ E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_\omega) - E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_{\omega-1}))$$

$$\tilde{\Delta}L_2 = \sum_{s=0}^{\omega-2} k_s\big(E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_\omega) \cdot \big(\tilde{v}^{-1}(\omega-1, \omega)\big)\big)$$
$$+ \sum_{s=0}^{\omega-2} k_s(E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_\omega) - E(\tilde{v}(\omega-1, \omega+s)|\mathcal{G}_{\omega-1}))$$

where the first term is the unwinding of the discount and the second term is the yield curve risk stemming from the discounting of the loss reserves. We have thus

$$\tilde{\Delta}L_2 = \tilde{R}_L \cdot L$$

where $L = \sum_{s=0}^{\omega-2} k_s E(\tilde{v}(\omega - 1, \omega + s)|\mathcal{G}_{\omega-1})$ is the total discounted loss reserves at the beginning of financial year ω and $\tilde{R}_L$ is the yield for financial year ω of a bond portfolio with the amounts k_s maturing at the end of financial year $\omega + s$ $(s = 0, 1, ..., \omega - 2)$. $\tilde{R}_L$ is the rate of return of a bond portfolio with the same maturities as the liabilities of the company. $\tilde{\Delta}L_2$ can thus be perfectly hedged through asset liability matching.

In summary the loss reserve risk consists of two parts

$$\tilde{\Delta}L = (\tilde{\Delta}L_1 - \ell_1) + \tilde{R}_L \cdot L$$

a loss reserve development risk $(\tilde{\Delta}L_1)$ and a yield curve risk $(\tilde{R}_L \cdot L)$.

4. General Model Including Asset Risk

4.1. Optimality Criterion

We have obtained the following representation for the return of the company during the financial year

$$\tilde{\Delta}u = \left(E(\tilde{S}) + \ell - \tilde{S}\right) + (\ell_1 - \tilde{\Delta}L_1) - \tilde{R}_L \cdot L + \tilde{\Delta}A$$

The first two terms are **insurance risks** (underwriting and loss reserve development risk), the last two terms are **financial risks** (yield curve risk and asset risk).

It is assumed that there are n different categories of assets. $\tilde{R}_j$, a random variable, denotes the return of asset category j. A_j denotes the amount invested by the company in asset category j. We have

$$\tilde{\Delta}A = \sum_{j=1}^{n} \tilde{R}_j \cdot A_j \, .$$

Let ρ_0 denote the return of the risk free asset. We obtain the following representation for the excess return of the company

$$\tilde{\Delta}u - \rho_0 u = \left(E(\tilde{S}) + \ell - \tilde{S}\right) + (\ell_1 - \tilde{\Delta}L_1) - (\tilde{R}_L - \rho_0) \cdot L + \sum_{j=1}^{n} (\tilde{R}_j - \rho_0) \cdot A_j$$

where we have used the fact that the sum of the liabilities of the company is equal to the sum of its assets

$$L + u = \sum_{j=1}^{n} A_j$$

Let

$$E(\tilde{S}) + \ell - \tilde{S} = \sum_{i=1}^{m} E(\tilde{X}_i) + \ell_i - \tilde{X}_i$$

$$\ell_1 - \tilde{\Delta} L_1 = \sum_{i=1}^{m'} \ell_i' - \tilde{X}_i' \qquad \left(E(\tilde{X}_i') = 0\right)$$

and

$$(\tilde{R}_L - \rho_0) \cdot L = \sum_{i=1}^{m'} (\tilde{R}_i' - \rho_0) \cdot L_i$$

be a split of the underwriting risk, the loss reserve development risk and the yield curve risk into individual risks (e.g. lines of business, market segments, etc.). We assume that company keeps a share α_i ($\alpha_i \in [0,1]$) of each individual underwriting risk and cedes $1 - \alpha_i$ via quota share reinsurance. Similarly the company retains a share β_j of loss reserve development risk and of the yield curve risk j. The excess profit of the company now reads

$$\tilde{\Delta} u - \rho_0 u = \sum_{i=1}^{m} \alpha_i \cdot \left(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\right) + \sum_{j=1}^{m'} \beta_j \cdot \left((\ell_j' - \tilde{X}_j') - (\tilde{R}_j' - \rho_0) \cdot L_j\right)$$
$$+ \sum_{i=1}^{n} (\tilde{R}_j - \rho_0) \cdot A_j$$

And it is seen that portfolio optimization amounts to an 'optimal' choice of the α's, β's and A's. We now define the optimality criterion.

Let

$$\tilde{\delta}(u) = \frac{\tilde{\Delta} u - \rho_0 u}{u}, \qquad \mu(u) = E\left(\tilde{\delta}(u)\right), \qquad \sigma^2(u) = Var\left(\tilde{\delta}(u)\right)$$

The objective of the company is to maximize

$$2\tau\mu(u) - \sigma^2(u), \qquad \text{with} \quad \tau \geq 0.$$

(For a discussion of the above objective function see section 2.2). As in section 2.2 we have

$$\mu(u) = \frac{R(\underline{\alpha}, \underline{\beta}, \underline{A})}{u}, \qquad \sigma^2(u) = \frac{V(\underline{\alpha}, \underline{\beta}, \underline{A})}{u^2}$$

Thus the same arguments apply and it is seen that the efficient frontier is defined by maximizing the risk return ratio (Sharpe's ratio).

$$\frac{\mu(u)}{\sigma(u)} = \frac{L(\underline{\alpha}, \underline{\beta}, \underline{A})}{\left(V(\underline{\alpha}, \underline{\beta}, \underline{A})\right)^{\frac{1}{2}}} = r(\underline{\alpha}, \underline{\beta}, \underline{A})$$

Hence the following

Definition
A portfolio is **optimal** if and only if the corresponding risk return ratio $r(\underline{\alpha}, \underline{\beta}, \underline{A})$ is maximal. In addition $\underline{\alpha}$ and $\underline{\beta}$ are such that the net retained insurance profit is maximized.

Usually $r(\underline{\alpha}, \underline{\beta}, \underline{A})$ is maximized under certain constraints such as $\alpha_i \in [0, 1]$ and $\beta_j \in [0, 1]$ and, if the company is not allowed to issue securities $A_i \geq 0$.

Once the company portfolio has been determined, the risk return ratio and the efficient border of the company are given. The company still has to choose a specific point on the efficient frontier. This choice is equivalent to the choice of the amount of capital of the company which in turn is defined by the risk tolerance τ (see section 2.2).

Let $\tilde{\Delta u} = \sum_{i=1}^{n} \tilde{Z}_i$ be any split of the total risk of the company into individual risks. Since the amount of capital required to assume the total risk $\tilde{\Delta u}$ is proportional to

$$Var(\tilde{\Delta u}) = \sum_{i=1}^{n} Cov(\tilde{Z}_i, \tilde{\Delta u})$$

We allocate to each individual risk $\tilde{Z}_i$ an amount of capital u_i, which is proportional to the contribution of that risk to the overall volatility of the result of the company

$$u_i = k \cdot Cov(\tilde{Z}_i, \tilde{\Delta u}).$$

Since $\sum_{i=1}^{n} u_i = u$, we obtain

$$u_i = u \cdot \frac{Cov(\tilde{Z}_i, \tilde{\Delta u})}{Var(\tilde{\Delta u})} .$$

The excess profit which the company expects to achieve for assuming the risk $\sigma^2(\tilde{\Delta u})$ is $(\rho - \rho_0) \cdot u$. It is fair to split the excess profit proportionally to the allocated capital. Thus

Definition
The **fair loading** of risk $\tilde{Z}_i$ is

$$(\rho - \rho_0) \cdot u_i = (\rho - \rho_0) \cdot u \cdot \frac{Cov(\tilde{Z}_i, \tilde{\Delta u})}{Var(\tilde{\Delta u})}$$

Remark
If the $\tilde{Z}_i'$s are uncorrelated the fair loading amounts to the variance principle. The multiple of the variance, which must be loaded, is derived from the company portfolio, capitalization level and return objective:

$$(\rho - \rho_0) \cdot u \cdot Var^{-1}(\tilde{\Delta} u).$$

If in addition the amount of equity is optimal

$$u = \tau^{-1} \frac{Var(\tilde{\Delta} u)}{(\rho - \rho_0)u}$$

the loading factor is equal to $(\tau u)^{-1}$.

4.2. Portfolio Optimization

The excess profit of the company is

$$\begin{aligned} \tilde{\Delta} u - \rho_0 u = & \sum_{i=1}^{m} \alpha_i \cdot \left(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\right) \\ & + \sum_{j=1}^{m'} \beta_j \cdot \left((\ell_j' - \tilde{X}_j') - (\tilde{R}_j' - \rho_0) \cdot L_j\right) \\ & + \sum_{i=1}^{n} (\tilde{R}_j' - \rho_0) \cdot A_j \end{aligned}$$

and our objective is to maximize the risk return ratio of the company.

In a first step we have to maximize the risk return ratio of the underwriting and loss reserve subportfolio through reinsurance buying. This leads to more homogeneous and less catastrophe exposed portfolios and hence to higher risk return ratios of the subportfolios. It also leads to distributions which are close to multivariate normal. This process is discussed in section 2.

We now turn to the second step which consists in the optimization of the global portfolio, i.e. in maximizing the risk return ratio as a function of the $\alpha's$, $\beta's$ and $A's$.

Let

$$\begin{aligned} \underline{x}' &= (\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m, A_1, \ldots, A_n) \\ \underline{\mu}' &= (\ell_1, \ldots, \ell_n, \ell_1' - (R_1' - \rho_0) \cdot L_1, \ldots, \ell_{m'}' - (R_{m'}' - \rho_0) \cdot L_{m'}, \\ & \qquad R_1 - \rho_0, \ldots, R_n - \rho_0) \\ \Sigma &= Cov(-\tilde{X}_1, \ldots, -\tilde{X}_m, -\tilde{X}_1' - \tilde{R}_1' L_1, \ldots, -\tilde{X}_{m'} - \tilde{R}_{m'}' L_{m'}, \tilde{R}_1, \ldots, \tilde{R}_n) \end{aligned}$$

The optimization problem now reads

$$r = \frac{\underline{x}' \cdot \mu}{(\underline{x}' \cdot \Sigma \cdot \underline{x})^{\frac{1}{2}}} = \max_{\underline{x}}!$$

with the conditions

$$x_i = \alpha_i \in [0,\ 1] \qquad i = 1,\ \ldots,\ m$$
$$x_i = \beta_i \in [0,\ 1] \qquad i = m+1,\ \ldots,\ m+m'$$

and if the company is not able to issue securities

$$x_i = A_i \geq 0 \qquad i = m+m'+1,\ \ldots,\ m+m'+n$$

This is a standard mathematical programming problem. The solution of which can be derived through standard algorithms.

Remarks

1. We restrict the reinsurance agreements to genuine quota shares. The company is not allowed to take a short position in any insurance subportfolio – which would be unrealistic – or to increase its share of any insurance subportfolio beyond 100% – which would attract important acquisition costs.
2. In order for any portfolio to be feasible the amount of liabilities must exceed the amount of assets

$$u + \sum_{i=1}^{m} \alpha_i L_i \geq \sum_{i=1}^{n} A_j$$

   If this is a true inequality, the assets corresponding to the excess liabilities can be invested in the risk free asset. This amounts to a restriction in the choice of the amount of capital

$$u \geq \sum_{i=1}^{n} A_j - \sum_{i=1}^{m} \alpha_i L_i$$

   We refer to the right hand side of the inequality as to the **amount of net invested assets**.
3. Within the framework of our model we can simultaneously optimize the reinsurance policy and the investment policy of the company. The model allows for a symmetrical treatment of the insurance risks and of the asset risks.

Theorem

We assume that Σ is a regular matrix

1. The unrestricted optimum, i.e. the vector $\underline{x}$ which maximizes

$$r = \frac{\underline{\mu}'\underline{x}}{(\underline{x}'\Sigma\underline{x})^{\frac{1}{2}}}$$

is given by

$$\underline{x} = c \cdot \Sigma^{-1}\underline{\mu}$$

(By definition $\underline{x}$ is only defined up to a constant factor c.)
The unrestricted optimal risk return ratio is equal to

$$r_{\max} = \left(\underline{\mu}'\Sigma^{-1}\underline{\mu}\right)^{\frac{1}{2}}$$

2. $\underline{x}$ is the unrestricted optimum if and only if all the actual loadings are equal to the fair loadings.

Proof

1. We have to maximize

$$r = \frac{\underline{\mu}'\underline{x}}{(\underline{x}'\Sigma\underline{x})^{\frac{1}{2}}}$$

equating to derivatives with respect to x_i to zero, we obtain

$$\frac{dr}{dx_i} = \frac{\mu_i(\underline{x}'\Sigma\underline{x})^{\frac{1}{2}} - \underline{\mu}'\underline{x} \cdot \frac{1}{2}(\underline{x}'\Sigma\underline{x})^{-\frac{1}{2}} \cdot 2\left(\sum_j \sigma_{ij}x_j\right)}{\underline{x}'\Sigma\underline{x}} = 0$$

$i = 1, \ldots, m+n$ where $(\sigma_{ij}) = \Sigma$.
After rearranging terms

$$\mu_i(\underline{x}'\Sigma\underline{x}) = \left(\underline{\mu}'\underline{x}\right)\sum_j \sigma_{ij}x_j, \quad \text{all } i$$

$$\underline{\mu} = k \cdot \Sigma \cdot \underline{x}$$

and since Σ is regular

$$\underline{x} = c \cdot \Sigma^{-1}\underline{\mu}$$

Plugging in the above definition of $\underline{x}$ we obtain

$$r_{\max} = \frac{c \cdot \underline{\mu}'\Sigma^{-1}\underline{\mu}}{\left(c\underline{\mu}'\Sigma^{-1} \cdot \Sigma \cdot c\Sigma^{-1}\underline{\mu}\right)^{\frac{1}{2}}} = \left(\underline{\mu}'\Sigma^{-1}\underline{\mu}\right)^{\frac{1}{2}}$$

2. All the actual loadings are equal to the fair loadings if and only if the following equations are satisfied

$$\alpha_i(\ell_i + \ell'_i) = k \cdot Cov\Big(-\alpha_i(\tilde{X}_i + \tilde{X}_{i'}), \tilde{\Delta} u\Big) \qquad i = 1, \ldots, m$$

$$(A_1 - L)(R_1 - \rho_0) = k \cdot Cov\Big((A_1 - L)\tilde{R}_1, \tilde{\Delta} u\Big)$$

$$A_j(R_j - \rho_0) = k \cdot Cov(A_j \tilde{R}_j, \tilde{\Delta} u) \qquad j = 2, \ldots, n$$

Using the above notation, this is equivalent to

$$x_i \mu_i = k \cdot Cov(x_i \tilde{Z}_i, \tilde{\Delta} u) \qquad i = 1, \ldots, m+n$$

for an appropriate choice of $\tilde{Z}_i$.
Hence

$$\mu_i = k \cdot \sum_j \Sigma_{ij} x_j \qquad i = 1, \ldots, m+n$$

$$\underline{\mu} = k \cdot \Sigma \cdot \underline{x}$$

which proves the 2nd statement of the theorem

q.e.d.

Remarks
1. The 2nd statement of our theorem is a further justification for our capital allocation formula.
2. The theorem is a generalisation of the theorem of section 2.5.

Example 1
We now turn to a numerical example. The company has two underwriting risks and two loss reserve risks which correspond to the different customer segments of the company. The risks and returns are as follows

Underwriting portfolio	*Risk*	ℓ	σ	$\frac{\ell}{\sigma}$
Private customers	$\tilde{X}_1$	4.5	15	30%
Industrial customers	$\tilde{X}_2$	14.4	30	48%

Note that we do not give the premium income since it is irrelevant.
Let $Corr(\tilde{X}_i, \tilde{X}_j) = \delta_{ij}$ where δ_{ij} is the Kronecker Symbol

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases} \quad .$$

Loss Reserve Portfolio	*Risk*	*L*	ℓ	σ	$\frac{\ell}{\sigma}$
Private customers	$\tilde{X}_1$	400	0.5	5	10%
Industrial customers	$\tilde{X}_2$	600	1.6	10	16%
		1'000			

with $Corr(\tilde{X}'_i, \tilde{X}'_j) = \delta_{ij}$ and $Corr(\tilde{X}_i, \tilde{X}'_j) = \delta_{ij} \cdot 0.40$.

Note that both in the case of the private customer and of the industrial customer portfolio the ratio between loading and variance is the same for the underwriting and for the loss reserve risk.

There are four different asset categories with risks and returns as defined below

Asset Category	*Risk*	$R_i - \rho_0$	σ	$\frac{R_i - \rho_0}{\sigma}$
Bond portfolio with medium term duration ($\tilde{R}_1 = \tilde{R}_L$)	$\tilde{R}_1$	1%	4%	25%
Bond portfolio with long term duration	$\tilde{R}_2$	2%	6%	33%
Equity portfolio	$\tilde{R}_3$	10%	20%	50%
Real Estate portfolio	$\tilde{R}_4$	8%	20%	40%

The correlation matrix of the different asset categories is as follows

$$Corr(\tilde{R}_i, \tilde{R}_j) = \begin{bmatrix} 1 & 0.9 & 0.4 & 0.4 \\ & 1 & 0.4 & 0.4 \\ & & 1 & 0.4 \\ & & & 1 \end{bmatrix}$$

It is assumed that insurance risks and asset risks are uncorrelated

$$Corr(\tilde{X}_i, \tilde{R}_j) = 0 \qquad \text{for all } i,j.$$

Without any loss of generality we assume

$$\tilde{R}_i = \tilde{R}'_i \qquad i = 1,\ 2.$$

This amounts to choosing bond portfolios with maturities matching the expected maturities of the respective liability portfolios.

We have

$$\underline{\mu}' = (4.5,\ 14.4,\ -3.5,\ -10.4,\ 0.01,\ 0.02,\ 0.10,\ 0.08)$$

and

$$\Sigma = \begin{bmatrix} Cov(\tilde{X}_i, \tilde{X}_j) & Cov(\tilde{X}'_i, \tilde{X}'_j) & 0 \\ Cov(\tilde{X}'_i, \tilde{X}_j) & Cov(\tilde{X}'_i, \tilde{X}'_j) + L_iL_j \cdot Cov(\tilde{R}'_i, \tilde{R}_j) & -L_i \cdot Cov(\tilde{R}'_i, \tilde{R}_j) \\ 0 & -L_j \cdot Cov(\tilde{R}_i, \tilde{R}'_j) & Cov(\tilde{R}_i, \tilde{R}_j) \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 225 & 0 & 30 & 0 & 0 & 0 & 0 & 0 \\ 0 & 900 & 0 & 120 & 0 & 0 & 0 & 0 \\ 30 & 0 & 281 & 518.4 & -0.64 & -0.864 & -1.28 & -1.28 \\ 0 & 120 & 518.4 & 1396 & -1.296 & -2.16 & -2.88 & -2.88 \\ 0 & 0 & -0.64 & -1.296 & 0.0016 & 0.00216 & 0.0032 & 0.0032 \\ 0 & 0 & -0.864 & -2.16 & 0.00216 & 0.0036 & 0.0048 & 0.0048 \\ 0 & 0 & -1.28 & -2.88 & 0.0032 & 0.0048 & 0.04 & 0.016 \\ 0 & 0 & -1.2 & -2.88 & 0.0032 & 0.0048 & 0.016 & 0.04 \end{bmatrix}$$

The unconstrained solution is

$$\underline{x} = c \cdot \Sigma^{-1} \cdot \underline{\mu} = (1,\ 0.8,\ -0.23,\ -0.18,\ -568.1,\ 245.1,\ 94.8,\ 54.4)'$$

which is not admissible because it entails taking a short position in the two loss reserve risks and issuing the short term bond portfolio for an amount of 568.1 monetary units. The constrained optimization problem is

$$c \cdot \underline{x}' \cdot \underline{\mu} - \underline{x}' \cdot \Sigma^{-1} \cdot \underline{x} = \max_{\underline{x}}!$$

with

$$x_3 = x_4 = x_5 = 0.$$

The associated objective function is

$$Z = c \cdot \underline{x}' \cdot \underline{\mu} - \underline{x}' \cdot \Sigma \cdot \underline{x} + \lambda_3 x_3 + \lambda_4 x_4 + \lambda_5 x_5 = \max!$$

where λ_3, λ_4 and λ_5 are the Lagrange multipliers associated with the above constraints. To solve the constrained optimization problem we must find $x_1, \ldots, x_N$ $(N = m + m' + n)$ such that

$$\frac{\partial Z}{\partial x_i} = 0 \quad i = 1, \ldots, N \quad \text{and} \quad \frac{\partial Z}{\partial \lambda_3} = \frac{\partial Z}{\partial \lambda_4} = \frac{\partial Z}{\partial \lambda_5} = 0.$$

This leads to the following set of equations

$$-c \cdot \mu_i + 2\sum_{j=1}^{N} \sigma_{ij} \cdot x_j - \sum_{j=3}^{5} \lambda_j \delta_{i,j} = 0$$
$$x_3 = x_4 = x_5 = 0$$

in matrix notation

$$\left[\begin{array}{cccccccccc}
 & & & & & & & 0 & 0 & 0\\
 & & & & & & & 0 & 0 & 0\\
 & & & & & & & -1 & 0 & 0\\
 & & & & 2\Sigma & & & 0 & -1 & 0\\
 & & & & & & & 0 & 0 & -1\\
 & & & & & & & 0 & 0 & 0\\
 & & & & & & & \vdots & \vdots & \vdots\\
 & & & & & & & 0 & 0 & 0\\
0 & 0 & 1 & 0 & \dots & \dots & \dots & 0 & & \\
0 & 0 & 0 & 1 & 0 & \dots & \dots & 0 & 0 & \\
0 & 0 & 0 & 0 & 1 & 0 & \dots & 0 & &
\end{array}\right]
\begin{bmatrix} x_1\\ x_2\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ x_8\\ \lambda_3\\ \lambda_4\\ \lambda_5 \end{bmatrix}
= c\cdot
\begin{bmatrix} \mu_1\\ \mu_2\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \vdots\\ \mu_8\\ 0\\ 0\\ 0 \end{bmatrix}$$

or

$$\Sigma^{*}\underline{x}^{*} = c\underline{\mu}^{*}$$

and the solution is

$$\underline{x}* = c(\Sigma^{*})^{-1}\underline{\mu}^{*}$$

The optimal constrained portfolio of the company is thus

Underwriting Risk	*Coefficient α, β or A*	*Expected Profit*	*Contribution to overall variance*
$-\tilde{X}_1 + \ell_1 + E(\tilde{X}_1)$	1	4.5	225.00
$-\tilde{X}_2 + \ell_2 + E(\tilde{X}_2)$	0.80	11.52	576.00
Loss Reserve Risk			
$-\tilde{X}_1' + \ell_1' - (\tilde{R}_1' - \rho_0)\cdot L_1$	0	0	0
$-\tilde{X}_2' + \ell_2' - (\tilde{R}_2' - \rho_0)\cdot L_2$	0	0	0
Asset Risks			
$\tilde{R}_1 - \rho_0$	0	0	0
$\tilde{R}_2 - \rho_0$	82.30	1.65	82.30
$\tilde{R}_3 - \rho_0$	94.14	9.41	470.68
$\tilde{R}_4 - \rho_0$	52.47	4.20	209.88
		31.28	1563.86

The optimal amount of capital is

$$u = \tau^{-1}\frac{V}{R}$$

with R and V are the expected profit and the contribution to overall variance respectively (see section 2.2.). Assuming τ we obtain $u = 199.98$.

The salient features of the optimal portfolio are the following
- The company cedes a 20% quota share of its industrial business.
- The company fully reinsures the loss reserve risk. As a consequence its balance sheet is not leveraged at all. The liability side of the balance sheet consists of equity only, there is no debt.
- The total amount of net invested assets is 228.91 which compares with an optimal amount of equity of 199.98. The optimal policy is only feasible if the company can raise an amount of debt of 28.93 monetary units at the risk free rate.
- The company invests a substantial part of its nets invested assets in shares and real estate (64%). The contribution to the expected profit and to the overall volatility from asset risks is substantial (49%).
- The optimal risk return ratio is $r = 0.791$.
- For the unconstrained risks (i.e. all the risks except x_3, x_4 and x_5) we have

$$\frac{\text{expected profit}}{\text{contribution to overall variance}} = \text{constant} = 0.020.$$

  For the constrained risks the above quantity is irrelevant.

Example 2
Based on the result of the section on loss reserves, the model assumes that the loadings ℓ_i and ℓ_i' are proportional to the variance of the corresponding risks $\sigma^2(\tilde{X}_i)$ and $\sigma^2(\tilde{X}_i')$. In practice however a loss portfolio transfer ($\beta_i = 0$) would probably command a much higher loading. Since there is no liquid reinsurance market for loss portfolio transfers we make the following

Assumption 7
$m' = m,\ \beta_i = \alpha_i \quad i = 1, \ldots, m.$
In addition we simplify the notation
$\tilde{X}_i + \tilde{X}_i'$ is replaced by $\tilde{X}_i$, and
$\ell_i + \ell_i'$ is replaced by ℓ_i.
The model now becomes

$$\tilde{\Delta}u - \rho_0 u = \sum_{i=1}^{m} \alpha_i\big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i - (\tilde{R}_i' - \rho_0)\cdot L_i\big) + \sum_{j=1}^{m} (\tilde{R}_j - \rho_0)\cdot A_j$$

We now reanalyze the preceding example. We have

Insurance Portfolio	*Risk*	L	ℓ	σ	$\frac{\ell}{\sigma}$
Private customers	$\tilde{X}'_1$	400	5	17.61	28%
Industrial customers	$\tilde{X}'_2$	600	16	35.12	45%

The other model parameters remain unchanged and we have

$$\mu' = (1,\ 4,\ 0.01,\ 0.02,\ 0.10,\ 0.08)$$

$$\Sigma = \begin{bmatrix} Cov(\tilde{X}_i, \tilde{X}_j) + L_i L_j Cov(\tilde{R}'_i, \tilde{R}'_j) & -L_i Cov(\tilde{R}'_i, \tilde{R}_j) \\ -L_i Cov(\tilde{R}_j, \tilde{R}'_i) & Cov(\tilde{R}_i, \tilde{R}_j) \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 566 & 518.4 & -0.64 & -0.864 & -1.28 & -1.28 \\ 518.4 & 2536 & -1.296 & -2.16 & -2.88 & -2.88 \\ -0.64 & -1.296 & 0.0016 & 0.00216 & 0.0032 & 0.0032 \\ -0.864 & -2.16 & 0.00216 & 0.0036 & 0.0048 & 0.0048 \\ -1.28 & -2.88 & 0.0032 & 0.0048 & 0.04 & 0.016 \\ -1.28 & -2.88 & 0.0032 & 0.0048 & 0.016 & 0.04 \end{bmatrix}$$

The unconstrained solution is

$$\underline{x} = c \cdot \Sigma^{-1} \underline{\mu} = (1,\ 0.8,\ -208.7,\ 935.3,\ 121.2,\ 69.6)$$

which entails a short position of 208.7 monetary units in the medium term bond (in addition to the 400 monetary units of loss reserves with the same return $\tilde{R}'_1 = \tilde{R}_1$). Within the framework of this model this is not admissible. We therefore introduce the side condition

$$x_3 = 0$$

which leads to the following objective function

$$Z = c \cdot \underline{x}' \cdot \underline{\mu} - \underline{x}' \Sigma \underline{x} + \lambda_3 x_3 = \max!$$

Deriving with respect to x_i and λ_3, we obtain the following matrix equation

$$\begin{bmatrix} & & & & & & 0 \\ & & & & & & 0 \\ & & 2\Sigma & & & & -1 \\ & & & & & & 0 \\ & & & & & & 0 \\ & & & & & & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ x_6 \\ \lambda_3 \end{bmatrix} = c \cdot \begin{bmatrix} \mu_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \mu_6 \\ 0 \end{bmatrix}$$

which is easily solved yielding the following optimal constrained portfolio

Insurance Risk	*Coefficients α_i*	*L*	*Expected profit*	*Contribution to overall variance*
$-\tilde{X}'_1 + \ell'_1 - (\tilde{R}'_1 - \rho_0) \cdot L_1$	1	400	1	57.95
$-\tilde{X}'_2 + \ell'_2 - (\tilde{R}'_2 - \rho_0) \cdot L_2$	0.75	450	2.99	173.35
		850		

Asset Risks	*Invested Amounts A_j*		*Expected profit*	*Contribution to overall variance*
$\tilde{R}_1 - \rho_0$	0		0	0
$\tilde{R}_2 - \rho_0$	776.2		15.52	899.64
$\tilde{R}_3 - \rho_0$	112.1		11.21	649.50
$\tilde{R}_4 - \rho_0$	63.8		5.10	295.70
	952.1		35.82	2076.14

Assuming $\tau = 0.25$, the optimal amount of equity is

$$u = \tau^{-1} \frac{V}{R} = 231.83$$

The salient features of the optimal portfolio are the following
- The company cedes a 25% quota share of its industrial business.
- The company keeps most of its loss reserves (850 monetary units out of a gross amount of 1000) thus leveraging its balance sheet.
- The total amount of net invested assets is 102.1 which compares with an optimal amount of equity of 231.83. The optimal policy is thus feasible without borrowing.

- The contribution to the expected profit and to the overall volatility from asset risks is much higher than the corresponding quantities from insurance risks (89% vs 11%). This is in particular due to the fact that the short position in yield curve risk acts as a hedge.
- The optimal risk return ratio $r = 0.786$.
- For unconstrained risks, we have

$$\frac{\text{expected profit}}{\text{contribution to overall variance}} = \text{constant} = 0.017$$

Example 3

So far we have assumed that the company may not issue securities, or in other words that $A_i \geq 0 \; i = 1, \ldots, n$.

We now make the following

Assumption 8

The company may issue securities, i.e. $A_i \; i = 1, \ldots, n$ are unconstrained.

Without loss of generality we also assume $n \geq m$ and

$$\tilde{R}_i' \equiv \tilde{R}_i \quad i = 1, \ldots, m$$

and we introduce the following notational simplification

$$B_j = A_j - \alpha_j L_j \quad j = 1, \ldots, m$$

The model can now be rewritten as

$$\tilde{\Delta} u - \rho_0 \cdot u = \sum_{i=1}^{m} \alpha_i \big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big) + \sum_{j=1}^{n} (\tilde{R}_j - \rho_0) \cdot B_j \, .$$

We have

$$\underline{\mu}' = (5,\ 16,\ 0.01,\ 0.02,\ 0.10,\ 0.08)$$

$$\Sigma = \begin{bmatrix} Cov(\tilde{X}_i, \tilde{X}_j) & 0 \\ 0 & Cov(\tilde{R}_i, \tilde{R}_j) \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 310 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1240 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0016 & 0.00216 & 0.0032 & 0.0032 \\ 0 & 0 & 0.00216 & 0.0036 & 0.0048 & 0.0048 \\ 0 & 0 & 0.0032 & 0.0048 & 0.04 & 0.016 \\ 0 & 0 & 0.0032 & 0.0048 & 0.016 & 0.04 \end{bmatrix}$$

The unconstrained solution

$$\underline{x} = c \cdot \Sigma^{-1}\underline{\mu} = (1,\ 0.8,\ -608.7,\ 455.3,\ 121.2,\ 69.6)$$

is now admissible and the optimal portfolio is

Insurance Risks	*Coefficients a_i*	*Expected profit*	*Contribution to overall variance*
$E(\tilde{X}_1) + \ell_1 - \tilde{X}_1$	1	5	310
$E(\tilde{X}_2) + \ell_2 - \tilde{X}_2$	0.8	12.8	793.6
Financial Risks	*Net invested Amounts B_j*		
$\tilde{R}_1 - \rho_0$	−608.7	−6.09	−377.4
$\tilde{R}_2 - \rho_0$	455.3	9.11	564.5
$\tilde{R}_3 - \rho_0$	121.2	12.12	751.7
$\tilde{R}_4 - \rho_0$	69.6	5.57	345.1
	37.4	38.51	2387.5

Assuming $\tau = 0.25$, the optimal amount of equity is

$$u = \tau^{-1}\frac{V}{R} = 247.99$$

The salient features of the optimal portfolio are
- The company cedes a 20% quota share of its industrial business.
- The gross invested amounts in asset category 1 and 2 are respectively −208.7 and 935.3 which are identical with the corresponding amounts pertaining to the (inadmissible) unconstrained solution of the preceding example.
- The amount of net invested assets is 37.4 which compares with an optimal amount of equity of 231.83.
- The contribution to the expected profit and to the overall volatility from financial risks (including short position in yield curve risk) is higher than the corresponding quantities from insurance risks (54% vs 46%).
- The optimal risk return ratio is $r = 0.788$.
- The ratio of expected profit to contribution to overall variance is the same for all risks (0.016).

Discussion of Assumption 8

A comparison between the last two examples shows that dropping the constraint $A_j \geq 0$ (for all j) leads to a higher risk return ratio and to a lower amount of net invested assets. In practice insurance companies are allowed to issue preferred shares or – through a holding company – obtain bank loans or issue corporate bonds. The amount of debt they are able to raise is usually limited and commands a spread over the risk free rate.

Generalization of Theorem 4.2

From the above example it is seen that the constrained optimum is obtained by computing the unconstrained solution

$$\underline{x} = c \cdot \Sigma^{-1} \cdot \underline{\mu}$$

and by choosing c in such a way that the retained insurance profit is maximized (i.e. $\max\limits_{i=1,\ldots,m+m'} x_j = 1$).

Let $i_1, \ldots, i_k$ be those indices for which $x_{i_j} < 0$. The constrained optimum is obtained by maximizing the following objective function

$$Z = \lambda \underline{x}' \cdot \underline{\mu} - \underline{x}' \Sigma \underline{x} - \lambda_{i_1} x_{i_1} - \ldots - \lambda_{i_k} x_{i_k}$$

where $\lambda_{i_1}, \ldots, \lambda_{i_k}$ are the Lagrange multipliers associated with the constraints

$$x_{i_j} = 0 \qquad (j = 1, \ldots, k).$$

This leads to the following set of equations

$$\frac{\partial Z}{\partial x_i} = -\lambda \mu_i + 2 \sum_j \sigma_{ij} x_j - \sum_{j=1}^{k} \delta_{i,i_j} \lambda_j = 0$$

$$x_{i_1} = \ldots = x_{i_k} = 0$$

In particular for unconstrained variables x_i, we have

$$\frac{2}{\lambda} \cdot \sum_j \sigma_{ij} x_j = \mu_i$$

which translates into

$$\ell_i = k \cdot Cov(-\tilde{X}_i, \tilde{\Delta} u)$$

$$\ell'_i - (R'_i - \rho_0) L_i = k \cdot Cov(-\tilde{X}'_i - \tilde{R}'_i \cdot L_i, \tilde{\Delta} u)$$

$$R_i - \rho_0 = k \cdot Cov(\tilde{R}_i, \tilde{\Delta} u)$$

i.e. the loading pertaining to unconstrained variables is equal to the fair loading. This is a further justification for our capital allocation and pricing formula.

4.3. Insurance Risk and Financial Risk

We consider the expression for the excess profit of the company which we have derived at the beginning of section 4.

$$\begin{aligned}\tilde{\Delta}u - \rho_0 u &= E(\tilde{S}) + \ell + \ell_1 - (\tilde{S} + \tilde{\Delta}L_1) - (\tilde{R}_L - \rho_0)L + \sum_{j=1}^{n}(\tilde{R}_j - \rho_0)A_j \\ &= \tilde{Z} + (\tilde{R} - \rho_0)A\end{aligned}$$

where

$$\tilde{Z} = E(\tilde{S}) + \ell + \ell_1 - (\tilde{S} + \tilde{\Delta}L_1)$$

$$\tilde{R} = \frac{\sum_j \tilde{R}_j \cdot A_j - R_L \cdot L}{A} \quad \text{with} \quad A = \sum_{j=1}^{n} A_j - L$$

$\tilde{Z}$ is the insurance risk, i.e. the sum of the underwriting risk and of the loss reserve development risk. $\tilde{R}$ is the rate of return of the financial risk and A is the amount of net invested assets. We introduce the following notation

$$\ell_z = E(\tilde{Z}) = \ell + \ell_1, \qquad \sigma_z^2 = Var(\tilde{Z})$$
$$\delta_R = E(\tilde{R}) - \rho_0, \qquad \sigma_R^2 = Var(\tilde{R}), \qquad \mathcal{K} = Corr(\tilde{Z}, \tilde{R})$$

The following theorem expresses the overall risk return ratio as a function of the insurance risk return ratio and of the financial risk return ratio.

Theorem
Let $\mathcal{K} \neq \pm 1$. The overall risk return ratio

$$r(A) = \frac{E(\tilde{\Delta}u) - \rho_0 u}{\sigma(\tilde{\Delta}u)} = \frac{\ell_z + \delta_R A}{\sqrt{\sigma_z^2 + (\sigma_R A)^2 + 2\mathcal{K}\sigma_z(\sigma_R A)}}$$

is maximized for the following amount of net invested assets

$$A = \frac{\ell_z}{\delta_R} \frac{\left(\frac{\delta_R}{\sigma_R}\right)^2 - \mathcal{K}\frac{\ell_z}{\sigma_z}\frac{\delta_R}{\sigma_R}}{\left(\frac{\ell_z}{\sigma_z}\right)^2 - \mathcal{K}\frac{\ell_z}{\sigma_z}\frac{\delta_R}{\sigma_R}}$$

and the corresponding risk return ratio is

$$r = r(A) = \left(\frac{\left(\frac{\ell_z}{\sigma_z}\right)^2 + \left(\frac{\delta_R}{\sigma_R}\right)^2 - 2\mathcal{K}\frac{\ell_z}{\sigma_z}\frac{\delta_R}{\sigma_R}}{1 - \mathcal{K}^2}\right)^{\frac{1}{2}}$$

Proof
We have

$$E(\tilde{\Delta}u) - \rho_0 u = \ell_z + \delta_R \cdot A, \qquad \sigma^2(\tilde{\Delta}u) = \sigma_z^2 + \sigma_R^2 A^2 + 2\mathcal{K}\sigma_z\sigma_R \cdot A$$

it follows that

$$r(A) = \frac{\ell_z + \delta_R A}{\sqrt{\sigma_z^2 + (\sigma_R A)^2 + 2\mathcal{K}\sigma_z\sigma_R A}} = \frac{\delta(A)}{\sqrt{V(A)}}$$

where $\delta(A)$ is $E(\tilde{\Delta}u) - \rho_0 \cdot u$ and $V(A)$ is $\sigma^2(\tilde{\Delta}u)$ considered as function of A.

Putting the derivative of r with respect to A equal to zero, we obtain

$$r'(A) = \frac{\delta'(A) \cdot V(A)^{\frac{1}{2}} - \delta(A)\frac{1}{2}V(A)^{-\frac{1}{2}}V'(A)}{V(A)} = 0$$

$$\delta'(A)V(A) - \frac{1}{2}\delta(A)V'(A) = 0$$

$$\delta_R(\sigma_z^2 + \sigma_R^2 A^2 + 2\mathcal{K}\sigma_z\sigma_R A) = (\ell_z + \delta_R A)(\sigma_R^2 A + \mathcal{K}\sigma_z\sigma_R)$$

$$A = \frac{\delta_R\sigma_z^2 - \mathcal{K}\ell_z\sigma_z\sigma_R}{\ell_z\sigma_R^2 - \mathcal{K}\delta_R\sigma_z\sigma_R} = \frac{\ell_z}{\delta_R}\,\frac{\left(\frac{\delta_R}{\sigma_R}\right)^2 - \mathcal{K}\frac{\ell_z}{\sigma_z}\frac{\delta_R}{\sigma_R}}{\left(\frac{\ell_z}{\sigma_z}\right)^2 - \mathcal{K}\frac{\ell_z}{\sigma_z}\frac{\delta_R}{\sigma_R}}$$

which proves the first statement of the theorem. In order to evaluate $r(A)$, we introduce the following notation

$$r_1 = \frac{\ell_z}{\sigma_z} \qquad r_2 = \frac{\delta_R}{\sigma_R}$$

and we restate the expression for A

$$A = \frac{\sigma_z}{\sigma_R}\frac{r_2 - \mathcal{K}r_1}{r_1 - \mathcal{K}r_2}$$

Thus obtaining

$$V(A) = \sigma_z^2 + \sigma_z^2\left(\frac{r_2 - \mathcal{K}r_1}{r_1 - \mathcal{K}r_2}\right)^2 + 2\mathcal{K}\sigma_z^2\frac{r_2 - \mathcal{K}r_1}{r_1 - \mathcal{K}r_2}$$

$$V(A) = \frac{\sigma_z^2}{(r_1 - \mathcal{K}r_2)^2}\Big((r_1 - \mathcal{K}r_2)^2 + (r_2 - \mathcal{K}r_1)^2 + 2\mathcal{K}(r_2 - \mathcal{K}r_1)(r_1 - \mathcal{K}r_2)\Big)$$

$$V(A) = \frac{\sigma_z^2}{(r_1 - \mathcal{K}r_2)^2}(1 - \mathcal{K}^2) \cdot (r_1^2 + r_2^2 - 2\mathcal{K}r_1 r_2)$$

$$r(A) = \frac{r_1 - \mathcal{K}r_2}{\sigma_z}\,\frac{\ell_z + \delta_R\frac{\sigma_z}{\sigma_R}\frac{r_2 - \mathcal{K}r_1}{r_1 - \mathcal{K}r_2}}{\sqrt{(1 - \mathcal{K}^2)(r_1^2 + r_2^2 - 2\mathcal{K}r_1 r_2)}} = \sqrt{\frac{r_1^2 + r_2^2 - 2\mathcal{K}r_1 r_2}{1 - \mathcal{K}^2}}$$

which proves the theorem.

q.e.d.

Remarks

1. From the proof of the theorem it is easily seen that for $\mathcal{K} = \pm 1$ we have $A = \mp \frac{\sigma_z}{\sigma_R}$ and $V(A) = 0$ i.e. the risk if fully eliminated.
2. For $\mathcal{K} = 0$ we have

$$A = \frac{\sigma_z}{\sigma_R} \frac{\left(\frac{\delta_R}{\sigma_R}\right)}{\left(\frac{\ell_z}{\sigma_z}\right)} \quad \text{and} \quad r(A) = \sqrt{\left(\frac{\ell_z}{\sigma_z}\right)^2 + \left(\frac{\delta_R}{\sigma_R}\right)^2}$$

and it is seen that the assumption of asset risk leads to a considerably higher risk return ratio. In practice we have $\mathcal{K} \simeq 0$ and the statement is thus true for all practical situations.

4.4. Realistic Example

We now turn to a more realistic example. The insurance portfolio of the company is broken down into four subportfolios corresponding to different lines of business and to different customer segments. The risks and returns of the combined underwriting and loss development risks are as follows

Insurance Subportfolio	*Risks*	*P*	*L*	σ	ℓ	$\frac{\ell}{\sigma}$
Motor	$\tilde{X}_1$	50	75	2.5	0.5	20%
Homeowners	$\tilde{X}_2$	20	10	3.2	0.8	25%
Industrial Fire	$\tilde{X}_3$	10	5	4	1	25%
General Third Party Liability	$\tilde{X}_4$	10	20	4	1.5	37.5%
		90	110		3.8	

L denotes the amount of loss reserves.

The premium volume is given for purely illustrative purposes. It is not used below. The ratio between standard deviation and premium volume as well as the ratio between loss reserves and premium are chosen in a realistic way. It is assumed that the motor and the homeowners portfolio are both exposed to storm and are therefore positively correlated.

$$Corr(\tilde{X}_1, \tilde{X}_2) = 0.20$$

The other correlations between insurance risks stem from the influence of the economic cycle and are treated below.

The different asset categories are as in the example of section 4.2.

Asset Category	*Risk*	$R_i - \rho_0$	σ	$\frac{R_i - \rho_0}{\sigma}$
Bond portfolio with medium term duration ($\tilde{R}_1 = \tilde{R}_L$)	$\tilde{R}_1$	1%	4%	25%
Bond portfolio with long term duration	$\tilde{R}_2$	2%	6%	33%
Equity portfolio	$\tilde{R}_3$	10%	20%	50%
Real Estate portfolio	$\tilde{R}_4$	8%	20%	40%

The correlation matrix of the different asset categories is as follows

$$Corr(\tilde{R}_i, \tilde{R}_j) = \begin{bmatrix} 1 & 0.9 & 0.4 & 0.4 \\ & 1 & 0.4 & 0.4 \\ & & 1 & 0.4 \\ & & & 1 \end{bmatrix}$$

During a boom phase of the economic cycle interest rates and therefore investment income from bonds are high, but so is the inflation rate which leads to an increased loss amount of the motor and of the general third party liability portfolio. Therefore we assume

$$Corr(-\tilde{X}_1, \tilde{R}_1) = Corr(-\tilde{X}_1, \tilde{R}_2) = -0.2$$
$$Corr(-\tilde{X}_4, \tilde{R}_1) = Corr(-\tilde{X}_4, \tilde{R}_2) = -0.2$$

and

$$Corr(\tilde{X}_1, \tilde{X}_4) = 0.2$$

When the economy goes into recession, equities and real estate depreciate, industrial fire results worsen – due to arson – and motor results improve – because people drive less. Thus

$$Corr(-\tilde{X}_1, \tilde{R}_3) = Corr(-\tilde{X}_1, \tilde{R}_4) = -0.2$$
$$Corr(-\tilde{X}_3, \tilde{R}_3) = Corr(-\tilde{X}_3, \tilde{R}_3) = 0.2$$

and

$$Corr(\tilde{X}_1, \tilde{X}_3) = -0.2$$

In summary we have the following correlations

	$-\tilde{X}_1$	$-\tilde{X}_2$	$-\tilde{X}_3$	$-\tilde{X}_4$	$\tilde{R}_1$	$\tilde{R}_2$	$\tilde{R}_3$	$\tilde{R}_4$
$\tilde{X}_1$	1	0.2	−0.2	0.2	−0.2	−0.2	−0.2	−0.2
$-\tilde{X}_2$	0.2	1	0	0	0	0	0	0
$-\tilde{X}_3$	−0.2	0	1	0	0	0	0.2	0.2
$-\tilde{X}_4$	0.2	0	0	1	−0.2	−0.2	0	0

Thus

$$\mu' = (0.5,\ 0.8,\ 1,\ 1.5,\ 0.01,\ 0.02,\ 0.10,\ 0.08)$$

$$\Sigma = \begin{bmatrix} 6.25 & 1.6 & -2 & 2 & -0.02 & -0.03 & -0.1 & -0.1 \\ 1.6 & 10.24 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & 0 & 16 & 0 & 0 & 0 & 0.16 & 0.16 \\ 2 & 0 & 0 & 16 & -0.032 & -0.048 & 0 & 0 \\ -0.02 & 0 & 0 & -0.032 & 0.0016 & 0.00216 & 0.0032 & 0.0032 \\ -0.03 & 0 & 0 & -0.048 & 0.00216 & 0.0036 & 0.0048 & 0.0048 \\ -0.1 & 0 & 0.16 & 0 & 0.0032 & 0.0048 & 0.04 & 0.016 \\ -0.1 & 0 & 0.16 & 0 & 0.0032 & 0.0048 & 0.016 & 0.04 \end{bmatrix}$$

and it is easily seen that the unconstrained solution

$$X = c\Sigma^{-1}\underline{\mu}$$

is a solution which satisfies the conditions $\alpha_i \in [0,\ 1]$ for $i = 1, 2, \ldots, 4$. Choosing c in such a way as to maximize the amount of business retained by the company we obtain the following optimal solution

Insurance Subportfolio	*Retention α_i*	*Expected Profit $\alpha_i \ell_i$*	*Contribution to overall Variance $Cov(-\alpha_i \tilde{X}_i, \tilde{\Delta}u)$*
Motor	1	0.5	4.47
Homeowners	0.54	0.43	3.87
Industrial Fire	0.44	0.44	3.93
GTPL	0.81	1.21	10.82
Asset Category	*Amount invested A_j*	*Expected Profit $A_j R_j$*	*Contribution to overall Variance $Cov(A_j \tilde{R}, \tilde{\Delta}u)$*
Medium bond	−69.3	−0.69	−6.20
Long bond	77.9	1.56	13.92
Equities	15.9	1.59	14.21
Real estate	8.5	0.68	6.04
		5.72	51.07

The risk return ratio is 0.80, the amount of net invested assets is 33.0 and the amount of net loss reserves is 98.3.

By perfect asset liability matching and by investing the equity into the risk free asset one can fully eliminate the asset risk. The vector of expected returns and the covariance matrix of the pure insurance risk are respectively

$$\underline{\mu}_0' = (0.5,\ 0.8,\ 1,\ 1.5)$$

and

$$\Sigma_0 = \begin{bmatrix} 6.25 & 1.6 & -2 & 2 \\ 1.6 & 10.24 & 0 & 0 \\ -2 & 0 & 16 & 0 \\ -2 & 0 & 0 & 16 \end{bmatrix}$$

and from the theorem of section 4.2 we know that the maximum risk return ratio which can be achieved in such a situation is

$$r = \left(\underline{\mu}_0' \Sigma^{-1} \underline{\mu}_0\right)^{1/2} = 0.53$$

which is considerably lower than risk return ratio obtained above. Thus, in this example too, it is seen that the assumption of asset risk leads to a considerable improvement of the risk return ratio of the portfolio.

Through quota share cessions the company has reduced the expected profit of its insurance portfolio from 3.8 to 2.58, i.e. it forgoes a substantial amount of profit in order to maximize its risk return ratio. As a comparison, we now look at the optimal portfolio assuming that the company cedes no quota share. In that case, we have the following vector of expected returns

$$\mu_1' = (3.8,\ 0.01,\ 0.02,\ 0.1,\ 0.08)$$

and covariance matrix

$$\Sigma_1 = \begin{bmatrix} 51.69 & -0.052 & -0.078 & 0.06 & 0.06 \\ -0.052 & 0.0016 & 0.00216 & 0.0032 & 0.0032 \\ -0.078 & 0.00216 & 0.0036 & 0.0048 & 0.0048 \\ 0.06 & 0.0032 & 0.0048 & 0.04 & 0.016 \\ 0.06 & 0.0032 & 0.0048 & 0.016 & 0.04 \end{bmatrix}$$

And the optimal solution excluding quota share cessions is

	a resp. A	*Expected Profit*	*Contribution to overall Variance*
Insurance Portfolio	1	3.8	50.18
medium bond ($A_1 - L$)	−104.2	−1.04	−13.75
long bond	114.0	2.28	30.10
equities	21.8	2.18	28.83
real estate	10.8	0.87	11.44
		8.09	106.79

The risk return ratio is now $r = 0.78$ which is only slightly lower than the optimal risk return ratio of 0.80. In practical circumstances an insurance company may prefer the above solution with the much higher expected profit of 8.09 (vs 5.72) to the optimal solution even if this entails a slight decrease of the risk return ratio.

The optimization method we have derived is nevertheless valuable since it provides us with a benchmark, the optimal portfolio, against which to measure any given portfolio.

5. Comparison with Other Results in Finance Theory

5.1. Markowitz's Portfolio Selection Method

The portfolio selection method presented here is based on the maximization of the same function as is used in the framework of Markowitz's mean variance method. There are however major differences. In the present model the amount of equity u supporting the business can be chosen by the company. The consequences of the introduction of this additional degree of freedom are discussed in section 2.2. The present model allows a simultanious optimization of a portfolio of risky assets and of insurance risks. The major difference between insurance and financial risks is that the latter are **easily traded** whereas the former are not. Financial risks are standardized securities for which there exist liquid and transparent secondary markets. The transaction costs are very low, the position of the company can be frequently adjusted at virtually no costs. (Hence the conditions $A_i \in (-\infty, \infty)$ or $A_i \geq 0$.) Insurance risks once taken on can only be traded on the reinsurance market which is neither liquid nor transparent. It is usually not possible to take a short position in an insurance risk. Increasing one's share of a risk beyond 100% leads to high transaction costs related to the acquisition of new blocks of business. (Hence the conditions $0 \leq \alpha_i, \beta_i \leq 1$, $i = 1, \ldots, n$.)

A further difference between insurance and asset risks is the fact that the optimization of insurance risks is a **two steps process**. Whilst it would in principle be possible to determine the optimal retention rates α_i and β_i of each policy, this would hardly be a tractable method in practice, given the fact that even a medium sized company has hundreds of thousands of customers who often buy more than one policy from the company. One has therefore to build insurance subportfolios (e.g. along lines of business and customer segments), to optimize those subportfolios individually (e.g. via surplus, and excess of loss reinsurance as illustrated in section 2) and to build an optimal global portfolio via appropriate quota share cessions. The process is therefore a two steps optimization process and the result depends on the sub portfolio structure which has been chosen.

Finally, the optimal portfolio of assets within the overall portfolio of the insurance company strongly depends on the portfolio of insurance risks. This is especially true since the loss reserve risk entails a short position in a bond portfolio. As a consequence, the portfolio of assets which pertains to the optimal overall portfolio is in general very different from the optimal portfolio of assets on a stand alone basis, as derived from Markowitz's method.

5.2. CAPM

5.2.1.

Each insurance company optimizes its overall portfolio of insurance and asset risks. The optimal portfolio of the company heavily depends on the gross insurance portfolio which varies considerably from company to company. As a consequence the optimal asset portfolios of different companies are not colinear and are different from the optimal asset portfolio according to the CAPM. Thus the optimal asset portfolio of the company is **not a market portfolio**, as in the CAPM, but a **company specific portfolio**. Given the weight of insurance companies and pension funds as institutional investors, the above result may explain why empirical evidence does not confirm the CAPM (see H.S. Houthakker and P.J. Williamson, 1996).

5.2.2.

A further difference between the CAPM and our general model is the fact that in our model insurance risks command a loading over and above the expected value of the losses they generate and this in spite of the fact that those risks are not market risks and can be diversified away. The reason why individuals are willing to pay such a loading is because they are risk averse and unable to diversify their risk. Closely held corporations are in a similar position. The case of firms with diffuse ownership is more complex. Stockholders and bondholders of such firms can diversify their claims and do not need to buy insurance. There are however other stakeholders such as employees, clients and suppliers who cannot diversify their claims. In the absence of insurance, employees and managers for instance would discount their expected future cash flows at a much higher interest rate to reflect the higher risk. It is therefore worthwhile for the firm to buy insurance even if the price is higher than the actuarially fair premium. Different other reasons such as a lowering of expected bankruptcy costs and a lowering of the company's expected tax liabilities also explain why the **'free lunch'** enjoyed by insurance companies is consistent with finance theory. For a more detailed discussion of the topic see Mayers and Smith (1982).

5.2.3.

In addition to a free lunch insurance companies also enjoy a **free loan**. The assumption of the yield curve risk as part of the loss reserve risk is tantamount to issuing a bond without having to pay any spread. This allows the company to achieve a higher risk return ratio than would be possible if it could not issue securities or if had to pay a spread.

5.2.4.

Both in the case of the CAPM and of our model the **separation theorem** holds true. The composition of the optimal portfolio follows from objective factors: the expected returns and the covariance between the returns of individual risks. The decision of how much risk to assume, i.e. the choice of a point on the efficient frontier is a subjective decision, which is separate from the selection of the optimal portfolio.

5.2.5.

Within the framework of CAPM, the expected return of asset i (R_i) and the expected return of the market portfolio (R_M) satisfy the following relationship

$$R_i - \rho_0 = \beta_i \cdot (R_M - \rho_0) \quad \text{with} \quad \beta_i = \frac{Cov(\tilde{R}_i, \tilde{R}_M)}{Var(\tilde{R}_M)}$$

Within the framework of our model (see example 3 of section 4.2) the following formulae hold true for the optimal portfolio

$$\ell_i = \frac{Cov(-\tilde{X}_i, \tilde{\Delta}u)}{Var(\tilde{\Delta}u)} \cdot \left(E(\tilde{\Delta}u) - \rho_0 u\right)$$

$$R_i - \rho_0 = \frac{Cov(\tilde{R}_i, \tilde{\Delta}u)}{Var(\tilde{\Delta}u)} \cdot \left(E(\tilde{\Delta}u) - \rho_0 u\right)$$

We can rewrite the CAPM formula as

$$\frac{(R_i - \rho_0) \cdot B_i}{Cov(\tilde{R}_i \cdot B_i, \tilde{R}_M)} = \frac{R_M - \rho_0}{Var(\tilde{R}_M)}$$

where B_i $i = 1, ..., n$ are the coefficients pertaining to the optimal portfolio, hence

$$\tilde{R}_M = \sum_{i=1}^{n} \tilde{R}_i \cdot B_i \cdot \left(\sum_{i=1}^{n} B_i \right)^{-1}.$$

On the other hand the formulae in our model can be rewritten as

$$\frac{\alpha_i \cdot \ell_i}{Cov(-\alpha_i \tilde{X}_i, \tilde{\Delta} u)} = \frac{E(\tilde{\Delta} u) - \rho_0 u}{Var(\tilde{\Delta} u)}$$

$$\frac{(R_i - \rho_0) \cdot B_i}{Cov(\tilde{R}_i B_i, \tilde{\Delta} u)} = \frac{E(\tilde{\Delta} u) - \rho_0 u}{Var(\tilde{\Delta} u)}$$

In the special case where there are only asset risks we have

$$\tilde{R}_M = \tilde{\Delta} u \cdot \left(\sum B_i \right)^{-1}$$

i.e. the optimal company portfolio and the optimal market portfolio are identical.

Setting $u = \sum B_i$, the second formula can be rewritten as

$$\frac{(R_i - \rho_0) \cdot B_i}{Cov(\tilde{R}_i B_i, \tilde{R}_M)} = \frac{R_M - \rho_0}{Var(\tilde{R}_M)}$$

and it is seen that the formulae of our model are a generalisation of the CAPM formula. Both types of formulae state that the ratio of expected profit to contribution to the overall covariance is the same for each risk. In the case of the CAPM the formula applies to asset risks only, in the case of our model it applies to asset and insurance risks. In the first case, the reference portfolio is the market portfolio, in the second case it is the company portfolio.

5.2.6. Discount Rates

Definition

The **rate of return** of the company associated with a given value u of net asset value is

$$\tilde{R}_u = \frac{\tilde{\Delta} u}{u}$$

Thereby u must be at least equal to the amount of net invested assets, i.e.

$$u \geq u_0 = \sum_{i=1}^{n} A_i - \sum_{i=1}^{m} \alpha_i L_i$$

Theorem
Assuming that the company may issue securities (Assumption 8) and that insurance risks and financial risks are uncorrelated ($Cov(\tilde{X}_j, \tilde{R}_i) = 0$ all i, j) we have

$$\tilde{R}_{u_0} = \tilde{R}_M + \frac{\sum_{i=1}^{n} \alpha_i\big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big)}{u_0}$$

where $\tilde{R}_M$ is the market rate of return for financial risks according to the CAPM and u_0 is the amount of net invested assets.

Proof
Under the assumptions of the theorem we have

$$\tilde{\Delta} u - \rho_0 u = \sum_{i=1}^{m} \alpha_i\big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big) + \sum_{j=1}^{n} (\tilde{R}_j - \rho_0) B_j.$$

For any $u \geq u_0$ (see section 4.2). Hence

$$R_{u_0} = \frac{\tilde{\Delta} u}{u_0} = \frac{\sum \tilde{R}_j \cdot B_j}{\sum B_j} + \frac{\sum \alpha_i\big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big)}{u_0}$$

since $u_0 = \sum A_i - \sum \alpha_j L_j = \sum B_j$. And since investment risks and insurance risks are uncorrelated

$$\left(\sum_j \tilde{R}_j \cdot B_j\right) \cdot \left(\sum_j B_j\right)^{-1} = \tilde{R}_M$$

q.e.d.

Remark
Under the assumptions of the theorem we have

$$\tilde{R}_u = \frac{u_0}{u} \cdot \tilde{R}_M + \frac{\sum \alpha_i\big(E(\tilde{X}_i) + \ell_i - \tilde{X}_i\big)}{u}$$

According to the CAPM, the discount rate associated with $\tilde{R}_u = \dfrac{\tilde{\Delta} u}{u}$ is

$$\tilde{R}_d(u) = \rho_0 + \frac{Cov(\tilde{R}_u, \tilde{R}_M)}{Var(\tilde{R}_M)} (\tilde{R}_M - \rho_0)$$

Assuming that insurance risks and investment risks are uncorrelated, we obtain from the above representation of $\tilde{R}_u$

$$Cov(\tilde{R}_u, \tilde{R}_M) = \frac{u_0}{u} \cdot Var(\tilde{R}_M)$$

We have thus derived the following

Corollary 1
Under the assumptions of the preceding theorem, the discount rate of the company is

$$R_d(u) = \frac{u_0}{u} R_M + \left(1 - \frac{u_0}{u}\right) \rho_0$$

Corollary 2
The value of the company is

$$\frac{E(\tilde{\Delta} u)}{R_d(u)} = u + \frac{\sum \alpha_i \ell_i}{R_d(u)}$$

Proof

$$E(\tilde{\Delta} u) = \rho_0 u + \sum_i \alpha_i \ell_i + \sum_j B_j (R_j - \rho_0)$$

and since insurance and investment risks are uncorrelated, we have

$$\sum_j B_j (R_j - \rho_0) = \left(\sum_j B_j\right)(R_M - \rho_0) = u_0 (R_M - \rho_0)$$

hence

$$E(\tilde{\Delta} u) = u_0 \cdot R_M + (u - u_0)\rho_0 + \sum \alpha_i \ell_i$$
$$E(\tilde{\Delta} u) = R_d(u) \cdot u + \sum \alpha_i \ell_i$$

which proves the corollary.

q.e.d.

The value of the company is thus the sum of its net asset value (at market prices) and of the goodwill of the company

$$G = \frac{\sum \alpha_i \ell_i}{R_d(u)} = \frac{\sum \alpha_i \ell_i}{\frac{u_0}{u} R_M + \left(1 - \frac{u_0}{u}\right) \rho_0}$$

The goodwill depends on u and it is easily seen that $G'(u) > 0$ and $G''(u) < 0$.

Remark
The goodwill

$$G = \frac{\sum \alpha_i \ell_i}{R_d(u)}$$

is the maximum value one should be willing to pay for the access to the business, i.e. for the distribution network. It depends on the amount of net asset value which supports the business, since the higher the equity u, the more valuable the excess return $\tilde{\Delta}u - \rho_0 u$.

Assuming that the amount of equity is determined based on the risk tolerance τ of the owners of the company

$$u = \frac{1}{\tau}\frac{V}{R}$$

(where $V = Var(\tilde{\Delta}u)$ and $R = E(\tilde{\Delta}u) - \rho_0 u$ for the optimal portfolio), we obtain the following discount rate

$$R_d = \frac{\tau R}{V} \cdot u_0 R_M + \left(1 - \frac{\tau R}{V} u_0\right)\rho_0$$

and the goodwill of the company is arrived at by plugging this expression into the above formula. And it is seen that the discount factor is an increasing function of the risk tolerance. Hence the goodwill is a decreasing function of the risk tolerance.

Example
Example 3 of section 4.2 satisfies the conditions of the above theorem. We have

$$u_0 = 37.4$$

$$u = \tau^{-1}\frac{V}{R} = \tau^{-1} \cdot 62.00 = 248.0, \quad \text{for} \quad \tau = 0.25$$

let $\rho_0 = 5\%$, we have

$$R_M = \frac{20.71}{37.4} + 5\% = 60.4\%$$

and we obtain

$$R_d = \frac{u_0}{u} R_M + \left(1 - \frac{u_0}{u}\right)\rho_0 = 0.151 \cdot 60.4\% + 0.849 \cdot 5\% = 13.37$$

Hence

$$G = \frac{\sum \alpha_i \ell_i}{R_d(u)} = \frac{17.8}{0.1337} = 133.2$$

ACKNOWLEDGEMENT

I am greatly indebted to one of the referees for drawing my attention to a major inconsistency and allowing me to substantially improve the presentation of this article.

REFERENCES

HOUTHAKKER, H.S. and WILLIAMSON, P.J. (1996) *The Economics of Financial Markets*. Oxford University Press.

MAYERS, D. and SMITH, C.W., Jr. (1982) On the Corporate Demand for Insurance. *Journal of Business*, Vol. 55, no. 2.

PANJER, H.H. et al. (1998) *Financial Economics*. The Actuarial Foundation.

SCHNIEPER, R. (1997) *Capital Allocation and Solvency Testing*. SCOR Notes.

SHARPE, W.F. (1970) *Portfolio Theory and Capital Markets*. McGraw-Hill.

RENÉ SCHNIEPER
Zurich Insurance Company, Reinsurance
P.O. Box, CH – 8022 Zurich
Switzerland

BOOK REVIEWS

M.A. COPPINI: *Incomes Redistribution Through Social Security.* Centro d'informazione e stampa universitaria (CISU) di Enzi Colamartini, Rome, ISBN 88 7975 201 4.

This is a recently published English translation (by David Giddings) of a book originally published in Italian 20 years ago. In spite of the time lapse, it does not appear to have been revised or updated and the references are all to work in the 1970s and earlier.

However, notwithstanding the fact that the basic principles of this work were presented as long ago as 1975 at the 6th ISSA Conference for Social Security Actuaries and Statisticians in Helsinki, it is likely that the methodologies will be unfamiliar to most actuaries, even to those working in the social security area.

The problems addressed are measuring the redistributive effects of a social security system and quantifying the effectiveness of the system in achieving redistributive objectives. The study of this sort of problem is perhaps more often associated with economists than actuaries, but the author has the advantage of being both an economist and an actuary.

Many readers may find the terminology and definitions somewhat hard to get to grips with. It may well be that some of the nuances are lost in translation, but it is often difficult to conceptualise what the notation is seeking to represent. The mathematics which follows is presented in full detail but requires careful study because of the definitional complexity.

Measuring the effects of redistribution presents many technical problems because of the complexity of the transfer of a social security system, which differ by branch (e.g. pensions, sickness, unemployment, health care, etc.) and have different impacts when looked at by individuals or by households, with effects which depend on earnings level, age, sex, duration of period of study, etc. The mathematics is complicated by the fact that most participants in the system are both contributors and beneficiaries, although not always at the same time. The author develops a generalised framework for examining and quantifying these effects and then elaborates a stochastic methodology, with conceptual roots in the insurance risk process and classical risk theory, as a way of providing practical solutions to a problem of rather daunting complexity.

Even more practical, perhaps, is the alternative approach of simulation which is offered by the author. However, dominant concerns in the book about the practicality of full simulation because of computing constraints more than anything serve to date the presentation. It seems unlikely that the application of these techniques would be much constrained today by availability of computing power, although I suspect that availability of raw data on earnings distributions, contribution density and other factors will prove more of a constraint in practical applications of the techniques.

The author does present, in a final chapter, some examples of practical applications in the context of the Italian social security system. Since the ground-breaking work on this methodology has been developed by the author and colleagues and students of his in Rome, it is not clear from the book that any equivalent studies have been carried out elsewhere, and I am not aware of a wider literature having developed since the presentation of these ideas at the Helsinki Conference and at a subsequent ISSA meeting in Rome in May 1984. The author himself points out that this is very much "work in progress", rather than a definitive text-book on the techniques. The proof of whether these techniques can be applied to improve understanding of redistributive effects in social security (and, for example, to confirm or rebut charges made by World Bank economists that traditional social security schemes do not redistribute nearly as much as it might be thought, for reasons such as differential mortality between high and low earners) will inevitably depend on further research into practical applications, most likely using simulation techniques.

CHRIS DAYKIN

Y.K. KWOK: *Mathematical Models of Financial Derivatives*. Springer Finance, Singapore, ISBN 981 3083 255 (hardcover), 981 3083 565 (softcover), 1998.

This book is described on the cover as being suitable for degree programs in mathematical and computational finance. As one who delivers a masters level course in derivative pricing to *maths* graduates I can see that this is indeed an appropriate audience. At the same time, I suspect that the typical masters student in finance with a first degree in a less numerate subject would struggle with this book.

The book is well written and maintains a consistent approach throughout. Apart from an early mention of the martingale approach to the pricing of derivatives and risk-neutral valuation the author sticks firmly with the partial differential equation (PDE) approach. Whether one should take the PDE approach or the martingale approach is really a matter for personal preference which often is the result of the what background a student or researcher comes from (applied maths or applied probability). However, my own preference is for the martingale approach, not just because of my personal background but also because the martingale approach gives much more insight into the subject. In particular, the martingale approach makes it much easier, at least initially, to tackle any new problem which is thrown at you. The book also tends to avoid rigorous technical development and this can leave students less well prepared for new, perhaps more complex derivative-pricing problems.

My overall impression of the book is, therefore, that it was not one which I would recommend to students as the core textbook in a course on derivative pricing. However, it is one which I would happily recommend as supplementary text. There are a number of reasons why I make this recommendation. First, the book, throughout, has good descriptive introductions to each topic. This carries through many of the essentially more technical sections where the author includes descriptive passages which turn an abstract problem and analysis into something more understandable. Second, each chapter ends with a comprehensive set of exercises which, again, is very useful for students wishing to reinforce what they are learning about the subject.

Chapter 1 gives a general introduction to the subject of derivative pricing, and presents essentially model-free results such as put-call parity and lower and upper bounds for prices. It then proceeds to introduce the models, tools and concepts used in the remainder of the book.

Chapters 2 to 6 deal with equity options. Chapter 2 works on the basic European option with the Black-Scholes model and formula taking centre stage. There is also what is essentially a statement of the Greeks without much intuitive explanation of what they are or how they should be used. Chapter 3 looks at multi-asset options. Chapter 4 considers how to price American options. This includes good non-technical descriptions of the various issues. Chapter 5 deals with various numerical methods for tackling these problems. It is a well written section and, of course, relates well to the dominant PDE approach in the book. Chapter 6 looks at a number of different exotic options.

Finally, Chapter 7 gives a short introduction (unfortunately common to many books in this field) to bond pricing and interest-rate derivatives.

In summary, therefore, this book is not perfect but there are many good things in it, so it is a worthwhile purchase.

ANDREW CAIRNS

E. KREMER: *Applied Risk Theory.*

This book of Erhard Kremer presents some selected chapters of risk theory. It starts with a nice 20 page summary of the basic results of probability theory.

Chapter two gives an overview about the common premium principles and proves their basic properties. Special emphasis is put on the "Swiss premium principle".

Chapter three deals with classical credibility theory. The reader might actually be mislead by the title "rating theory", since beside credibility theory no other rating methods are discussed.

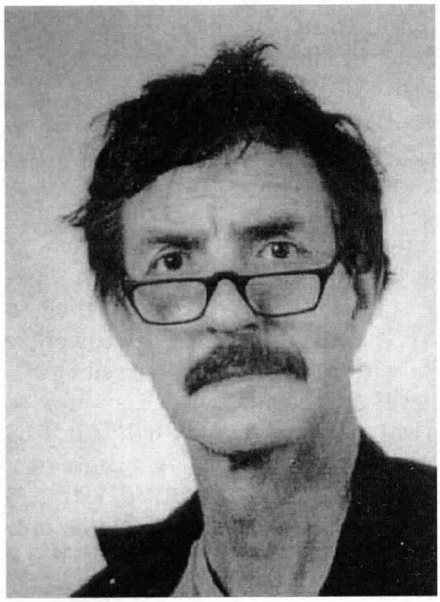
In the introduction the author mentions that in his view reinsurance is one of the main and nicest topics in Risk Theory. Unfortunately, my high expectations raised by these comments were not fully met. Although the material treated in chapter four covers some important concepts, e.g. Panjer's algorithm or optimality of reinsurance covers, the link to practical applications is fully missing. Beside the classical material I would also have expected some words on today's hot topics like extreme value theory or modelling of correlated risks.

The last chapter gives a short introduction to some methods used in life insurance, including applications of martingale theory.

In my opinion this book can be of value for actuarial students in order to quickly get a first idea about risk theory and some important actuarial principles. I wouldn't recommend it, however, to a practical actuary, who is usually interested in quite different types of questions.

PETER ANTAL

BOB ALTING VON GEUSAU
1946-1999



Bob Alting von Geusau died on November 4, 1999, only 53 years old. Though optimistic to the last, he knew he was terribly ill; with a final effort, he had given his goodbye lecture as a professor of actuarial science only a few weeks before. This impressive lecture, aptly called "The survival of the fittest", essentially described the plans he had had for the remainder of his career. It was attended by some 600 of his friends, colleagues and former students. It was a sad occasion, but endurable because of the way Bob handled his predicament, making everybody laugh at his stories and anecdotes.

By students and colleagues alike, he will be sorely missed at the actuarial department of the University of Amsterdam. Apart from being an excellent and inspiring teacher and a really pleasant colleague, he was also outstanding at promoting actuarial science in the Netherlands. He persuaded many students to choose our profession, by presenting them with exotic and inspiring tales at their schools while they were trying to select a career. A recurrent part of these sessions was "Around the world in 80 questions". A prospective student would be asked to name any country in the world, and Bob produced an interesting story about insurance relating to this particular country. For these, he could draw from his years of experience. He managed to make actuarial science sound so much more interesting than accountancy!

Bob has been involved with the department of actuarial science for a period of around 30 years. First as an assistent of the late Jaap van Klinken, then, after a period outside the university, as his successor. In the period between, he worked among others with the reinsurance company NRG, and later started his own actuarial bureau. He was a former chairman of the Dutch Actuarial Society. Also, he was the founder and for a long period the chairman of the section AFIR.

Apart from being the best PR-officer a university department could wish, his most important characteristic was perhaps that he was incredibly versatile: he has made contributions to actuarial science in life and in non-life insurance, as well as of course in the field in which he was appointed a professor, social security and pension funds. But he was also a pioneer promoting AFIR in the Netherlands. While the other lecturers at the department are mainly scientists of the "ivory tower" type, Bob has always maintained close links with actuarial practice; his job at the university was only part-time. Whenever a journalist needed an impartial and well-founded academic opinion on some matter newsworthy, Bob was able to give one. Invariably, he had presided some working group or conference attending just this matter a few years ago, had supervised a student writing his master's thesis about the subject, or had simply found the subject interesting. The Dutch Actuarial Society asked him to serve a second term as its chairman, in the period they celebrated their 100th anniversary. He was a really outstanding public speaker, and great at organizing meetings where he could display this ability.

Astineers will fondly remember Bob from many ASTIN-colloquia, presenting a paper or being the chairman of a session. He was not only fluent in English and French, but also quite at home in Italian and some other languages. When he did give a talk in English, he pleased the part of the audience not quite so fluent in this language by using transparencies in French. One time, giving a talk on an application of compound Poisson distributions in Lausanne, in the few minutes allotted to him he meticulously explained to the audience the ins and outs of a traditional Dutch family game "sjoelbakken", a variant of shovel-board. In later years, his interests had shifted from ASTIN-subjects to the topics he was teaching, the Dutch system of social security and pension funds.

It will be hard to fill his position at our department. Our hearts go out to his wife Hedwig and their children Karen and Niels.

Rob Kaas

GUIDELINES TO AUTHORS

1. Papers for publication should be sent in quadruplicate to the Editor:

Paul Embrechts,
Department of Mathematics, ETHZ,
CH-8092 Zurich, Switzerland.
embrechts@math.ethz.ch

or to one of the Co-Editors:

René Schnieper,
Zurich Insurance Company,
P.O. Box, CH-8022 Zurich, Switzerland.
rene.schnieper@zurichre.com

Andrew Cairns
Department of Actuarial Mathematics
and Statistics
Heriot-Watt University
Edinburgh EH 14 4AS, United Kingdom.
A.Cairns@ma.hw.ac.uk

Submission of a paper is held to imply that it contains original unpublished work and is not being submitted for publication elsewhere.
Receipt of the paper will be confirmed and followed by a refereeing process, which will take about three months.

2. Manuscripts should be typewritten on one side of the paper, double-spaced with wide margins. The basic elements of the journal's style have been agreed by the Editors and Publishers and should be clear from checking a recent issue of *ASTIN BULLETIN*. If variations are felt necessary they should be clearly indicated on the manuscript.

3. Papers should be written in English or in French. Authors intending to submit longer papers (e.g. exceeding 30 pages) are advised to consider splitting their contribution into two or more shorter contributions.

4. The first page of each paper should start with the title, the name(s) of the author(s), and an abstract of the paper as well as some major keywords. An institutional affiliation can be placed between the name(s) of the author(s) and the abstract.

5. Footnotes should be avoided as far as possible.

6. Upon acceptance of a paper, any figures should be drawn in black ink on white paper in a form suitable for photographic reproduction with lettering of uniform size and sufficiently large to be legible when reduced to the final size.

7. References should be arranged alphabetically, and for the same author chronologically. Use a, b, c, etc. to separate publications of the same author in the same year. For journal references give author(s), year, title, journal (in italics, cf. point 9), volume (in boldface, cf. point 9), and pages. For book references give author(s), year, title (in italics), publisher, and city.

Examples

BARLOW, R.E. and PROSCHAN, F. (1975) *Mathematical Theory of Reliability and Life Testing*. Holt, Rinehart, and Winston, New York.

JEWELL, W.S. (1975a) Model variations in credibility theory. In *Credibility: Theory and Applications* (ed. P.M. KAHN), pp. 193–244, Academic Press, New York.

JEWELL, W.S. (1975b) Regularity conditions for exact credibility. *ASTIN Bulletin* **8**, 336–341.

References in the text are given by the author's name followed by the year of publication (and possibly a letter) in parentheses.

8. The address of at least one of the authors should be typed following the references.

Continued overleaf.

Guidelines to Authors *continued from inside back cover.*

9. Italics (boldface) should be indicated by single (wavy) underlining. Mathematical symbols will automatically be set in italics, and need not be underlined unless there is a possibility of misinterpretation.
   Information helping to avoid misinterpretation may be listed on a separate sheet entitled 'special instructions to the printer'. (Example of such an instruction: Greek letters are indicated with green and script letters with brown underlining, using double underlining for capitals and single underlining for lower case.)

10. Electronic Typesetting using Word Perfect 5.1 and Word 97 is available. Authors who wish to use this possibility should ask one of the editors for detailed instructions.

11. Authors will receive from the publisher two sets of page proofs together with the manuscript. One corrected set of proofs plus the manuscript should be returned to the publisher within one week. Authors may be charged for alterations to the original manuscript.

12. Authors will receive 50 offprints free of charge. Additional offprints may be ordered when returning corrected proofs. A scale of charges will be enclosed when the proofs are sent out.