AGLM as an Area of Investigation

Suguru & Iwahiro
From Japan

May 20th, 2021

About the Speakers

Suguru Fujita, FIAJ, CERA

- Guy Carpenter Japan, Inc.
- Life (3yr) -> Non-Life (1.5yr) -> Reinsurance (3.5yr)
- M.S./B.E. – Applied Mathematics

Iwahiro (Hirokazu Iwasawa), FIAJ

- Teacher of actuarial science
- Guest Professor of Waseda University, etc.
- Wrote 9 math books, among others
- Board member of JARIP

IAJ: Institute of Actuaries of Japan

- ASTIN-related study group
- Data Science-related research group

公益社団法人 日本アクチュアリー会

Think the Future, Manage the Risk

Agenda

1. Introduction

2. AGLM Tour

3. Further Voyage

1. Introduction

- What is AGLM?
- AGLM Project
- History and Development

What is AGLM?

Our proposed model, which is..

- A hybrid modeling method of **GLM** and **Data Science techniques**

- Aiming for well-balanced model in terms of both **Interpretability** and **Prediction accuracy**

AGLM Project

Current team members – 6 actuaries!

- Suguru Fujita
- Tsudoi Kaminaka
- Toyoto Tanaka
- Takahiro Kobayashi
- Kazuhisa Takahashi
- Hirokazu Iwasawa

Special thanks to Kenji Kondo as the author and maintainer of the R package for AGLM

Activities:

- Research Development
- Writing Papers
- R Package
- Presentation/ Workshops

History and Development

- 2017

AGLM Idea

2018 -

Project Launched!

IAJ Conference

公益社団法人 日本アクチュアリー会
Think the Future, Manage the Risk

2019 -

R Package

Joint ASTIN/CAS Seminar

(Tokyo)

2020 -

IAA Sections Virtual Colloquium

IAJ Workshop

SECTIONS VIRTUAL COLLOQUIUM 2020

2021 - Present

ASTIN Online Colloquium

We are here!

Future

Next chapter..??

2. AGLM Tour

- **Definition**
- **Model Pipeline**
- **Non-linearity Treatment**
- **R Package `aglm`**

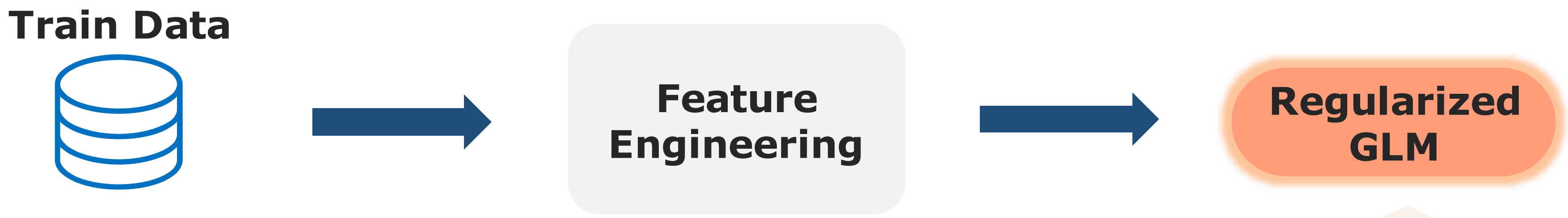
Definition

- AGLM consists of three techniques:

Regularized GLM + **Discretization** + **O Dummy/L Variables**

- What does '**A**' stand for?
  - "Accurate" - expect higher prediction accuracy than GLM
  - "Actuarial," "Accountable," etc. - see it as a somewhat symbolic letter representing other words as well

Model Pipeline



Notation -

$\boldsymbol{y}$	Response variable
$\boldsymbol{x}$	Features
$\boldsymbol{\beta}$	Regression coefficients
n	# observations
p	# features
g	Link function
L	Likelihood function

GLM:

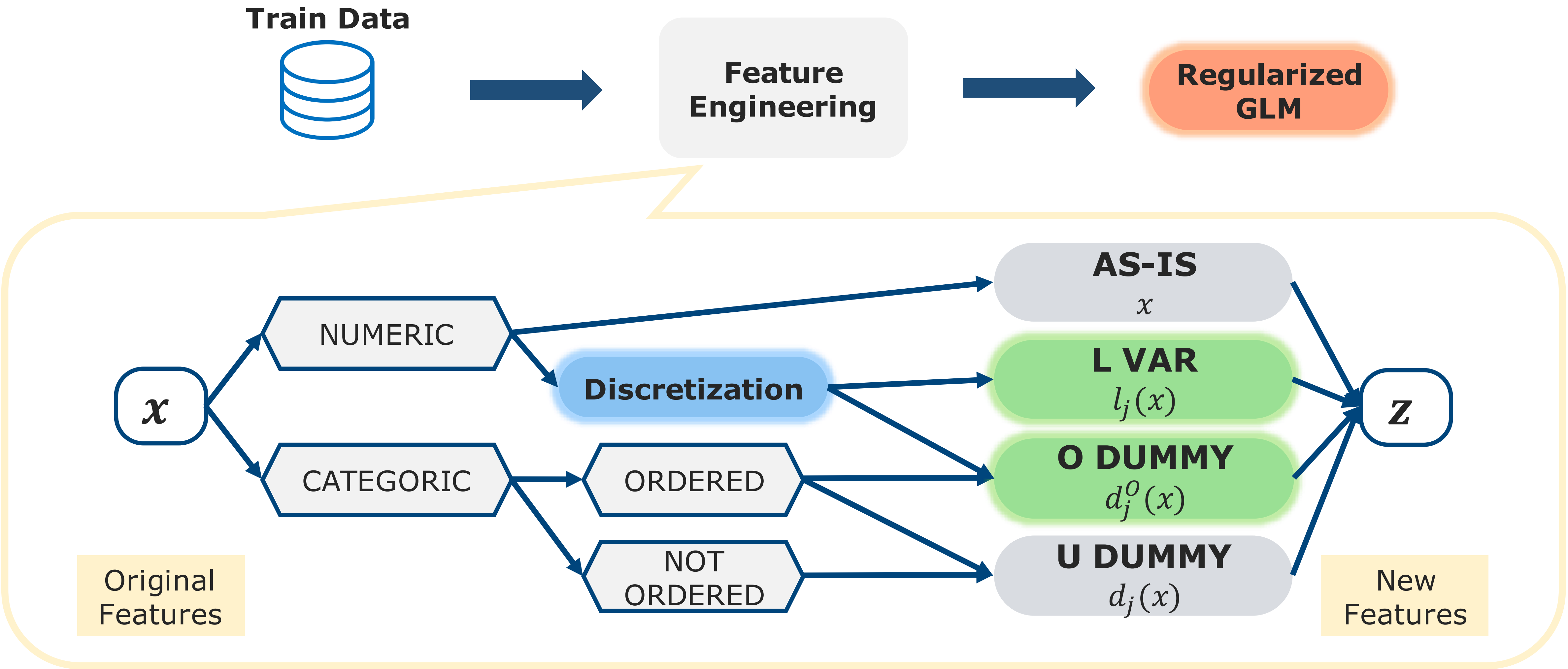
$$\mathrm{E}[y_i] = g^{-1}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \quad (i = 1, \cdots, n)$$

Optimization:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{-\log L(\boldsymbol{\beta}) + R(\boldsymbol{\beta};\lambda)\}$$

$R(\boldsymbol{\beta};\lambda)$: regularization term (lasso, ridge, elastic net, etc.)

Model Pipeline



Non-linearity Treatment

O Dummy Variables

- Elaborate on dummy variables

Assume the discretized feature x takes m levels $\{1,2,\cdots,m\}$ $(\ni j)$

U (Usual) Dummy

$$d_j(x) = \begin{cases} 1 & \text{if } x = j; \\ 0 & \text{otherwise.} \end{cases}$$

O (Ordinal) Dummy

$$d_j^{\mathrm{O}}(x) = \begin{cases} 1 & \text{if } x > j; \\ 0 & \text{otherwise.} \end{cases}$$

X	$d_1(x)$	$d_2(x)$	$d_3(x)$	$\cdots$	$d_{m-1}(x)$	$d_m(x)$
1	1	0	0	$\cdots$	0	0
2	0	1	0	$\cdots$	0	0
3	0	0	1	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$m-1$	0	0	0	$\cdots$	1	0
m	0	0	0	$\cdots$	0	1

X	$d_1^{\mathrm{O}}(x)$	$d_2^{\mathrm{O}}(x)$	$d_3^{\mathrm{O}}(x)$	$\cdots$	$d_{m-1}^{\mathrm{O}}(x)$	$d_m^{\mathrm{O}}(x)$
1	0	0	0	$\cdots$	0	0
2	1	0	0	$\cdots$	0	0
3	1	1	0	$\cdots$	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$m-1$	1	1	1	$\cdots$	0	0
m	1	1	1	$\cdots$	1	0

Non-linearity Treatment

L Variables

- Further elaborate on numerical features

Assume the discretized feature x takes m levels $\{1,2,\cdots,m\}$ $(\ni j)$

where b_j is boundary of bin

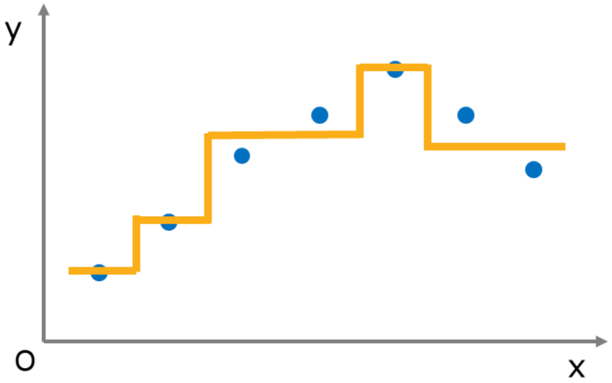
L (Linear) Variables

$$l_j(x) = \begin{cases} |x - b_j| & (j = 1, \ldots, m-1); \\ x & (j = 0 \text{ as a linear term}). \end{cases}$$

To illustrate..

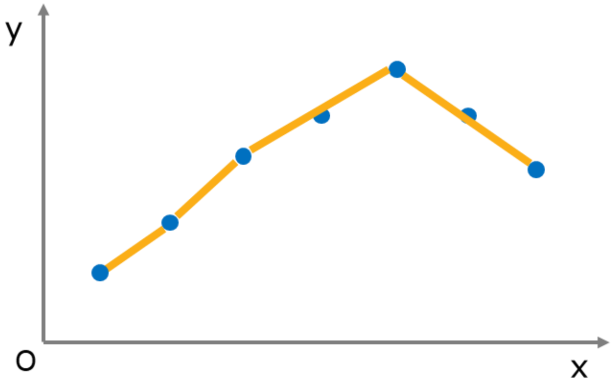
O Dummy: Step-wise Function

$$x \mapsto \sum_j \beta_j d_j^O(x)$$



L Variables: Piece-wise Linear Function

$$x \mapsto \sum_j \beta_j l_j(x)$$



Non-linearity Treatment

- AGLM virtually covers the existing regularization terms:

O Dummy + L1 Regularization -> Fused Lasso Effect

L Variables + L1 Regularization -> Trend Filter-like Effect

Will deep-dive into this topic in the next section! ☺

R Package `aglm`

A handy R package for AGLM *(since Jan. 2019)*

- GitHub - https://github.com/kkondo1981/aglm

What's New

- Wider range of distributions are available (incl. Gamma/Negative binomial/Tweedie) with the update of `glmnet` ver.4.0* *(May 2020)*

Will give an R demo later during the break! ☺

* https://glmnet.stanford.edu/articles/glmnet.html

3. Further Voyage

Advantages to be extended

- From the viewpoint of implementation, AGLM is a set of modeling methods realized by using the `glmnet` algorithm efficiently.
- Thus, AGLM has the following advantages:
  - Resulting models can be constructed reliably and relatively very fast.
  - As well as L1 and L2 regularizations, the Elastic Net regularization is available from the beginning.
  - Since May in 2020, all GLM families of distributions are available, including actuaries' favorite Gamma distribution and Tweedie distribution.
- The advantages can be vastly extended by adding varieties of simple devices to the present `aglm`.

Two approaches to expand AGLM

- Recall that, in the cases of O Dummy and L Variable, there are two steps in implementation of them,

  i) binning and ii) regularization.
- For the first step, varieties of feature engineering other than binning may lead to new methods.
- For the second step, there is, in fact, a common form to be noted.
- So, there are two kinds of approaches:
  i. Other feature engineering than binning
  ii. A common form → To be discussed first in what follows

A common form of expanding AGLM

- Our problem has the general form:

$$\min_{\beta} \; -\frac{1}{n}\ell(y, X\beta) + \lambda\|h(\beta)\|$$

  Here ℓ is a log-likelihood function and $\|\cdot\|$ is a norm of the Elastic Net including L1 and squared L2. The resulting model is called "Generalized Lasso" when, typically assuming normally distributed and homoscedastic, the norm is L1 and $h(\beta)$ is of $D\beta$ where D is a matrix.

- Generally speaking, the `glmnet` can be used as the backend for expanding AGLM if there is a vector γ and a kind of design matrix X' such that

$$\left(\min_{\gamma} \; -\frac{1}{n}\ell(y, X'\gamma) + \lambda\|\gamma\|\right) = \left(\min_{\beta} \; -\frac{1}{n}\ell(y, X\beta) + \lambda\|h(\beta)\|\right).$$

Regular Generalized Elastic Net

- When $h(\beta) = D\beta$ with some regular matrix D, let $\gamma = D\beta$ and $X' = XD^{-1}$, then

$$\left(\min_{\gamma} -\frac{1}{n}\ell(y, X'\gamma) + \lambda\|\gamma\|\right) = \left(\min_{\beta} -\frac{1}{n}\ell(y, X\beta) + \lambda\|D\beta\|\right).$$

- E.g.,

$$D = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -1 & 1 & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix} \Rightarrow D^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 1 & \cdots & \cdots & 1 \end{pmatrix}$$

(O dummies for ordered variables)

- Find a meaningful regular matrix D in data analytics, you'll get a nice fast regularized GLM modeling method.

Generalized (Generalized) Ridge

- For the L2 regularization, it's not required that D is regular but only that $\mathrm{rank}(D) \geq p$. In this case, let $\gamma = D\beta$ and $X' = XD^{+}$ in which D^{+} represents the pseudo-inverse of D, then
$$\left(\min_{\gamma} -\frac{1}{n}\ell(y, X'\gamma) + \lambda\|\gamma\|_2^2\right) = \left(\min_{\beta} -\frac{1}{n}\ell(y, X\beta) + \lambda\|D\beta\|_2^2\right).$$
- It means that the AGLM approach allows us to obtain any Generalized Ridge model corresponding to a Generalized Lasso with any GLM family distribution in a simple and reasonable way. For another approach, refer to the literature of "Laplacian Filter".

Examples of regular matrices for D

$D=(d_{ij})$	Existing methods with similar effects	Other notes
$d_{ij}=\begin{cases}1 & (i=j)\\ -1 & (i=j+1)\\ 0 & (\text{Others})\end{cases}$	Fused Lasso (L1 norm) AGLM's O Dummy	Same as O Dummy for an ordered variable
$d_{ij}=\begin{cases}-2 & (i=j)\\ 1 & (i=j\pm 1)\\ 0 & (\text{Others})\end{cases}$	Trend Filter (L1 norm) Hodrick-Prescott Filter (Squared L2 norm) AGLM's L variable	
$d_{ij}=\begin{cases}1 & (i=j=1)\\ 0 & (i=1\neq j,\ j=1\neq i)\\ -\sum_{k=2}^{p} a_{i,k} & (i=j\neq 1)\\ a_{i,j} & (\text{Others})\end{cases}$	Graph Trend Filter (L1 norm) Laplacian Filter (Squared L2 norm)	$A=(a_{ij})$ is the adjacency matrix

An example idea for Generalized Ridge —Dealing with periodicity

- Suppose a periodic variable has p levels. Then the following D's may be nice candidates for Generalized Ridge to deal with periodicity of the variable.
  - $d_{ij} = \begin{cases} 1 & (i = j) \\ -1 & (i \equiv j + 1 \pmod{p}) \\ 0 & (\text{Others}) \end{cases}$
  - $d_{ij} = \begin{cases} -2 & (i = j) \\ 1 & (i \equiv j \pm 1 \pmod{p}) \\ 0 & (\text{Others}) \end{cases}$

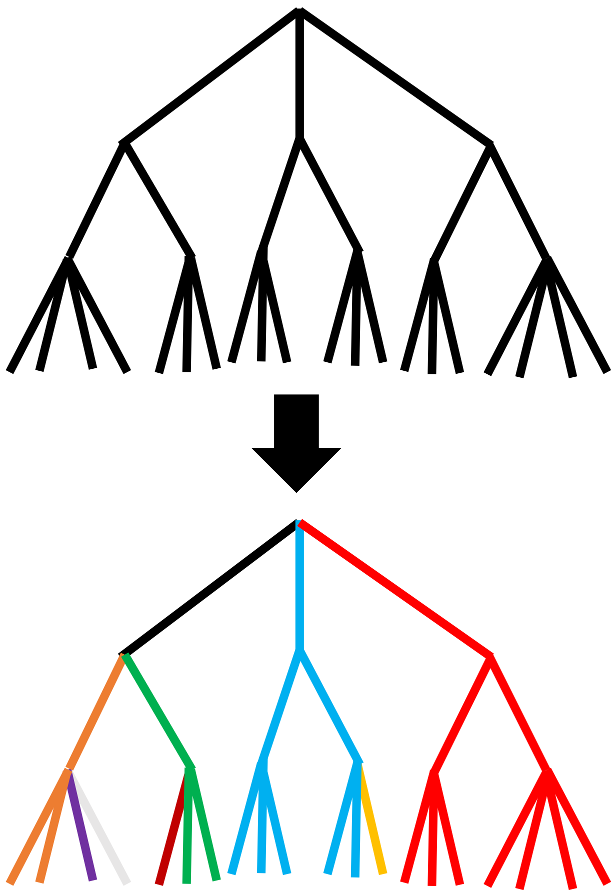
An example idea of feature engineering —Dealing with hierarchy

- Thanks to the function of variable selection via regularization, simple one-hot encoding for a hierarchical structure may work well.
- E.g., suppose there are three layers for a variable, say, a vehicle type with company, brand, and model. Then three sets of variables:

$$x_i = \begin{cases} 1 & (\text{company} = i\text{th company}) \\ 0 & (\text{Others}) \end{cases}$$

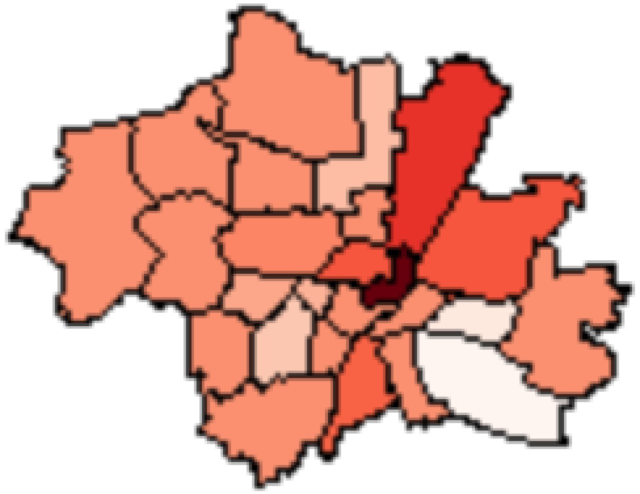
$$x_{ij} = \begin{cases} 1 & (\text{brand} = j\text{th brand of } i\text{th company}) \\ 0 & (\text{Others}) \end{cases}$$

$$x_{ijk} = \begin{cases} 1 & (\text{model} = k\text{th model of } i\text{th company}, j\text{th brand}) \\ 0 & (\text{Others}) \end{cases}$$

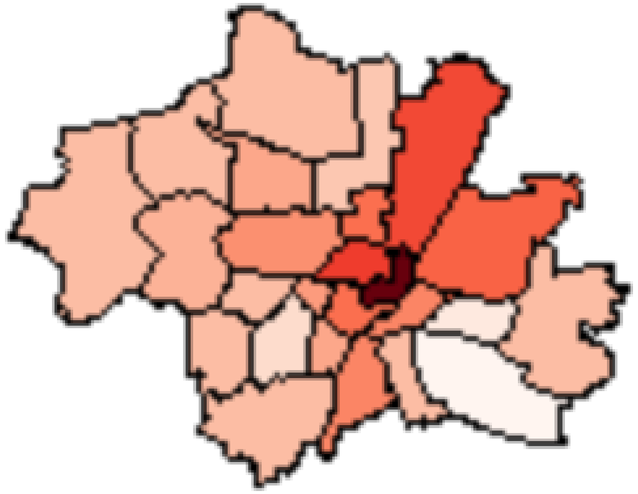


Application examples for spatial information

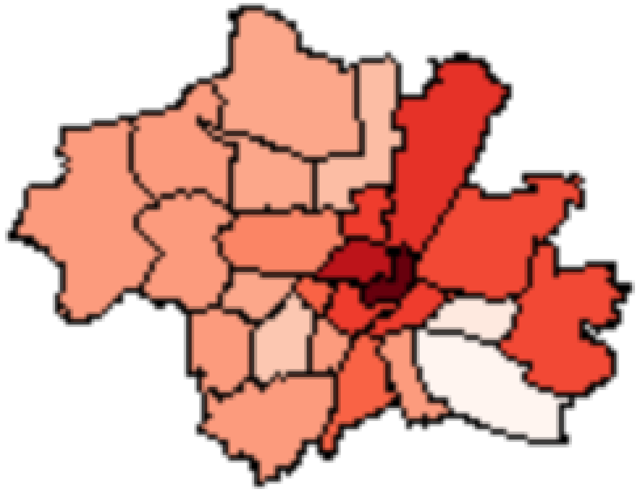
No spatial information



Graph-Trend-Filter-like



With hierarchical structure



- The dataset used is `catdata::rent` in CRAN.
- The response variable is `rent`. Explanatory variables are all others but `rentm` in the dataset and each area's population density from another source.
- All three models uses L1 norm and respectively select λ via cross validation.
- The hierarchical structure adopted here is a tentative one without domain knowledge.

Conclusion

- We hope you enjoyed the AGLM trip!
- Our conclusion message is:

"Would you like to go on an AGLM voyage with us?"

- Please feel free to tell your interest, or ask any question to
  suguru.fujita@guycarp.com and/or iwahiro@bb.mbn.or.jp

Thank you! ☺