An information-geometrical path algorithm for Poisson regression

Yoshihiro Hirose
The University of Tokyo, Tokyo, Japan - hirose@mist.i.u-tokyo.ac.jp

Abstract

We consider Poisson regression for counting data and compare three estimation methods. The first method, *bisector regression* (BR), was proposed by our previous work, which is based on the information geometry of the statistical model. The second, *differential geometric least angle regression* (DGLARS), is another method based on the information geometry. The third is the l_1-regularization method. These methods are related with the least angle regression (LARS) algorithm, which is an important algorithm in the normal linear regression. BR and DGLARS are directly motivated by LARS and they are extensions of LARS for settings more than the normal linear regression. In the normal linear regression setting, the l_1-regularization method, *lasso*, is known to be closely related with LARS. Three methods output the sequence of parameter estimates, each estimate of which is corresponding to a submodel. This fact means that the methods select a model in addition to parameter estimation. In the paper, we compare three methods by analyzing datasets and numerical experiments. The methods make similar but different results.

Keywords: bisector regression; information geometry; least angle regression; regularization.

1. Introduction

We consider Poisson regression for counting data and compare three estimation methods. The first method, *bisector regression* (BR), was proposed by Hirose & Komaki (2010), which is based on the information geometry of the statistical model. The second, *differential geometric least angle regression* (DGLARS, Augugliaro et al. 2013), is another method based on the information geometry. The third is the l_1-regularization method (Park & Hastie 2007). BR has not been compared with DGLARS in the literature.

In Hirose & Komaki (2010, 2013), we extended the least angle regression (LARS) algorithm (Efron et al. 2004) using the information geometry of dually flat spaces. The LARS algorithm estimates parameter values and simultaneously selects explanatory variables in the normal linear regression. The LARS algorithm is described in terms of the geometry of the Euclidean space spanned by explanatory variable vectors. In the LARS algorithm, the bisectors and distance in the Euclidean space play an important role. BR uses a curve in a dually flat space corresponding to a bisector in the Euclidean space.

Recently, other information-geometrical approaches for LARS were given in Yukawa & Amari (2011), Amari & Yukawa (2013) and Augugliaro et al. (2013). In Amari & Yukawa (2013), the BR algorithm was compared theoretically with the method proposed by Amari & Yukawa (2013).

The connection of LARS with the l_1-regularization method, *lasso* (Tibshirani 1996), is known in the normal linear regression. Park & Hastie (2007) proposed methods to make paths of estimated parameters in the generalized linear regression.

The information geometry (Amari & Nagaoka 2000; Kass & Vos 1997) is a generalization of the Euclidean geometry and a dually flat space is a generalization of the Euclidean space. In a dually flat space, geodesics and divergence correspond to straight lines and distance in the Euclidean space, respectively. The model manifold of an exponential family of distributions is known to be a dually flat space. An exponential family of distributions appears naturally in the generalized linear regression. One of advantages of using the information geometry is that it captures the nature of the problem in a geometric way. For example, the lasso method is based on just an optimization problem, which cannot capture the essence of the problem. It is the case for various extensions of lasso. BR makes an invariant estimate with respect to scale transformation of variables while lasso is essentially based on the scale of variables.

In this paper, we compare three methods for Poisson regression, which are BR, DGLARS and l_1-regularization. In Section 2, we describe the problem treated and the BR algorithm briefly. The minimal preliminaries and explanation are presented. In Section 3, we show the results of BR for counting data. We compare the result

of BR with those of other methods. In Section 4, we present the conclusions.

2. Problem and estimation method

In Subsection 2.1, the problem treated in the paper is presented. We give a short explanation on the information geometry in Subsection 2.2. In Subsection 2.3, we describe the BR algorithm briefly. For the detail of the BR algorithm including the geometrical aspect, see Hirose & Komaki (2010).

2.1. Poisson regression

The data $\{\mathbf{y}, \mathbf{X}\}$ is given, where $\mathbf{y} = (y_1, \dots, y_n)^\top \in \{0, 1, 2, \dots\}^n$ is a response variable vector, $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^n)^\top$ $(1 \leq i \leq d)$ is a explanatory variable vector, and $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_d) = (x_i^a)_{1 \leq a \leq n, 1 \leq i \leq d}$ is a design matrix. The response $\mathbf{y}$ consists of counting data for n samples. The design matrix $\mathbf{X}$ consists of d explanatory variables for n samples. We assume $n \geq d + 1$.

In the Poisson regression setting, we assume that each components of $\mathbf{y}$ is independently distributed according to Poisson distributions. The following relationship is assumed: $\log \boldsymbol{\lambda} = \mathbf{X}\boldsymbol{\theta} + \theta^0 \mathbf{1} = \widetilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}$, where $\mathbf{1} = (1, \dots, 1)^\top \in \mathbf{R}^n$, $\log \boldsymbol{\lambda} = (\log \lambda_1, \log \lambda_2, \dots, \log \lambda_n)^\top$, $\boldsymbol{\lambda} = (\lambda_a)_{1 \leq a \leq n} = (\mathrm{E}[y_a])_{1 \leq a \leq n}$ is the expectation of $\mathbf{y}$, $\widetilde{\mathbf{X}} = (\mathbf{X}|\mathbf{1})$ is the extended design matrix, $\boldsymbol{\theta} = (\theta^1, \dots, \theta^d)^\top \in \mathbf{R}^d$ and $\theta^0 \in \mathbf{R}$ are parameters, and $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\top, \theta^0)^\top \in \mathbf{R}^{d+1}$.

Let $\mathcal{F}$ denote the set of all Poisson distributions, that is, $\mathcal{F} = \{p(\cdot|\, \boldsymbol{\lambda}) | \, \boldsymbol{\lambda} \in (0, \infty)^n\}$, where $p(\mathbf{y}|\, \boldsymbol{\lambda}) = \prod_{a=1}^n e^{-\lambda_a} \lambda^{y_a} / y_a!$ is the probability function of a Poisson distribution. Let $\mathcal{S}$ be the full model of Poisson regression, that is, $\mathcal{S} = \left\{ p(\cdot|\, \boldsymbol{\lambda}) \middle| \, \log \boldsymbol{\lambda} = \widetilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}, \, \tilde{\boldsymbol{\theta}} \in \mathbf{R}^{d+1} \right\} \subseteq \mathcal{F}$. Our aim is to estimate the parameter $\tilde{\boldsymbol{\theta}}$ with simultaneously selecting explanatory variables.

2.2. Information geometry for estimation

We briefly introduce some tools of the information geometry. For details, refer to Amari & Nagaoka (2000) and Kass & Vos (1997). Our main tools are geodesics connecting two points and a projection onto a submodel in addition to the Kullback–Leibler divergence.

We introduce some notations. We define $\boldsymbol{\xi} = (\xi^a)_{1 \leq a \leq n} \in \mathbf{R}^n$ with $\xi^a = \log \lambda_a$, $\boldsymbol{\eta} = \mathbf{X}^\top \boldsymbol{\lambda} \in \mathbf{R}^d$, and $\tilde{\boldsymbol{\eta}} = \widetilde{\mathbf{X}}^\top \boldsymbol{\lambda} \in \mathbf{R}^{d+1}$. The parameter $\boldsymbol{\xi}$ is the natural parameter and $\boldsymbol{\lambda}$ is the expectation parameter for $\mathcal{F}$. These parameters are different way to identify a Poisson distribution. It is known that $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ are useful coordinate systems in $\mathcal{F}$. Strictly speaking, $\boldsymbol{\xi}$ is an e-affine coordinate system and $\boldsymbol{\lambda}$ is an m-affine coordinate system in terms of the information geometry. Similarly, it is also known that $\tilde{\boldsymbol{\theta}}$ is an e-affine coordinate system and $\tilde{\boldsymbol{\eta}}$ is an m-affine coordinate system in $\mathcal{S}$. As is the following, affine coordinate systems enable us to treat the problem as a problem described in terms of the Euclidean geometry. Pairs of parameters, $(\boldsymbol{\xi}, \boldsymbol{\lambda})$ and $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}})$, are useful tools. For $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$, $p(\cdot|\, \boldsymbol{\xi})$ and $p(\cdot|\, \boldsymbol{\lambda})$ indicate the same Poisson distribution in $\mathcal{F}$. Similarly, $p(\cdot|\, \tilde{\boldsymbol{\theta}})$ and $p(\cdot|\, \tilde{\boldsymbol{\eta}})$ indicate the same Poisson distribution in $\mathcal{S}$.

We introduce some definitions in terms of $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ in $\mathcal{F}$. The Kullback–Leibler divergence $D_{\mathcal{F}}(\boldsymbol{\lambda}|\, \boldsymbol{\lambda}')$ between two Poisson distributions, $p(\cdot|\, \boldsymbol{\lambda})$ and $p(\cdot|\, \boldsymbol{\lambda}')$, is defined by

$$D_{\mathcal{F}}(\boldsymbol{\lambda}|\, \boldsymbol{\lambda}') = \sum_{a=1}^n \lambda_a \left(\log \lambda_a - \log \lambda_a' \right) - \sum_{a=1}^n \left(\lambda_a - \lambda_a' \right).$$

Equivalently, for $p(\cdot|\, \boldsymbol{\xi})$ and $p(\cdot|\, \boldsymbol{\xi}')$, it holds that

$$D_{\mathcal{F}}(\boldsymbol{\xi}|\, \boldsymbol{\xi}') = \sum_{a=1}^n e^{\xi^a} \left(\xi^a - \xi'^a \right) - \sum_{a=1}^n \left(e^{\xi^a} - e^{\xi'^a} \right).$$

The Kullback–Leibler divergence is corresponding to the square of the Euclidean distance. The m-geodesic $l_{\mathrm{m}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ connecting $p(\cdot|\, \boldsymbol{\lambda})$ and $p(\cdot|\, \boldsymbol{\lambda}')$ is defined by

$$l_{\mathrm{m}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \{ p(\cdot|\, \boldsymbol{\lambda}(t)) | \, \boldsymbol{\lambda}(t) = t\boldsymbol{\lambda} + (1 - t)\boldsymbol{\lambda}', \, t \in [0, 1] \}.$$

Similarly, the e-geodesic $l_{\mathrm{e}}(\boldsymbol{\xi}, \boldsymbol{\xi}')$ connecting $p(\cdot|\, \boldsymbol{\xi})$ and $p(\cdot|\, \boldsymbol{\xi}')$ is defined by

$$l_{\mathrm{e}}(\boldsymbol{\xi}, \boldsymbol{\xi}') = \{ p(\cdot|\, \boldsymbol{\xi}(t)) | \, \boldsymbol{\xi}(t) = t\boldsymbol{\xi} + (1 - t)\boldsymbol{\xi}', \, t \in [0, 1] \}.$$

The m-projection $p(\cdot|\,\bar{\boldsymbol{\lambda}})$ of $p(\cdot|\,\boldsymbol{\lambda}) \in \mathcal{F}$ onto $\mathcal{F}' \subset \mathcal{F}$ is defined by

$$\bar{\boldsymbol{\lambda}} = \arg \min_{\boldsymbol{\lambda}' :\, p(\cdot|\,\boldsymbol{\lambda}') \in \mathcal{F}'} D(\boldsymbol{\lambda}|\,\boldsymbol{\lambda}').$$

We introduce some submodels in $\mathcal{S}$ for our purpose. Suppose that we have observed $\{\mathbf{y}, \mathbf{X}\}$. The submodel $\mathcal{M}$ is defined as $\mathcal{M} = \{p(\cdot|\,\boldsymbol{\eta})|\,\eta_0 = (\hat{\boldsymbol{\eta}}_{\mathrm{MLE}})_0\} \subseteq \mathcal{S}$ depending on the data $\{\mathbf{y}, \mathbf{X}\}$, where $\hat{\boldsymbol{\eta}}_{\mathrm{MLE}}$ is the maximum likelihood estimator (MLE) of the full model. Introducing $\mathcal{M}$ is corresponding to estimate the intercept in the normal linear regression setting. We define the submodel $\mathcal{M}(I) \subseteq \mathcal{M}$ for $I \subseteq \{1, 2, \ldots, d\}$ as

$$\mathcal{M}(I) = \left\{ p(\cdot|\,\boldsymbol{\xi}) \middle|\, \boldsymbol{\xi} = \mathbf{X}_I \boldsymbol{\theta}_I + \theta^0 \mathbf{1},\, \boldsymbol{\theta}_I \in \mathbf{R}^{|I|}, \eta_0 = (\hat{\boldsymbol{\eta}}_{\mathrm{MLE}})_0 \right\},$$

where $\mathbf{X}_I = (x_i^a)_{1 \le a \le n, i \in I}$ and $\boldsymbol{\theta}_I = (\theta^i)_{i \in I}$. The subset I indicates which variables are left for constructing the estimator in the BR algorithm.
As is the case of $\mathcal{F}$, we define the Kullback–Leibler divergence, geodesics, and m-projection in terms of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\eta}}$ in $\mathcal{S}$. The Kullback–Leibler divergence $D_{\mathcal{S}}(\tilde{\boldsymbol{\theta}}|\,\tilde{\boldsymbol{\theta}}')$ between $p(\cdot|\,\tilde{\boldsymbol{\theta}})$ and $p(\cdot|\,\tilde{\boldsymbol{\theta}}')$ is defined by

$$D_{\mathcal{S}}(\tilde{\boldsymbol{\theta}}|\,\tilde{\boldsymbol{\theta}}') = D_{\mathcal{F}}(\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}|\,\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}'),$$

where the right-hand side is written in terms of $\boldsymbol{\xi}$. The m-geodesic $l_{\mathrm{m}}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}}')$ connecting $p(\cdot|\,\tilde{\boldsymbol{\eta}})$ and $p(\cdot|\,\tilde{\boldsymbol{\eta}}')$ is defined by

$$l_{\mathrm{m}}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\eta}}') = \{p(\cdot|\,\tilde{\boldsymbol{\eta}}(t))|\,\tilde{\boldsymbol{\eta}}(t) = t\tilde{\boldsymbol{\eta}} + (1-t)\tilde{\boldsymbol{\eta}}',\, t \in [0,1]\}.$$

Similarly, the e-geodesic $l_{\mathrm{e}}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')$ connecting $p(\cdot|\,\tilde{\boldsymbol{\theta}})$ and $p(\cdot|\,\tilde{\boldsymbol{\theta}}')$ is defined by

$$l_{\mathrm{e}}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}') = \left\{ p(\cdot|\,\tilde{\boldsymbol{\theta}}(t))|\,\tilde{\boldsymbol{\theta}}(t) = t\tilde{\boldsymbol{\theta}} + (1-t)\tilde{\boldsymbol{\theta}}',\, t \in [0,1] \right\}.$$

The m-projection $p(\cdot|\,\bar{\boldsymbol{\theta}})$ of $p(\cdot|\,\tilde{\boldsymbol{\theta}}) \in \mathcal{S}$ onto $\mathcal{M}(I)$ is defined by

$$\bar{\boldsymbol{\theta}} = \arg \min_{\tilde{\boldsymbol{\theta}}' :\, p(\cdot|\,\tilde{\boldsymbol{\theta}}') \in \mathcal{M}(I)} D(\tilde{\boldsymbol{\theta}}|\,\tilde{\boldsymbol{\theta}}'),$$

which is corresponding to the MLE.

2.3. BR algorithm

The BR algorithm works in the submodel $\mathcal{M}$ introduced in Subsection 2.2, and it is sufficient to treat only $\boldsymbol{\theta}$ rather than $\tilde{\boldsymbol{\theta}}$. The intercept θ^0 is deceided by the restriction of $\tilde{\boldsymbol{\theta}} \in \mathcal{M}$. The BR algorithm outputs a sequence $\{\hat{\boldsymbol{\theta}}_{(k)}\}_{0 \le k \le d}$ of estimates of the parameter $\boldsymbol{\theta}$. The estimate $\hat{\boldsymbol{\theta}}_{(k)}$ includes k zeros as its components, which means that BR selects explanatory variables. The first estimate is $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, the MLE of the full model S, and, in general, includes no zero component. It is known that the MLE is contained by the submodel $\mathcal{M}$. The BR algorithm iteratively eliminate one variable of the estimator, which means that one component of the parameter becomes zero at each iteration. The elimination is based on the projection onto a submodel and the intersection of submodels. The number of the iterations is that of explanatory variables.
The BR algorithm is as follows. Steps 2 to 6 are iterated. Note that a point $\boldsymbol{\theta}$ indicates the Poisson distribution $p(\cdot|\,\boldsymbol{\theta}) \in \mathcal{M}$. The index set I indicates which variables are used for constructing the estimator. k is the iteration number and indicates how many elements are zero. The submodel $\mathcal{M}(i, I)$ is the submodel using only variables corresponding to $I \setminus \{i\}$, that is, $\mathcal{M}(i, I) = \mathcal{M}(I \setminus \{i\})$.

1. Set $I = \{1, 2, \ldots, d\}$, $\hat{\boldsymbol{\theta}}_{(0)} := \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, and $k = 0$.
2. Set $\mathcal{M}(i, I) = \{p(\cdot|\,\boldsymbol{\theta})|\,\theta^i = 0, \theta^j = 0\,(j \notin I)\} = \mathcal{M}(I \setminus \{i\})\,(i \in I)$ and calculate the m-projection $\bar{\boldsymbol{\theta}}(i, I)$ of $\hat{\boldsymbol{\theta}}_{(k)}$ onto $M(i, I)$.
3. Calculate $t^* = \min_{i \in I} D^{[I]}(\hat{\boldsymbol{\theta}}_{(k)}|\,\bar{\boldsymbol{\theta}}(i, I))$ and $i^* = \arg\min_{i \in I} D^{[I]}(\hat{\boldsymbol{\theta}}_{(k)}|\,\bar{\boldsymbol{\theta}}(i, I))$.
4. For $i \in I$, let $l_{\mathrm{m}}(i, I)$ be the m-geodesic connecting $\hat{\boldsymbol{\theta}}_{(k)}$ and $\bar{\boldsymbol{\theta}}(i, I)$, and calculate $\boldsymbol{\theta}(t^*, i, I) \in l_{\mathrm{m}}(i, I)$ satisfying $D(\hat{\boldsymbol{\theta}}_{(k)}|\,\boldsymbol{\theta}(t^*, i, I)) = t^*$.

5. Set $\hat{\theta}^i_{(k+1)} = \theta^i(t^*, i, I)\,(i \in I)$ and $\hat{\theta}^j_{(k+1)} = 0\,(j \notin I)$.
6. If $k+1 = d-1$, then go to step 7. If $k+1 < d-1$, then go to step 2 with $k := k+1, I := I \backslash \{i^*\}$.
7. Set $\hat{\boldsymbol{\theta}}_{(d)} = \mathbf{0}$ and quit the algorithm.

In step 2, the m-projection is corresponding to the MLE for each $\mathcal{M}(i, I)$ if the k-th estimate $\hat{\boldsymbol{\theta}}_{(k)}$ is regarded as the data. The submodel $\mathcal{M}(i^*, I)$ defined in step 3 is the nearest submodel from $\hat{\boldsymbol{\theta}}_{(k)}$. The minimization includes only one variable. In step 4, the bisection method algorithm is available for searching $\boldsymbol{\theta}(t^*, i, I)$ because this process also includes only one variable. The next estimate $\hat{\boldsymbol{\theta}}_{(k+1)}$ is defined in step 5, which is the intersection of submodels. Note that $\hat{\theta}^{i^*}_{(k+1)} = \theta^{i^*}(t^*, i^*, I) = 0$ in step 5. It means that one component of $\hat{\boldsymbol{\theta}}$ becomes 0 in each iteration and that BR selects variables in addition to parameter estimation. The BR algorithm runs until all explanatory variables are excluded from the index set I. The number of the iteration is that of the explanatory variables.

3. Examples

In Subsection 3.1, we analyze the real data. In Subsection 3.2, the results of numerical experiments are given. We compare BR with DGLARS and l_1-regularization. BR and DGLARS are based on the information geometry and the third one is not. In Augugliaro et al. (2013), the connection of DGLARS and l_1-regularization is pointed out. The result of DGLARS is calculated by the function `dglars.fit()` in the R package `dglars`. The result of l_1-regularization is calculated and output by functions in the R package `glmpath`. For the detail of DGLARS and l_1-regularization, refer to Augugliaro et al. (2013) and Park & Hastie (2007).

3.1. Data analysis

We analyzed two datasets by three methods. The first dataset is available in the software R. The second dataset is used in Fahrmeir et al. (2013).

First, the `gala` dataset in the R package `faraway` is analyzed, which is the data on the species diversity on the galapagos islands. The dataset consists of six explanatory variables and one response for 30 samples (islands). Variables are as following; x_1 is the number of endemic species, x_2 is the area of the island, x_3 is the highest elevation of the island, x_4 is the distance from the nearest island, x_5 is the distance from Santa Cruz island, x_6 is the area of the adjacent island, and y is the number of plant species found on the island. The results of three methods are presented in Figures 1, 2 and 3. The horizontal axis indicates l_1-norm of each estimate $|\hat{\boldsymbol{\theta}}|$ and the vertical axis indicates parameter values $\hat{\theta}^i$. Each line, which is an approximation to a theoretical curve, indicates the change of a regression coefficient, that is, a path. In the BR procedure, the estimator starts from the MLE of the full model corresponding to the right-most estimate in Figure 1 and finally reaches the origin, the left-most estimate. The explanatory variables are eliminated in turn of x_6, x_5, x_3, x_2, x_4, x_1. In the DGLARS procedure, the estimator starts from the origin corresponding to the left-most estimate in Figure 2. The explanatory variables are added in turn of x_1, x_4, x_6, x_2, x_5, x_3 as moving from left to right in the figure. In the l_1-regularization, the estimator moves from the left-most, the origin, to the right-most, the MLE of the full model, as the Lagrangian parameter becomes larger. The explanatory variables are added in turn of x_1, x_4, x_2, x_6, x_3, x_5. The BR paths are different from others although they all are roughly similar. The DGLARS and l_1-regularization results are similar as pointed out in Augugliaro et al. (2013). However, it should be noted that the two methods selected different sets of variables in their procedures.

Second, we analyze the `patentdata` dataset, which is the data on the patent opposition. The dataset consists of eight explanatory variables and one response for 4866 samples. Variables are as following; x_1 is the patent opposition ("yes"= 1 or "no"= 0), x_2 is the patent from biotech/pharma ("yes"= 1 or "no"= 0), x_3 is the US twin patent exists ("yes"= 1 or "no"= 0), x_4 is the patent holder from the USA ("yes"= 1 or "no"= 0), x_5 is the patent holder from Germany, Switzerland, or Great Britain ("yes"= 1 or "no"= 0), x_6 is the grant year, x_7 is the number of designated states for the patent, x_8 is the number of claims, and y is the number of citations for the patent. The results of three methods are presented in Figures 4, 5 and 6. In the BR procedure, the explanatory variables are eliminated in turn of x_4, x_3, x_7, x_5, x_8, x_2, x_1, x_6 as moving from right to left in the figure. In the DGLARS procedure, the explanatory variables are added in turn of x_6, x_1, x_8, x_2, x_5, x_4, x_3, x_7 as moving from left to right in the figure. In the l_1-regularization procedure, the

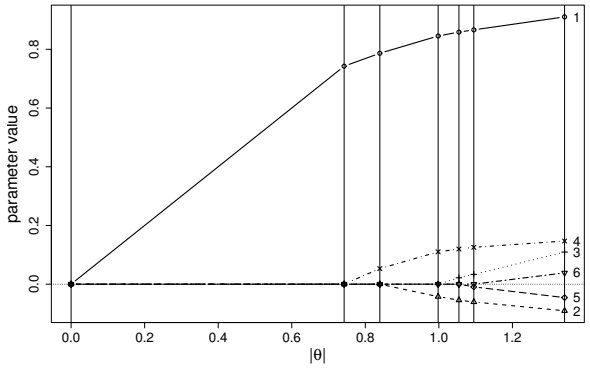

Figure 1: BR for `gala`.

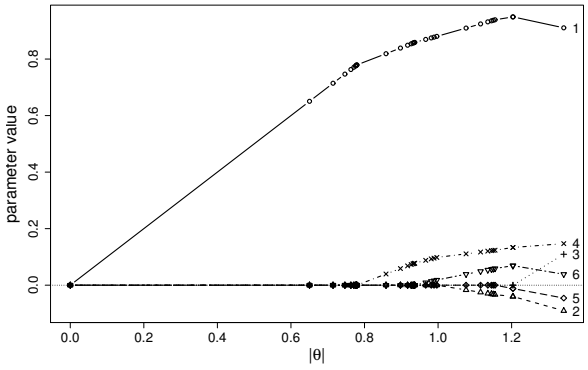

Figure 2: DGLARS for `gala`.

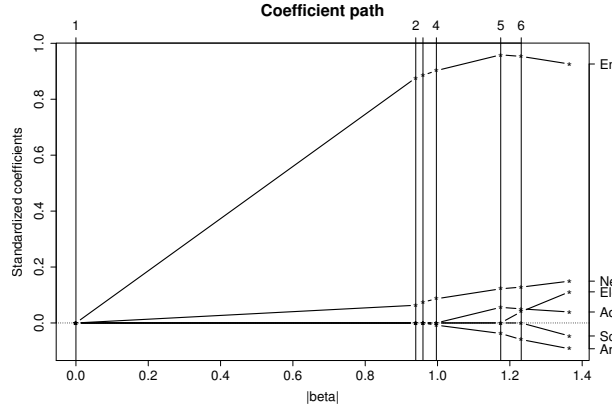

Figure 3: l_1-regularization for `gala`.

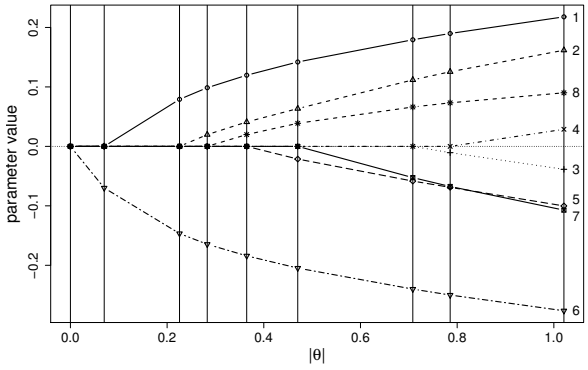

Figure 4: BR for `patentdata`.

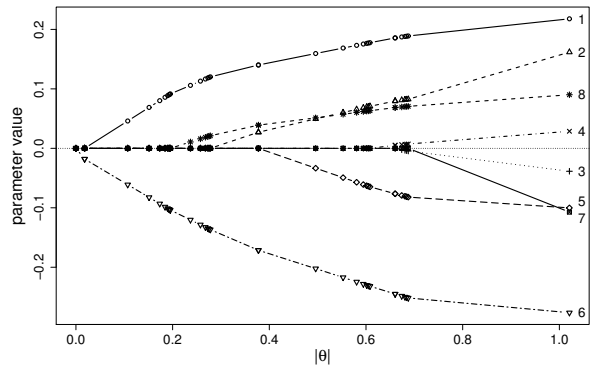

Figure 5: DGLARS for `patentdata`.

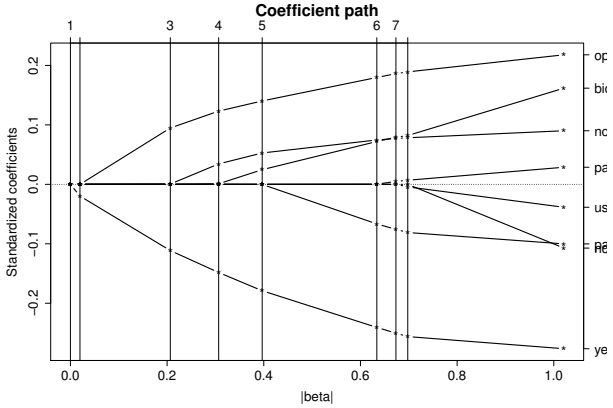

Figure 6: l_1-regularization for `patentdata`.

explanatory variables are added in turn of x_6, x_1, x_8, x_2, x_5, x_4, x_3, x_7 as moving from left to right in the figure. Three results are a little different, which is similar to the trend observed in the result for the `gala` dataset. The order of variables added are the same for DGLARS and l_1-regularization.

3.2. Numerical experiments

We give the result of numerical experiments. We are interested in whether three methods select explanatory variables correctly in their paths.

We set $d = 5$, $n = 10$ and the true parameter $\boldsymbol{\theta} = (1, 1, 1, 0, 0)^\top$. A design matrix and a response vector were generated randomly 1000 times, and three methods were applied. The sequence of the BR estimates included the true model $\{x_1, x_2, x_3\}$ for 241 trials of total 1000 trials. The DGLARS sequence included the true set for 114 trials of 1000 trials. The l_1-regularization sequence contained the set for 71 trials of 1000 trials. This means that BR is better than others. For another set of design matrices and responses, DGLARS and l_1-regularization worked better than BR. The result of variable selection varied much depending on simulation datasets made randomly. Three methods sometimes worked very well. Each of them succeeded for over 900 trials of 1000 trials.

We tried the same procedure for other settings, which included the case of the larger size of design matrices and responses. For example, considered is the case that $d = 20$, $n = 100$ and $\boldsymbol{\theta} = (1, 1, 1, 1, 1, 0, 0, \ldots, 0)^\top$. The similar trend described above was observed. As is implied by the result in Subsection 3.1, three methods do not necessarily selects a similar set of variables although they generate similar paths.

4. Conclusions

We treated Poisson regression for counting data and compared three methods. The first method, bisector regression (BR) proposed in Hirose & Komaki (2010), is motivated by the LARS algorithm proposed in Efron et al. (2004). The second is DGLARS proposed in Augugliaro et al. (2013), which is a different approach to LARS based on the information geometry. The third is the l_1-regularization method for the generalized linear models (Park & Hastie 2007), which is also related with LARS.

We presented the BR algorithm briefly for the Poisson regression setting. The BR algorithm captures the nature of the problem in a geometrical way. Note that, unlike l_1-regularization and other regularization methods, a geometrical approach is robust for the scaling of variables.

The results of three methods were given for two datasets. Three methods made similar paths while they actually came from different ideas. The result of BR was a little different from others. From the view point of variable selection, three methods selected different sets of explanatory variables although they generated similar paths. Through numerical experiments, we examined whether the methods select the correct set of variables. As was indicated by the results for the datasets, the methods selected different sets of variables.

It is one of our future works to clarify the connection between BR and DGLARS theoretically. The method in Amari & Yukawa (2013) is also related with them. On the application of the BR algorithm, it is possible for BR to treat not only the standard Poisson regression but also the zero-truncated Poisson regression with a slight modification.

References

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry.* Oxford University Press.

Amari, S., & Yukawa, M. (2013). Minkovskian gradient for sparse optimization. *IEEE J. Sel. Top. Signal. Process.*, **7**, 576–585.

Augugliaro, L., Mineo, A. M., & Wit, E. C. (2013). Differential geometric least angle regression: A differential geometric approach to sparse generalized linear models. *J. R. Stat. Soc. B*, **75**, 471–498.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Stat.*, **32**, 407–499.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications.* Springer.

Hirose, Y., & Komaki, F. (2010). An extension of least angle regression based on the information geometry of dually flat spaces. *J. Comput. Graph. Stat.*, **19**, 1007–1023.

Hirose, Y., & Komaki, F. (2013). Edge selection based on the geometry of dually flat spaces for Gaussian graphical models. *Stat. Comput.*, **23**, 793–800.

Kass, R., & Vos, P. (1997). *Geometrical foundations of asymptotic inference.* John Wiley.

Park, M. Y., & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *J. R. Stat. Soc. B*, **69**, 659–677.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Yukawa, M., & Amari, S. (2011). On extensions of lars by information geometry: Convex objectives and l_p-norm. In *APSIPA Annu. Summit. Conf: Special Session on Recent Advances in Adaptive/Sparse Signal Processing.*