Model Selection and Active Learning for High-Dimensional Data Analysis

梅津, 佑太

https://doi.org/10.15017/1654664

出版情報：九州大学, 2015, 博士（機能数理学）, 課程博士
バージョン：

Model Selection and Active Learning for High-Dimensional Data Analysis

Yuta Umezu

Graduate School of Mathematics

Kyushu University

2016

Acknowledgment

I have been able to complete this doctoral thesis with a lot of support from many people. First of all, I would like to express my sincere gratitude of my supervisor, Professor Yoshiyuki Ninomiya, for his valuable comments, suggestions and encouragement through the whole of my research.

I would like to acknowledge valuable advice and various helps given by Professor Ryuei Nishii, Yoshihiko Maesono, Hiroki Masuda and the members of group of statistics in Kyushu University.

I also would like to thank Professor Hidefumi Kawasaki of Kyushu University and Dr. Hidetoshi Matsuoka and Dr. Hiroshi Ikeda of FUJITSU LABORATORIES LTD. for my internship.

Finally, I would like to give special thanks to my parents for their enormous encouragement and financial assistance, and my friends for enormous support and encouragement.

Yuta Umezu

Janualy, 2016, Fukuoka

Contents

Chapter 1

Introduction

From the 1990s onward, high-dimensional data analysis has been known as an important problem in a number of fields, such as mathematical statistics and machine learning. The "high-dimensional" means that the number of the feature is large compared to the sample size, and it is difficult to apply the classic statistical technique. Let us consider, for example, that estimating the population mean $\boldsymbol{\mu}$ based on p-dimensional independent samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$. Then the sample mean $\bar{\boldsymbol{X}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$ is a simple estimator of the population mean $\boldsymbol{\mu}$ and is a good estimator satisfying a unbiasedness and consistency when p is small. However, assuming that the population variance is I_p for the sake of the simplicity, the mean squared error of $\bar{\boldsymbol{X}}$ is given by

$$\mathrm{E}[\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|^2] = \frac{p}{n},$$

and thus we cannot assure the consistency of $\bar{\boldsymbol{X}}$ when p or p/n is large. Moreover, since it is rare to evaluate an expectation explicitly in practice, we need to evaluate it numerically. Unfortunately, to evaluate the expectation, we need to evaluate the p-dimensional integral, and handling the high-dimensional data becomes more difficult also from the computational perspective.

The genome sequence data using the next-generation sequencer and image data may be widely known as typical examples of such high-dimensional data. Then the problem that mentioned in the above becomes critical; such a data could have features from several tens of thousand to several million although the sample size is not so much large. Moreover,

another example of the high-dimensional data is that the questionnaire data using to construct the recommendation system in many Web applications (e.g, Amazon, Netflix, iTunes, and MovieLens). Such a kind of data is a matrix consisting of the user and item, and has many missing values. Considering the recent spread of the internet and the development of the Web marketing, there are considerable amount of users and items, i.e., the data we should consider is a large-scale matrix, and so it is also difficult to deal such a data from the point of view of the estimation accuracy and computation time.

Basic ideas to overcome such a problem are the following:

- To increase the samples by experiments.
- To decrease the number of features by removing the unnecessary variables.

The former is the idea coping with the high-dimensional data by increasing the sample size. Accordingly, it is expected that the estimation accuracy can be improved though the dimension of the variable is large. On the other hand, it is better that the number of sampling times are less whenever possible because it takes time and cost to experiment. Such kind of methods are well known as the experimental design method or active learning method. These methods are kinds of the sampling method in order to compensate for the lack of the sample size against the dimension of the variables, and permit an efficient sampling in some sense. The experimental design method is pioneered by R. A. Fisher in the 1920s for agricultural applications of statistical methods, and has been studied for many years (Fienberg and Hinkley 1980). At the time, it had been actively that the study of the experimental design method for the analysis of variance model by using the latin square. Then following the study of Kiefer (1959, 1961) or Kiefer and Wolfowitz (1959) for a linear regression model, the experimental design method which can be applicable for the non-linear model such as the generalized linear model (McCullagh and Nelder 1983) has been developed. Active learning method is an expression used in the context of machine learning, and fundamentally is the sampling method same as the experimental

design method. By using such an idea, Umezu et al. (2015) proposed the simple active learning method for estimating a distribution of defective product obtaining in the product manufacturing process evaluated for its defect rate.

The latter is the idea that makes an inference by using several features being relevant to the data of p-dimensional variable. This is a kind of the variable selection problem or model selection problem, and we need to explore the number of combination of 2^p candidates usually. To avoid such a computational difficulty, the sparse regularization method is frequently used. Concretely speaking, sparse regularization method is a technique for parameter estimation imposing a penalty term, which is not differentiable at the origin, on an estimating function and the estimates is called a sparse solution. One of an advantage of the sparse regularization method is that we can do parameter estimation and variable selection simultaneously. The Lasso (Tibshirani 1996) imposing an ℓ_1 norm as a penalty term to induce a sparse solution is an elementary and important one. Moreover, the penalty term of the Lasso is a convex function, and thus we can solve the Lasso problem as a convex optimization problem if an estimating function is also convex. However, the efficiency of the parameter estimation based on the Lasso is not necessarily large, because the Lasso shrinks the estimator to the zero vector too strongly. It is known that the Bridge (Frank and Friedman 1993), SCAD (Fan and Li 2001), and MCP (Zhang 2010) can avoid such a problem although the penalty term of these methods is no longer convex. Recently, Loh and Wainwright (2015) showed that it is possible to ensure an estimation accuracy for a non-convex optimization problem such as SCAD and MCP even if the estimates are local optima, and the study on these methods would become more and more important. A trace norm, which is an ℓ_1 penalty about the singular values of matrix, is known as a penalty inducing the sparsity, and is used for a low-rank matrix reconstruction. This norm is also known as a convex relaxation of a rank restriction, and so we can use a commonly convex optimization problem like the Lasso. Such regularization method has get a lot of attention since it is useful for constructing a recommendation system (see,

e.g., Candès and Recht 2009; Koltchinskii et al. 2011). Another important problem of the sparse regularization method is to choose a tuning parameter which is included in the estimator. Umezu and Ninomiya (2015) have considered a sparse regularization method such as the Bridge, SCAD, and MCP for generalized linear model, and have proposed to select a tuning parameter by minimizing an AIC-type information criterion.

In this thesis, we attempt to overcome the above mentioned problems by using the statistical methods that can be available for high-dimensional data. The rest of the thesis is organized as follows. In **chapter 2**, we consider a defect rate, which is one of a problem in the product manufacturing, and an active learning method. After describing a background and introducing some discriminant methods, we point out that even a modern active learning method such as ASVM (Tong and Koller 2002) is inappropriate for our data. Then the efficient sampling method to obtain a discriminant boundary, which classify good and defective products, is proposed. Our proposed active learning method here is simple but is very useful in this kind of data. Moreover, after checking the validity of our method through simulation studies and real data analysis, we evaluate a defect rate by using a importance sampling.

We model a questionnaire data obtained from evaluating the product of company and derive a model selection criterion for this model in **chapter 3**. Concretely speaking, we consider a model, which can be thought of some extension of the generalized multivariate analysis of variance model to 3 dimensional array data, for the evaluated values of several items. Then we see that the estimator is given explicitly by using a dummy matrix representing the existence of the missing value, although the data have many missing values. In addition, we consider the Ridge-type estimator to assure the stability of the estimator, and propose a C_p-type information criterion as an unbiased estimator of a standardized mean squared error to select tuning parameters.

In **chapter 4**, we consider a tuning parameter selection problem which is important in sparse regularization method. The tuning parameter is a positive constant controlling

the penalty level, and many of the estimators are shrunken to 0 exactly, when it is large. Therefore the objective selection of the tuning parameter is a severe problem because data analysts can choose it arbitrarily and do model selection as they want. In chapter 4, we consider a generalized linear model same as Ninomiya and Kawano (2014). Then we derive an asymptotic distribution of the penalized maximum likelihood estimator using a non-convex penalty including the Bridge, SCAD, and MCP, and propose an AIC-type information criterion as an asymptotic unbiased estimator of the Kullback-Leibler divergence.

Because the bias correction term of information criteria are depend on an unknown value with respect to the true parameter, we usually plug in the consistent estimator of it in order to define such criteria. However, it is difficult for an ℓ_1-type regularization method such as the SCAD or MCP to derive a consistent estimator of a bias correction term of the AIC obtained in chapter 4. Then in **chapter 5**, we consider a condition assuring the consistency of the AIC for non-concave penalized likelihood method. To be concrete, we focus on an oracle property (Fan and Li 2001; Fan and Peng 2004). We describe that the bias correction term of AIC is reduced to an active set of the parameter when the estimator has a sparsity, and in addition, its consistent estimator is given by an active set of the estimator when the estimator has a variable selection consistency. The sparsity and variable selection consistency are one of a desirable property and are defined in chapter 5.

The several matrix algebra, which is frequently used in chapter 3, and proofs are relegated to **chapter 6**.

Chapter 2

Defect Rate Evaluation via Simple Active Learning

In the preparatory stage of product manufacturing, its defect risk is often evaluated by checking whether experimentally manufactured products cause the defect or not. The experimentally manufacturing is conducted for various values of variables which may related the defect, but manufacturing products for all combinations of the values will cost a lot especially when the number of variables is large. To overcome this problem, active learning methods which may be able to evaluate the defect risk efficiently by selecting values purposefully are considered. In this chapter, it is pointed out that even a modern active learning method is inappropriate if the nonlinearity of the relation between the variables and the defect is strong and if the defect rate is small. And then a simple active learning method which can work well for such a case is proposed. Through simulation studies and real data analysis, the validity of the proposed method is checked.

2.1 Background

Let us consider a product with the risk of having a defect at the time of manufacturing. We assume that the risk depends on the values of various variables such as the temperature, the amount of an ingredient or operating time. In particular, letting $y_i = -1$ and $y_i = 1$ mean that the i-th product respectively has and does not have a defect and letting $\boldsymbol{x}_i$ ($\in \mathcal{D} \subset \mathbb{R}^p$) be the values of such p variables, we assume that the defect risk for $\boldsymbol{x}_i$ is given

by

$$\mathrm{P}(y_i = -1|\boldsymbol{x}_i) = \frac{\exp(h(\boldsymbol{x}_i))}{1 + \exp(h(\boldsymbol{x}_i))}$$

for an appropriate function $h(\boldsymbol{x})$. In product manufacturing, it is indispensable to evaluate $\mathrm{P}(y_i = -1|\boldsymbol{x}_i)$ because we can avoid to yield defects which may cause a severe damage in manufacturing company if we know $\mathrm{P}(y_i = -1|\boldsymbol{x}_i)$ (see e.g., Katayama et al. 2010; Sun and Li 2014). Let us consider the appropriate function as $h(\boldsymbol{x})$. Needless to say, the products are manufactured not to have any defect, and so y tends to be one if $\boldsymbol{x}$ is in the central zone of the domain $\mathcal{D}$. That is, the region $\{\boldsymbol{x} \mid h(\boldsymbol{x}) > 0\}$ should be small and at the edge of $\mathcal{D}$. In addition, considering that the cause of the defect is usually multiple, such regions, which produces a defect more likely than not, tend to be scattered at the edge of $\mathcal{D}$. Namely, such regions should not only have a strongly nonlinear boundary but also tend to be separated although they will not be far from each other. Therefore, we suppose that $h(\boldsymbol{x})$ has a strong nonlinearity although it will not have a drastic fluctuation.

Under this situation, we consider to estimate the decision rule $\mathrm{sgn}(h(\boldsymbol{x}))$ along with the defect risk by sampling $\{(y_i, \boldsymbol{x}_i) \mid \boldsymbol{x}_i \in \mathcal{X};\ i = 1, 2, \ldots, n\}$ appropriately. Here $\mathcal{X}$ $(\subset \mathcal{D})$ is a set of candidates of $\boldsymbol{x}_i$ we can sample. This is $\mathcal{D}$ itself in some cases and a finite set in other cases. As a versatile method for giving nonlinear decision rules, recently the SVM (support vector machine; e.g., Cristianini and Shawe-Taylor 2000; Scholkopf and Smola 2001) becomes a standard tool. In addition, the Gaussian process regression method (e.g., Rasmussen and Williams 2006) is known to have comparable performance. Although these methods are capable of dealing with strong nonlinearity, a lot of samples are required to deal with it as a matter of course. This requirement becomes evident when the dimension of $\boldsymbol{x}$ is large. Therefore, an appropriate selection of samples from $\mathcal{X}$ is important to estimate $\mathrm{sgn}(h(\boldsymbol{x}))$ efficiently for the case where the sampling cost is not necessarily low and we can select samples purposefully. This type of appropriate selection is called optimal design in classic statistical area and active learning in machine learning

area.

While the active learning versions for the above-mentioned nonlinear discriminant methods are not sufficiently developed, the ASVM (active learning SVM) proposed by Tong and Koller (2002) gets a lot of attention. The ASVM, an active learning method specialized for the SVM, can be implemented easily for considerably huge data and has a high computational efficiency. Roughly speaking, however, the sampling scheme in the ASVM is to select samples close to the decision boundary estimated by already gotten data, and it is often the case that the ASVM does not work well for our problem owing to the sampling scheme. Also the method in Umezu and Ninomiya (2013) does not work well for our problem although it was proposed to overcome the weak point of the ASVM. Our main purpose is to show the lack of the development in active learning methods for a kind of discriminant problems even though they are simple and general problems. In addition, as the first step of the development, we try to provide a simple and computer-efficient method capable of such discriminant problems.

The rest of the chapter is as follows. In Sections 2.2 and 2.3, we will introduce the above-mentioned nonlinear discriminant methods and active learning methods, respectively. In Section 2.4, after explaining why such methods are not suitable for our problem, we will propose a simple but effective method. The method is shown to be valid through a simulation study in Section 3.4, and then we report the result in which the method is applied to real data in Section 3.5. We suggest how to evaluate the rate of producing defects in Section 2.7, and some concluding remarks are presented in Section 2.8.

2.2 Existing discriminant methods

In this section, we introduce the SVM and the Gaussian process regression method for discriminant problems, which is also known as Gaussian process classification. These methods are capable of dealing with strong nonlinearity and are essentially based on the kernel defined by the inner product between two feature maps (e.g., Cristianini and

Shawe-Taylor 2000; Rasmussen and Williams 2006). In particular, the feature map $\boldsymbol{\phi}(\cdot)$ is a map from the space $\mathcal{X}$ of the input $\boldsymbol{x}$ to a higher dimensional so-called feature space $\mathcal{F}$, which satisfies $\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) = k(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, where $k(\cdot, \cdot)$ is a symmetric and positive definite kernel. Note that the kernel can be also defined as the inner product on some function space such as probability space.

In each methods, we can obtain a decision boundary given by $\sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$ for an optimal $\alpha_1, \alpha_2, \ldots, \alpha_n$. Roughly speaking, the differences between these two methods are whether it has a sparsity and whether it can be evaluated the variance of the estimator. Indeed, the estimator obtained from the SVM is sparse, i.e., some of the α_i's reduces exactly to 0, but we can not evaluate its variance because we have no evaluation formula for it. On the other hand, because the estimator obtained from the Gaussian process classification is based on the Bayesian method, we can obtain its posterior distribution and thus we can evaluate its variance although it is not sparse.

In the following, we describe these details and examples of commonly used kernel.

2.2.1 SVM

The SVM is a classifier whose decision rule is

$$y = \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}))$$

for unknown input $\boldsymbol{x}$, where $\mathrm{sgn}(\cdot)$ is the sign function that returns 1 or -1 when the argument is respectively positive or negative, and $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ is called a discriminant function. Here, $\boldsymbol{w}$ $(\in \mathcal{F})$ is an unknown coefficient for $\boldsymbol{\phi}(\boldsymbol{x})$.

The optimal coefficient $\hat{\boldsymbol{w}}$ is given by maximizing a so-called margin. For a dataset of n-tuple $\{(y_i, \boldsymbol{x}_i) \mid i = 1, 2, \ldots, n\}$ consisting of an input $\boldsymbol{x}_i$ $(\in \mathcal{D})$ and its output y_i $(\in \{1, -1\})$, the maximization problem reduces to

$$\max_{\boldsymbol{w} \in \mathcal{V}} \min_{i \in \{1,2,\ldots,n\}} \{y_i \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i)\},$$

where

$$\mathcal{V} \equiv \{\boldsymbol{w} \in \mathcal{F} \mid ||\boldsymbol{w}|| = 1;\ \forall i,\ y_i \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) > 0\} \tag{2.1}$$

is called the version space. You may think this optimization problem is hard to solve because $\mathcal{F}$ is a high dimensional space, but it can be shown by the representer theorem (e.g., Shawe-Taylor and Cristianini 2004) that the optimal coefficient of $\boldsymbol{w}$ provides a simple estimated discriminant function as

$$\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{\alpha}_i k(\boldsymbol{x}, \boldsymbol{x}_i), \tag{2.2}$$

where $\hat{\alpha}_i$ is given by the following optimization problem:

$$\begin{aligned} &\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{subject to} \ \sum_{i=1}^{n} \alpha_i y_i = 0 \ \text{ and } \ \forall i,\ \alpha_i \geq 0. \end{aligned} \tag{2.3}$$

Since this optimization problem is convex with respect to the variables to be optimized, we can use a popular method of convex optimization (e.g., Boyd and Vandenberghe 2004).

Note that in general $\boldsymbol{\phi}(\boldsymbol{x})$ is nonlinear with respect to $\boldsymbol{x}$, the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ and the decision boundary $\{\tilde{\boldsymbol{x}} \mid \hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\tilde{\boldsymbol{x}}) = 0\}$ are also nonlinear.

2.2.2 Gaussian process classification

In this section, the output is assumed to be 0 or 1 without loss of generality, and we consider the following probabilistic model:

$$p_i = \mathrm{P}(y_i = 1 | Z(\boldsymbol{x}_i) = z_i) = \frac{\exp(z_i)}{1 + \exp(z_i)}, \tag{2.4}$$

where $\{Z(\boldsymbol{x}_i)\}$ is a Gaussian process with mean function 0 and covariance function $\mathrm{Cov}[Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The Gram matrix is denoted by $K = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1,2,\ldots,n}$. We assume that $y_i | Z(\boldsymbol{x}_i)$ and $y_j | Z(\boldsymbol{x}_j)$ are independent for any $i \neq j$ ($i, j \in \{1, 2, \ldots, n\}$).

The relation between kernel and covariance function may seem irrelevant, but it can be obtained from a simple example, the inference of a linear function $Z(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) =$

$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$ with $\boldsymbol{w} \sim N(\boldsymbol{0}, I)$ and feature map $\boldsymbol{\phi}(\boldsymbol{x})$. In this simple model, we have

$$\mathrm{E}[Z(\boldsymbol{x})] = \mathrm{E}[\boldsymbol{w}]^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) = 0,$$
$$\mathrm{Cov}[Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)] = \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}]\boldsymbol{\phi}(\boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j),$$

and this covariance function is just the kernel.

Our goal in this section is to evaluate the distribution of unknown output y conditional on $\{(\boldsymbol{x}_i, y_i) \mid i = 1, 2, \ldots, n\}$ and unknown input $\boldsymbol{x}$ in a simple form. Let f be a probability density function of random variable. By letting $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, we have

$$\mathrm{P}(y = 1|\boldsymbol{y}) = \int \mathrm{P}(y = 1|Z(\boldsymbol{x}) = z) f(z|\boldsymbol{y}) dz = \int \frac{\exp(z)}{1 + \exp(z)} f(z|\boldsymbol{y}) dz.$$

Unfortunately, we can not evaluate $f(z|\boldsymbol{y})$ analytically, so we need to approximate it. Let us denote $(Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2), \ldots, Z(\boldsymbol{x}_n))^{\mathrm{T}}$ by $\boldsymbol{z}$. We then have

$$f(z|X, \boldsymbol{y}, \boldsymbol{x}) = \int f(z|\boldsymbol{z}) f(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z}.$$

By using the Laplace approximation, we see that $\boldsymbol{z}|\boldsymbol{y}$ is approximately distributed according to $\mathrm{N}(\hat{\boldsymbol{z}}, (\hat{P} + K^{-1})^{-1})$, where $\hat{\boldsymbol{z}} = \mathrm{argmin}_{\boldsymbol{z}} \log f(\boldsymbol{z}|\boldsymbol{y})$ and $\hat{P} = \mathrm{diag}(\hat{p}_i(1 - \hat{p}_i))_{i=1,2,\ldots,n}$ for $\hat{p}_i = \exp(\hat{z}_i)/(1 + \exp(\hat{z}_i))$. Note that $\hat{\boldsymbol{z}}$ can be obtained by the Newton's method. In fact, we can update in the following: let $\boldsymbol{z}^{(0)}$ be the initial value of $\boldsymbol{z}$ and let $\boldsymbol{p}^{(0)} = (\exp(z_i^{(0)})/\{1+\exp(z_i^{(0)})\})_{i=1,2,\ldots,n}$ and $P^{(0)} = \mathrm{diag}(p_i^{(0)}(1-p_i^{(0)}))_{i=1,2,\ldots,n}$. Then, for t-steps, we update as

$$\begin{aligned}
\boldsymbol{z}^{(t+1)} &= K(I + P^{(t)}K)^{-1}(\boldsymbol{y} - \boldsymbol{p}^{(t)} + P^{(t)}\boldsymbol{z}^{(t)}) \\
\boldsymbol{p}^{(t+1)} &= \left(\frac{\exp(z_i^{(t+1)})}{1 + \exp(z_i^{(t+1)})} \right)_{i=1,2,\ldots,n} \\
P^{(t+1)} &= \mathrm{diag}(p_i^{(t+1)}(1 - p_i^{(t+1)}))_{i=1,2,\ldots,n},
\end{aligned} \tag{2.5}$$

and we repeat this until convergence. The derivation of (2.5) is in Section 6.2.1.

Let us denote $\hat{\boldsymbol{p}}$ and $\boldsymbol{k}(\boldsymbol{x})$ by $(\exp(\hat{z}_i)/\{1 + \exp(\hat{z}_i)\})_{i=1,2,\ldots,n}$ and $(k(\boldsymbol{x}, \boldsymbol{x}_i))_{i=1,2,\ldots,n}$, respectively. Then we can evaluate the distribution of $Z(\boldsymbol{x})$ conditional on $\boldsymbol{y}$ as

$$Z(\boldsymbol{x})|\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{p}}), k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}(\hat{P}^{-1} + K)^{-1}\boldsymbol{k}(\boldsymbol{x})) \tag{2.6}$$

since $Z(\boldsymbol{x})|\boldsymbol{z}$ is a Gaussian distributed random variable with mean $\boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}K^{-1}\boldsymbol{z}$ and variance $k(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}K^{-1}\boldsymbol{k}(\boldsymbol{x})$ by a simple calculation.

We use the fact that $\exp(z)/\{1+\exp(z)\}$ is approximated by $\Phi(\lambda z)$, where $\Phi(\cdot)$ is a cumulative distribution function of the standard Gaussian distribution and $\lambda = \sqrt{\pi/8}$ (see Bishop 2006 for more details). Finally, by letting $\mathrm{E}[Z(\boldsymbol{x})|\boldsymbol{y}] = \nu$ and $\mathrm{V}[Z(\boldsymbol{x})|\boldsymbol{y}] = \tau^2$, we can approximate $\mathrm{P}(y=1|\boldsymbol{y})$ as follows:

$$\begin{aligned}\mathrm{P}(y=1|\boldsymbol{y}) &= \int \frac{\exp(z)}{1+\exp(z)} f(z|\boldsymbol{y})dz \simeq \int \Phi(\lambda z) f(z|\boldsymbol{y})dz \\ &= \Phi\left(\frac{\lambda\nu}{(1+\lambda^2\tau^2)^{1/2}}\right) \simeq \frac{\exp(g(\boldsymbol{x}))}{1+\exp(g(\boldsymbol{x}))},\end{aligned}$$

where

$$g(\boldsymbol{x}) = \left\{1 + \frac{\pi}{8}(\sigma^2 - \boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}(\hat{P}^{-1}+K)^{-1}\boldsymbol{k}(\boldsymbol{x}))\right\}^{-1/2} \cdot \boldsymbol{k}(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\hat{\boldsymbol{p}})$$

Note that the decision boundary $\{\tilde{\boldsymbol{x}};\ g(\tilde{\boldsymbol{x}}) = 0\}$ is also nonlinear since $\boldsymbol{k}(\boldsymbol{x})$ is nonlinear with respect to $\boldsymbol{x}$.

2.2.3 Examples of kernel

The decision boundary obtained from the SVM or the Gaussian process classification is, in general, depend on the kernel, so we need to choose it. Generally speaking, the kernel represent the similarity between p-dimensional inputs $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, but an arbitrary function of $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ will not be a valid kernel because the kernel must be symmetric and positive definite. Here, we describe some examples of several commonly used kernels.

polynomial kernel

The polynomial kernel may be considered as the most simple kernel with the form of

$$k(\boldsymbol{x},\tilde{\boldsymbol{x}}) = (c + \boldsymbol{x}^{\mathrm{T}}\tilde{\boldsymbol{x}})^d,$$

where $c \geq 0$ and $d \in \mathbb{N}$. Especially, it is known as a so-called linear kernel and a quadratic kernel when $d = 1$ and $d = 2$, respectively. The linear kernel appears in the inference of

a linear function as in the above mentioned. Note that this kernel is a finite dimensional kernel. For example, when $d = 2$, we have

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sum_{j=1}^{p} x_j^2 \tilde{x}_j^2 + 2 \sum_{1 \le i < j \le p} x_i x_j \tilde{x}_i \tilde{x}_j + 2c \sum_{j=1}^{p} x_i \tilde{x}_j + c^2 = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\phi}(\tilde{\boldsymbol{x}}).$$

From this, the feature map $\boldsymbol{\phi}(\boldsymbol{x})$ is given by

$$\boldsymbol{\phi}(\boldsymbol{x}) = ((x_j^2)_{1 \le i \le p}^{\mathrm{T}}, (\sqrt{2} x_i x_j)_{1 \le i < j \le p}^{\mathrm{T}}, (\sqrt{2c} x_i)_{1 \le i \le p}^{\mathrm{T}}, c)^{\mathrm{T}} \in \mathbb{R}^{(p+1)(p+2)/2}.$$

Gaussian kernel

While there are so many kinds of kernels , we can say that one of the most frequently used kernels is Gaussian kernel with the form of

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp\left(-\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{2\ell^2} \right),$$

where ℓ (> 0) is a tuning parameter which controls the dispersion of the Gaussian kernel. The Gaussian kernel can be obtained from the limit of a specific random function: let us consider, for simplicity, an 1-dimensional random function as

$$\psi_m(x) = \frac{1}{m} \sum_{j=1}^{m} w_j \exp\left(- \left(x - \frac{j}{m} \right)^2 \right),$$

where w_j is an i.i.d. Gaussian random variable with mean 0 and variance 1. This can be interpreted as the Bayesian linear regression model with prior $(w_j)_{j=1,2,\ldots,m}$ and the feature map $(\exp(-(x - j/m)^2))_{j=1,2,\ldots,m}$. Then we have

$$\psi(x) = \lim_{m \to \infty} \psi_m(x) = \int w(t) \exp(-(x - t)^2) dt,$$

where $w(t)$ is a Gaussian distributed random variable with mean 0 and variance 1 for any t, and is independent even for any infinitesimal interval $[t, t + dt)$. Note that this type of random process is studied in the area of white noise analysis (see e.g., Chapter 8 in Hida and Si 2008). Moreover, the expectation and the covariance of $\psi(\cdot)$ is given by

$$\mathrm{E}[\psi(x)] = 0 \quad \text{and} \quad \mathrm{Cov}[\psi(x)\psi(\tilde{x})] \propto \exp\left(-\frac{1}{2}(x - \tilde{x})^2 \right),$$

respectively.

rational quadratic kernel

The rational quadratic kernel is defined by

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left(1 + \frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{2\alpha\ell^2}\right)^{-\alpha},$$

where $\alpha,\ \ell > 0$. It is easy to see that the rational kernel converges to the Gaussian kernel when $\alpha \to \infty$ and ℓ is fixed. Another interpretation of the rational quadratic kernel is that it can be seen as a scale mixture, i.e. an infinite sum, of Gaussian kernels with deferent dispersions: let $\gamma = \ell^{-2}$ and $f(\gamma|\alpha, \beta) \propto \gamma^{\alpha-1}\exp(-\alpha\gamma/\beta)$ be a density function of a gamma distribution with parameter $\alpha, \beta > 0$. Then we have

$$\int f(\gamma|\alpha, \beta)\exp\left(-\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{2\ell^2}\right)d\gamma \propto \left(1 + \frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2}{2\alpha\ell^2}\right)^{-\alpha},$$

where we set $\beta = \ell^2$.

γ-exponential kernel

The γ-exponential kernel

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp\left(-\left(\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\ell}\right)^{\gamma}\right) \qquad (\ell > 0,\ 0 < \gamma \leq 2)$$

includes the Gaussian kernel in the special case when $\gamma = 2$. When $\gamma = 1$ and $p = 1$ (the dimension of input), it is well known as the exponential kernel which is the covariance function of the Ornstein-Uhlenbeck process.

Matérn kernel

The Matérn kernel is defined by

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{2^{1-\alpha}}{\Gamma(\alpha)}\left(\frac{\sqrt{2\alpha}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\ell}\right)^{\alpha} K_\alpha\left(\frac{\sqrt{2\alpha}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\ell}\right),$$

where $\alpha,\ \ell > 0,\ \Gamma(\cdot)$ is a gamma function and K_α is a modified Bessel function of the second kind of order α. It is known that the Matérn kernel is $\lfloor\alpha\rfloor$-times differentiable, where $\lfloor\alpha\rfloor$ is the largest integer not greater than α , and it includes the Gaussian kernel

(when $\alpha \to \infty$) and the exponential kernel (when $\alpha = 1/2$). When $\alpha = d + 1/2$ with $d \in \mathbb{N}$, the Matérn kernel is simply expressed as

$$k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp\left(-\frac{\sqrt{2d+1}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\ell}\right) \frac{\Gamma(d+1)}{\Gamma(2d+1)} \sum_{i=0}^{d} \frac{(d+i)!}{i!(d-i)!} \left(\frac{\sqrt{4(2d+1)}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|}{\ell}\right)^{p-i}$$

by the definition of K_α.

It is known that the conical combination of kernels or the product of kernels are also satisfy the property of kernel, so we can construct new kernel from several kernels. Especially, the sum of kernels is called multiple kernel and is studied in machine learning area (e.g., Bach et al. 2004; Kloft et al. 2011). Although there are many candidates of kernels, in practice, we need to choose the optimal value of a tuning parameter which is included in the kernel. Usually, it is selected by the computational method such as cross validation (CV: Stone 1974). The discriminant function obtained from the SVM and the Gaussian process classification is illustrated in Figure 2.1. The data can be available on the package `mlbench` in `R`. By using a proper kernel, it is seen that the classifier obtained from the model can deal with a non-linear discriminant function. In this chapter, as usual, we use the Gaussian kernel and select the optimal value of the tuning parameter by cross-validation.

2.3 Active learning

The active learning method (optimal design method) is to design inputs to improve a learning (estimation) accuracy for the case where we can design the inputs purposefully. Before introducing the following active learning methods, we first describe classic optimal design methods, the A-optimal and D-optimal designs.

Generally speaking, for a parameter vector $\boldsymbol{\theta}$ and its estimator $\hat{\boldsymbol{\theta}}$, the mean squared error matrix $\mathrm{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}}]$ is a natural index for the estimation accuracy. It is divided into the variance-related term $\mathrm{E}[(\hat{\boldsymbol{\theta}} - \mathrm{E}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - \mathrm{E}[\hat{\boldsymbol{\theta}}])^{\mathrm{T}}]$ and the bias-related term

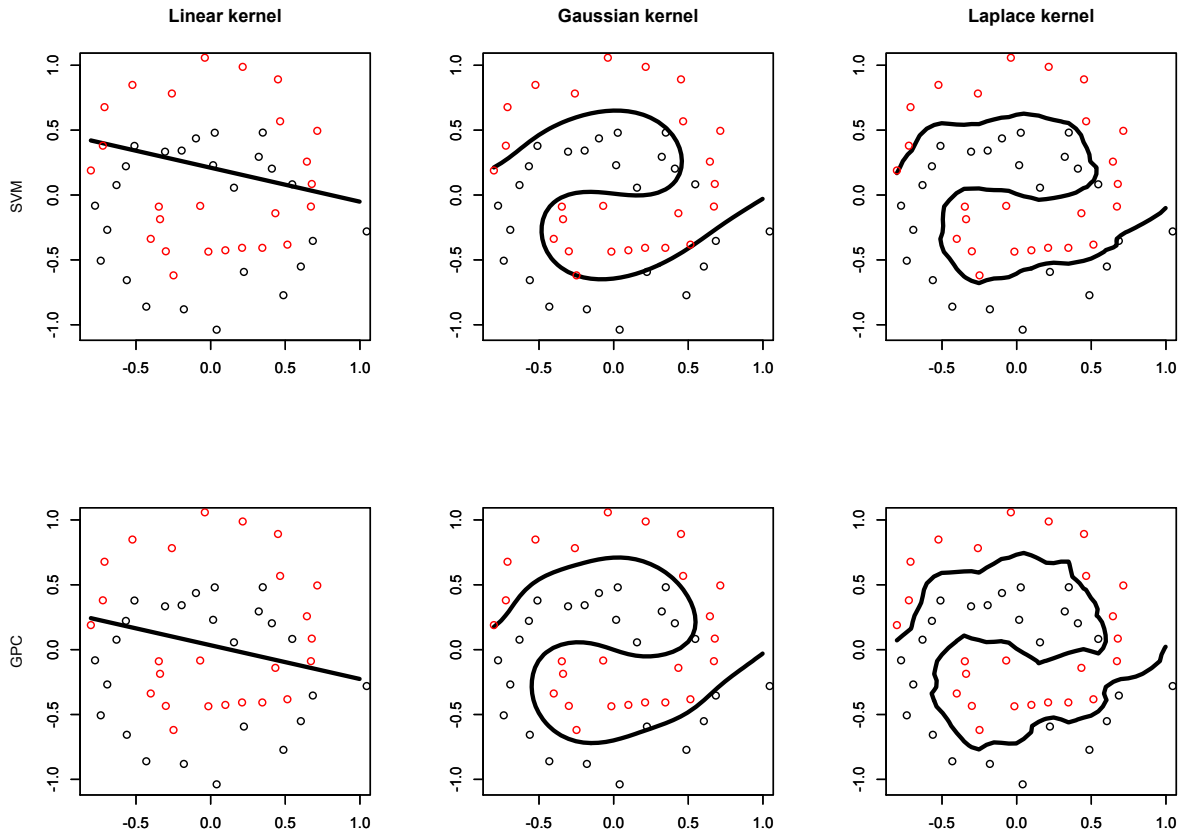

Figure 2.1: **Example of discriminant functions.** The red circles and the black circles represente the label of each class, respectively. In the top, the discriminant function given by the SVM is drawn, and in the bottom, those given by the Gaussian process classification (GPC) is drawn. The solid line is the decision boundary obtained from each methods. The kernel used in these models is, from the left to the right, the linear kernel, the Gaussian kernel, and the Laplace kernel, respectively. The tuning parameter included in each kernels is selected by the cross validation.

$(\mathrm{E}[\hat{\boldsymbol{\theta}}]-\boldsymbol{\theta})(\mathrm{E}[\hat{\boldsymbol{\theta}}]-\boldsymbol{\theta})^{\mathrm{T}}$, and in well-used estimation methods such as the maximum likelihood method, the former becomes the main term asymptotically. In classic optimal design methods, it is proposed to give a new input which minimize the trace or maximize the determinant of the main term $\mathrm{E}[(\hat{\boldsymbol{\theta}}-\mathrm{E}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}}-\mathrm{E}[\hat{\boldsymbol{\theta}}])^{\mathrm{T}}]$, and it is called A-optimal design or D-optimal design, respectively (e.g., Kiefer 1959, 1961; Kiefer and Wolfowitz 1959). Because we cannot evaluate this matrix explicitly in general, it is common to use the inverse of Fisher's information matrix in place of it, which is asymptotically equivalent to it under some regularity conditions. Note that the A-optimal and D-optimal designs are

equivalent under a setting of linear regression. Also note that in the D-optimal design, the input giving the maximum prediction variance of its output is selected under some conditions, which is an important property. That is, letting $\hat{y}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ be the predictive value of the output for $\boldsymbol{x}$, the D-optimal design selects $\mathrm{argmax}_{\boldsymbol{x}} \mathrm{V}[\hat{y}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})]$ as a new input, and so it is regarded as a method which gradually reduces region which gives unstable prediction.

2.3.1 Entropy based optimal design

In the classic optimal design methods, we can only apply it to parametric models because these optimality is based on the variance of the estimator of the parameter. Moreover, there is a problem so-called model misspecification. It is easy to imagine that we cannot improve a estimation accuracy by using a linear model when the true model is nonlinear. These are critical problem when we do not have a proper parametric model what we use and when the data includes the strong non-linearity.

To overcome such a problem, by considering the important property of the D-optimal design explained above, Umezu and Ninomiya (2013) proposed a new optimal design method which selects samples with the maximum prediction instability measured by an entropy by using the model (2.4). They considered the optimality defined by

$$\boldsymbol{x}_{n+1} = \underset{\boldsymbol{x} \in \mathcal{X}}{\mathrm{argmax}} \, \mathrm{E}[-\log f(y, Z(\boldsymbol{x}) | \boldsymbol{y})], \tag{2.7}$$

and they call it H-optimal design. The H-optimality is interpreted as that the new input is sampled from the most unstable region because the entropy measures the instability of an event.

Here, we derive a lower bound of the objective function in (2.7) in a simple form because it cannot be in general evaluated explicitly, and then we consider maximizing the lower bound instead of the objective function in (2.7). The objective function in (2.7) can be rewritten as

$$\mathrm{E}[-\log f(y, Z | \boldsymbol{y})] = \mathrm{E}[-\log f(y | \boldsymbol{y}, Z)] + \mathrm{E}[-\log f(Z | \boldsymbol{y})]$$

Let us define ν and τ^2 as in Section 2.2.2, and let us denote $Z(\boldsymbol{x})$ by Z for simplicity of the notation. Note that ν and τ^2 are depend on an input $\boldsymbol{x}$. The predictive distribution of Z is given by (2.6), and thus the second term in the above equality is given by

$$\mathrm{E}[-\log f(Z(\boldsymbol{x})|\boldsymbol{y})] = \frac{1}{2}\log(2\pi\tau^2) + \frac{1}{2}. \tag{2.8}$$

The first term in the above equality is reduced to

$$\begin{aligned}
\mathrm{E}[-\log f(y|\boldsymbol{y}, Z)] &= \mathrm{E}[yZ - \log(1+\exp(Z))|\boldsymbol{y}] \\
&= -\mathrm{E}\left[\frac{Z\exp(Z)}{1+\exp(Z)}\middle|\boldsymbol{y}\right] + \mathrm{E}[\log(1+\exp(Z))|\boldsymbol{y}] \\
&= -\frac{1}{\tau^2}\mathrm{E}[Z(Z-\nu)\log(1+\exp(Z))|\boldsymbol{y}] + 2\mathrm{E}[\log(1+\exp(Z))|\boldsymbol{y}].
\end{aligned}$$

In the third equality, we just use

$$\frac{z\exp(z)}{1+\exp(z)} = z\frac{d\log(1+\exp(z))}{dz} \quad \text{and} \quad \frac{df(z|\boldsymbol{y})}{dz} = -\frac{z-\nu}{\tau^2}f(z|\boldsymbol{y}).$$

Moreover, we can evaluate the expectations in the last line as

$$\mathrm{E}[Z(Z-\mu)\log(1+e^Z)] \leq \tau^2\log 2 + \frac{2\tau(\mu^2+\tau^2)}{\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\tau^2}\right) + 2\tau^2\max\{\mu, 0\} \tag{2.9}$$

and

$$\mathrm{E}[\log(1+e^Z)|\boldsymbol{y}] \geq \log 2 + \frac{1}{2}\nu, \tag{2.10}$$

respectively. From (2.8), (2.9) and (2.10), because $\log 2 + 1/2$ is a positive constant, we obtain a lower bound of the joint entropy as

$$\mathrm{E}[-\log f(y, Z(\boldsymbol{x})|\boldsymbol{y})] \geq -|\nu| - \frac{2(\nu^2+\tau^2)}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{\nu^2}{2\tau^2}\right) + \frac{1}{2}\log(2\pi\tau^2). \tag{2.11}$$

The derivation of (2.9) and (2.10) are in Section 6.2.1.

Let $H(\boldsymbol{x})$ be the right-hand side in (2.11), it is easy to see that when $|\nu| \to \infty$ and τ^2 is fixed or $\tau^2 \to 0$ and ν is fixed, then $H(\boldsymbol{x}) \to -\infty$. Thus, we can sample not only around the discriminant function but also the region which is unstable, and this is a kind of desirable property. Although the H-optimality is derived from the same idea of the classic optimal design, it has a considerably computational task to maximize $H(\boldsymbol{x})$ when the dimension of x are large.

2.3.2 Active learning via SVM

Unfortunately, the SVM cannot use the classic optimal design methods because we have no evaluation formula for the variances of parameter estimators in the SVM setting as mentioned above. Actually as seen from (2.3), the numbers of parameters and samples are the same, and so we have no evaluation formula even in an asymptotical form. Under this situation, Tong and Koller (2002) propose a new criterion for sampling based on the version space for the SVM.

After getting a dataset of n-tuple $\{(y_i, \boldsymbol{x}_i) \mid i = 1, 2, \ldots, n\}$, the version space is defined as in (2.1). In this definition, each $y_i \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) > 0$ represents a half space in $\mathcal{F}$ and $\mathcal{V}$ represents the polyhedral body which is the product set of the half spaces. As the $(n+1)$-th new sample, they propose to select $\boldsymbol{x}_{n+1}$ $(\in \mathcal{X})$ such that the hyperplane $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}_{n+1}) = 0$ divides $\mathcal{V}$ into two parts as equally as possible. It is indicated in Tong and Koller (2002) that the $(n+1)$-th new sample is close to

$$\underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmin}} |\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})|$$

for the estimated discriminant function in (2.2). Therefore, this method selects $\boldsymbol{x}_{n+1}$ such that $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}_{n+1})$ is close to 0, in other words, $\boldsymbol{x}_{n+1}$ which is close to the decision boundary.

2.4 Proposed method

In this section, first we will point out a severe problem in the ASVM under our situation. Next we will propose a simple active learning method which avoids the problem.

For simplicity, we assume that the dimension p of inputs is 2, and let us consider the example in which all candidates of the inputs and their outputs are as in the left of Figure 2.2. We consider the situation where the region of $\boldsymbol{x}$ in which $y = -1$ is given more likely than $y = 1$ tends to be separated and near the edge but they will not be far from each other. As written in Section 3.1, this situation is natural for the defect rate evaluation problem in product manufacturing. Actually, our real data treated later have

this situation while p is much larger.

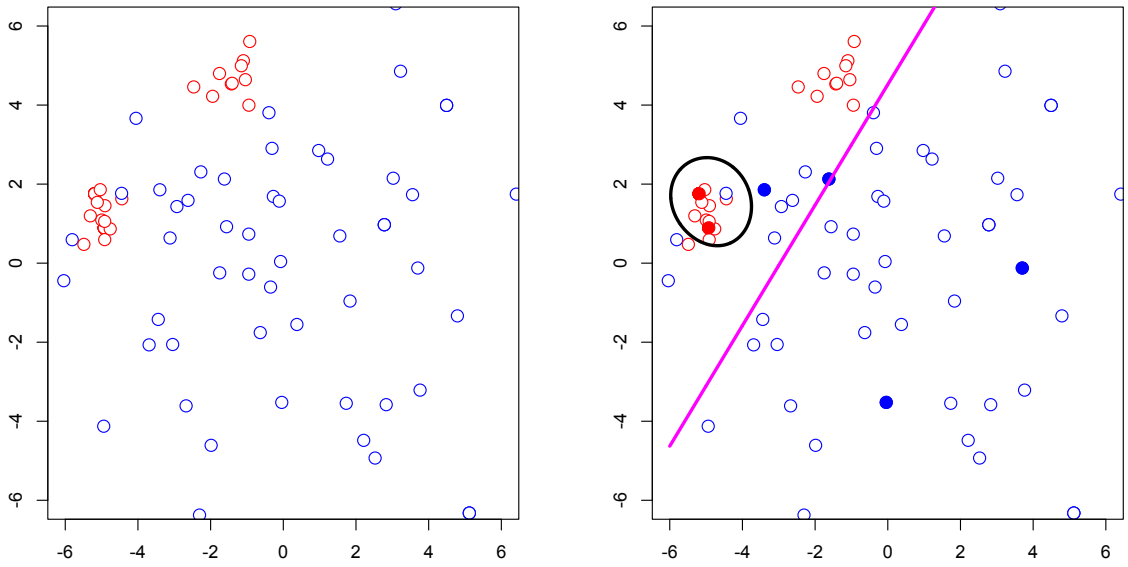

Figure 2.2: **Example of two-dimensional case.** In the left, the blue and red open circles represent the candidates of inputs with $y = 1$ and $y = -1$, respectively. In the right, the filled circles and the curve represent the sampled inputs, the black solid line is the estimated decision boundary via SVM, and the purple solid line is the linear discriminant boundary.

Let us imagine that the ASVM is applied to this example. Although the ASVM is an active learning method, we cannot get samples actively in the first stage, and so we get them completely at random. Suppose that they are filled points in the right of Figure 2.2. If we estimate the decision boundary based on them by the SVM, the curve in the figure is obtained. After that, we get samples according to the sampling scheme of the ASVM. Then decision boundary for the left group of $y = -1$ will be improved step by step because inputs close to the estimated decision boundary must be selected as explained in Section 2.3. On the other hand, the inputs with $y = -1$ in the right group are rarely sampled, and so the estimated decision boundary near the right group will not appear for a long time. Thus, on the whole, the discriminant accuracy will not be much improved even if the sampling is repeated.

When p is large and the number of outputs of $y = -1$ is small, the above phenomenon becomes apparent. The region giving $y = -1$ more likely than not tends to be more separated, and it is difficult to get any sample from a number of separated regions at the first random sampling. In addition, it takes longer time to get a sample from the separated regions where no inputs are sampled at the first stage. Thus, it is indicated that active learning methods which get samples near the estimated decision boundary are not suitable for this type of cases.

On the other hand, the Gaussian process regression method, which has comparable performance to the SVM, is based on a framework of classic statistics, and so the prediction accuracy can be evaluated. Considering the important property of the D-optimal design explained in Section 2.3, Umezu and Ninomiya (2013) proposed a new optimal design method which selects samples with the maximum prediction instability measured by an entropy. Using this method, we can select inputs by considering its closeness to not only the estimated decision boundary but also already sampled inputs. However, this can be regarded as a method between the ASVM and the method with sampling completely at random, and so it must be inappropriate for our problem.

Hence, we consider a method which does not depend on the estimated decision boundary. After sampling an input uniformly at random from $\mathcal{X}$ and obtaining an output according to the input, which is repeated till at least one output with $y = -1$ is obtained, we consciously forget the nonlinearity of our discriminant problem and conduct a linear discriminant analysis. Let $\tilde{h}(\boldsymbol{x})$ be the linear discriminant function, and let $\mathcal{D}^- = \{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) < 0\}$. In this chapter, we consider a hyperplane which consists of the points such that the distances from the centers of $\mathcal{D}$ and of inputs with $y = -1$ are equal, and we define $\tilde{h}(\boldsymbol{x})$ so that $\{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) = 0\}$ is the hyperplane. Because the separated regions giving $y = -1$ more likely than not are not large, not far from each other, and at the edge, it can be expected that most of such regions are included in $\mathcal{D}^-$ (see the right in Figure 1). Then we sample inputs uniformly at random from $\tilde{\mathcal{X}} \equiv \mathcal{D}^- \cap \mathcal{X}$ and

obtain outputs according to the inputs. By repeating this procedure, we can expect to get samples from all the separated regions. In this situation, since the area of $\mathcal{D}^-$ is not large in comparison with the area of $\mathcal{D}$, we will be able to get inputs with $y = -1$ efficiently. Finally we recall the nonlinearity of our discriminant problem, and then estimate the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ by applying the SVM. This procedure can be summarized as in Table 2.1.

Table 2.1: Procedure in LSVM.

1)	For $i = 1, 2, \ldots, N$, sample $\boldsymbol{x}_i$ uniformly at random from $\mathcal{X}$, and obtain $y_i \in \{\pm 1\}$ according to $\boldsymbol{x}_i$.
2)	Obtain a linear discriminant function $\tilde{h}(\boldsymbol{x})$ such that $\{\boldsymbol{x} \mid \tilde{h}(\boldsymbol{x}) = \boldsymbol{0}\}$ becomes the hyperplane which is equidistant from the centers of $\mathcal{D}$ and of inputs with $y = -1$.
3)	Set $\tilde{\mathcal{X}} = \{\boldsymbol{x} \in \mathcal{X} \mid \tilde{h}(\boldsymbol{x}) < 0\}$
4)	For $i = 1, 2, \ldots, M$, sample $\tilde{\boldsymbol{x}}_i$ uniformly at random from $\tilde{\mathcal{X}}$, and obtain $\tilde{y}_i \in \{\pm 1\}$ according to $\tilde{\boldsymbol{x}}_i$
5)	Iterate 2) to 4) for K times
6)	Estimate a discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ by SVM

2.5 Simulation study

To compare "Linear discrimination"-based active learning with the SVM (LSVM) proposed in Section 2.4, the method in Tong and Koller (2002) (ASVM), and "Random sampling" with the SVM (RSVM), we conduct a simulation study in this section. In the RSVM, we sample inputs from $\mathcal{X}$ completely at random without active learning and finally use the SVM to estimate the discriminant function. Because we must apply these methods many times in the simulation study, we set the dimension of inputs and the number of sampled inputs are small.

Concretely speaking, first we produce $2,000$ inputs with a negative output by

$$\boldsymbol{x} \sim \mathrm{Mix}(1/2, \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \;\Rightarrow\; y = -1$$

and $98,000$ inputs with a positive output by

$$\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3) \;\Rightarrow\; y = 1,$$

and pool them. Here, $\mathrm{Mix}(1/2, \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2))$ means the mixture distribution of $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with the mixing rate $1:1$. Letting $\boldsymbol{R}(\theta)$ be the two dimensional rotation matrix with the angle θ, we set

$$\begin{aligned}
&\boldsymbol{\mu}_1 = (0, 5)^{\mathrm{T}}, \quad \boldsymbol{\mu}_2 = \boldsymbol{R}(\theta)\boldsymbol{\mu}_1, \quad \boldsymbol{\mu}_3 = \boldsymbol{0}, \\
&\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_1^2, \sigma_2^2), \quad \boldsymbol{\Sigma}_2 = \boldsymbol{R}(\theta)\boldsymbol{\Sigma}_1\boldsymbol{R}(\theta)^{\mathrm{T}}, \quad \boldsymbol{\Sigma}_3 = 5I_2
\end{aligned}$$

as the values of the parameters. The inputs with $y = -1$ form two groups and the angle θ indicates their distance. Next we compare the methods by getting samples from the pooled data. In every method, first we get 50 samples completely at random, and then we iterate 25 times of samplings in which we get 10 samples at one time according to the procedure of each method. That is, in Table 2.1 for the LSVM, we set $N = 50$, $M = 10$ and $K = 25$.

In Table 2.2, by each designed value of $(\theta, \sigma_1^2, \sigma_2^2)$, we can check the transitions of FPRs (false positive rates) caused by increasing the number of iterations of sampling, where the FPR is defined by $\#\{i \mid \hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) > 0,\ y_i = -1\}/\#\{i \mid y_i = -1\}$. Here we do not report about the FNRs (false negative rates) because the FPR is more important to be checked than the FNR in our problem and because the FNRs for all methods were always very close to one and almost the same values. The values in the table are the averages and standard deviations of FPRs based on 50 simulations for each method. In every method, the FPR is decreasing when the number of iterations is increased.

First, it can be seen in every case that the RSVM provides much higher values of FPRs than those of the LSVM and the ASVM. This is because the RSVM can rarely get samples with $y = -1$ unlike the other two methods. Next, it can be seen by comparing the two methods that basically the LSVM is superior to the ASVM when the number of

Table 2.2: Transition of FPRs for simulated data.

θ	method	number of iterations				
		5	10	15	20	25
$\pi/9$	LSVM (sd)	0.912 (0.188)	0.022 (0.020)	0.016 (0.007)	0.021 (0.011)	0.031 (0.014)
	ASVM (sd)	0.102 (0.132)	0.027 (0.017)	0.036 (0.019)	0.037 (0.016)	0.026 (0.009)
	RSVM (sd)	1.000 (0.000)	1.000 (0.001)	0.932 (0.169)	0.654 (0.332)	0.352 (0.299)
$2\pi/9$	LSVM (sd)	0.994 (0.024)	0.360 (0.278)	0.068 (0.096)	0.043 (0.026)	0.045 (0.026)
	ASVM (sd)	0.540 (0.080)	0.503 (0.002)	0.503 (0.003)	0.504 (0.006)	0.504 (0.002)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.988 (0.061)	0.943 (0.127)	0.805 (0.227)
$\pi/3$	LSVM (sd)	0.979 (0.084)	0.439 (0.244)	0.085 (0.154)	0.033 (0.098)	0.027 (0.098)
	ASVM (sd)	0.532 (0.063)	0.502 (0.002)	0.503 (0.004)	0.503 (0.002)	0.505 (0.004)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.998 (0.013)	0.969 (0.123)	0.833 (0.255)
$4\pi/9$	LSVM (sd)	0.996 (0.028)	0.517 (0.177)	0.337 (0.227)	0.281 (0.250)	0.269 (0.247)
	ASVM (sd)	0.528 (0.037)	0.503 (0.003)	0.503 (0.003)	0.504 (0.003)	0.504 (0.003)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.940 (0.178)	0.841 (0.239)

(a) Case of $(\sigma_1^2, \sigma_2^2) = (0.1, 0.3)$.

iterations become large while the ASVM is superior to the LSVM when it is small. This is because the ASVM can quickly get samples with $y = -1$ close to the initially gotten sample with $y = -1$ but cannot get those far from it. For the case where the two groups of the inputs with $y = -1$ are close, e.g., $\theta = \pi/9$, the ASVM has a possibility of finding any of those, and so the two methods are comparable. In addition, for the case where the two groups are too far from each other, e.g., $\theta = 4\pi/9$, even the LSVM does not have a possibility of finding any of inputs with $y = -1$, and so the superiority of the LSVM becomes small. For the other cases, the LSVM is clearly better than the ASVM.

2.6 Real data analysis

In this section, we compare the methods through applying them to some trial data which is used in a real product manufacturing. The data consists of $97,740$ samples with $y = 1$

θ	method	number of iterations				
		5	10	15	20	25
$\pi/9$	LSVM (sd)	0.787 (0.349)	0.032 (0.024)	0.027 (0.018)	0.028 (0.015)	0.029 (0.017)
	ASVM (sd)	0.083 (0.107)	0.028 (0.016)	0.031 (0.020)	0.026 (0.016)	0.018 (0.009)
	RSVM (sd)	0.900 (0.303)	0.900 (0.303)	0.850 (0.322)	0.663 (0.362)	0.286 (0.257)
$2\pi/9$	LSVM (sd)	1.000 (0.003)	0.359 (0.280)	0.080 (0.109)	0.047 (0.038)	0.050 (0.041)
	ASVM (sd)	0.528 (0.043)	0.503 (0.002)	0.503 (0.004)	0.504 (0.003)	0.505 (0.003)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.994 (0.037)	0.941 (0.129)	0.798 (0.262)
$\pi/3$	LSVM (sd)	1.000 (0.000)	0.506 (0.291)	0.080 (0.162)	0.028 (0.098)	0.019 (0.070)
	ASVM (sd)	0.524 (0.037)	0.502 (0.001)	0.502 (0.002)	0.503 (0.003)	0.503 (0.003)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.992 (0.049)	0.946 (0.133)	0.791 (0.261)
$4\pi/9$	LSVM (sd)	0.990 (0.052)	0.512 (0.180)	0.239 (0.228)	0.192 (0.244)	0.182 (0.241)
	ASVM (sd)	0.535 (0.053)	0.503 (0.003)	0.502 (0.002)	0.503 (0.002)	0.504 (0.002)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.978 (0.072)	0.821 (0.232)

(b) Case of $(\sigma_1^2, \sigma_2^2) = (0.2, 0.2)$.

and $2,260$ samples with $y = -1$, and the dimension p of the inputs is 18. As in the situation we treated until now, the inputs with $y = -1$ form several groups at the edge of the domain $\mathcal{D}$. Note that we know all values of the outputs because this data is for trial. Using these known values, we can estimate a good discriminant function without active learning, but here we suppose to know only the values of outputs gotten by sampling. Needless to say, it is because we look ahead to apply the methods to non-trial data.

In every method, first we get 500 samples completely at random, and then we iterate 200 times of samplings in which we get 50 samples at one time according to the procedure of each method. That is, in Table 2.1 for the LSVM, we set $N = 500$, $M = 50$ and $K = 200$. In Figure 2.3, we plot the transition of the FPR for each method, which is measured by making test data from non-sampled data with $y = -1$. It can be seen that the values of the FPR for the LSVM are always smaller than those of the ASVM and become stable after about 50 times iterations while those of the ASVM are decreasing

θ	method	number of iterations				
		5	10	15	20	25
$\pi/9$	LSVM (sd)	0.901 (0.221)	0.046 (0.034)	0.033 (0.015)	0.030 (0.014)	0.024 (0.011)
	ASVM (sd)	0.061 (0.060)	0.029 (0.018)	0.022 (0.011)	0.016 (0.008)	0.011 (0.005)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.941 (0.159)	0.649 (0.338)	0.401 (0.347)
$2\pi/9$	LSVM (sd)	0.970 (0.096)	0.333 (0.236)	0.055 (0.061)	0.042 (0.017)	0.065 (0.029)
	ASVM (sd)	0.509 (0.046)	0.476 (0.111)	0.475 (0.110)	0.481 (0.091)	0.482 (0.092)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.995 (0.032)	0.943 (0.138)	0.834 (0.255)
$\pi/3$	LSVM (sd)	0.991 (0.061)	0.471 (0.245)	0.082 (0.137)	0.030 (0.068)	0.034 (0.068)
	ASVM (sd)	0.513 (0.017)	0.503 (0.002)	0.502 (0.001)	0.504 (0.003)	0.503 (0.004)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	0.944 (0.157)	0.827 (0.253)
$4\pi/9$	LSVM (sd)	0.991 (0.042)	0.493 (0.200)	0.241 (0.237)	0.179 (0.235)	0.159 (0.227)
	ASVM (sd)	0.531 (0.075)	0.504 (0.002)	0.503 (0.002)	0.503 (0.002)	0.503 (0.002)
	RSVM (sd)	1.000 (0.000)	1.000 (0.000)	0.991 (0.061)	0.971 (0.101)	0.852 (0.220)

(c) Case of $(\sigma_1^2, \sigma_2^2) = (0.3, 0.1)$.

slowly. About the RSVM, the values of the FPR become temporally smaller than those of the LSVM, but it will be by accident because the values considerably fluctuate after that. Moreover, the RSVM is superior to the ASVM in this case. It may be because that there are too many groups of the inputs with $y = -1$ to deal with by the ASVM. Actually the values do not become stable even after 100 times iterations.

2.7 Evaluation of defect rate

While efficient estimation of the discriminant function for a defect was discussed until now, it is often the case actually in product manufacturing that the estimation of its defect rate has more concern. Then we consider to evaluate

$$\mathrm{E}[y = -1] = \int \rho(\boldsymbol{x}) f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{2.12}$$

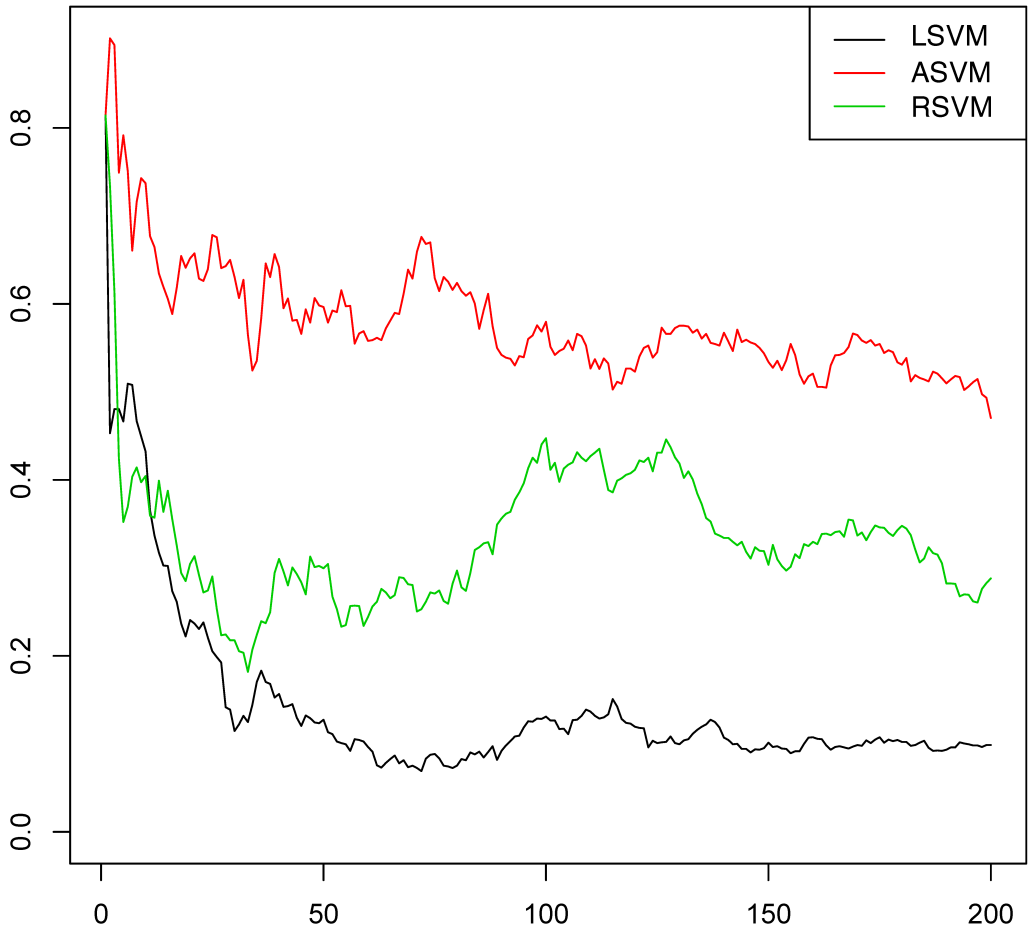

Figure 2.3: **Transition of FPRs for real data.** The horizontal axis indicates the number of sampling iterations and the vertical axis indicates the FPR.

where $f(\boldsymbol{x})$ is the probability density function of $\boldsymbol{x}$, and $\rho(\boldsymbol{x})$ is the probability of being $y = -1$ at $\boldsymbol{x}$.

First we model the local defect rate by

$$\rho(\boldsymbol{x}) = \frac{\exp(a\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b)}{1 + \exp(a\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x}) + b)} \tag{2.13}$$

using the discriminant function $\hat{\boldsymbol{w}} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ obtained by the SVM (e.g., Platt et al. (1999)). Here a and b are unknown parameters, and we estimate them by the maximum likelihood method under the setting where y_i is an independent sample from the Bernoulli distribution $\mathrm{Be}(\rho(\boldsymbol{x}_i))$. We substitute the maximum likelihood estimators $\hat{a}$ and $\hat{b}$ for the a and b in the right-hand side of (2.13), and we denote the substituted right-hand side by $\hat{\rho}(\boldsymbol{x})$ as an evaluated local defect rate.

From this, we can provide the value of the defect rate by evaluating the multiple integration in (2.12) numerically, but it is almost impossible if the dimension of $\boldsymbol{x}$ is large. Then, by simulating $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the distribution $f(\boldsymbol{x})$ at random, we consider to use Monte Carlo integration, that is, to provide $\sum_{i=1}^{\tilde{n}} \hat{\rho}(\tilde{\boldsymbol{x}}_i)/\tilde{n}$. However, a problem remains. The defect rate is tiny in general, i.e., $\hat{\rho}(\boldsymbol{x}) \approx 0$ for almost all $\boldsymbol{x}$, and so we cannot provide an accurate evaluation of the defect rate even if we simulate huge size of $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$.z

To overcome this difficulty, we try to simulate $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the region where $\hat{\rho}(\boldsymbol{x})$ is large, and then we evaluate the defect rate efficiently by an importance sampling. Concretely speaking, letting $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ be respectively the sample mean vector and sample variance-covariance matrix for a set of the inputs with a defect $\{\boldsymbol{x}_i \mid y_i = -1\}$, we simulate $\{\tilde{\boldsymbol{x}}_i \mid i = 1, 2, \ldots, \tilde{n}\}$ from the Gaussian distribution $\mathrm{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ at random. Then we evaluate the defect rate by

$$\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{f(\tilde{\boldsymbol{x}}_i)}{g(\tilde{\boldsymbol{x}}_i)} \hat{\rho}(\tilde{\boldsymbol{x}}_i), \tag{2.14}$$

where $g(\boldsymbol{x})$ is the probability density function of $\mathrm{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$. From the law of large numbers, this converges to our desired expectation in (2.12).

For the data treated in Section 3.5, we conducted this defect rate evaluation after 100 iterations of samplings. The estimates of a and b were -3.88 and 0.58, respectively. In

Figure 2.4, we can check the transition of the evaluations in (2.14) caused by increasing $\tilde{n}$. The evaluations become stable when $\tilde{n}$ is close to 10^6, and as a result we found that the defect rate is about 1.2×10^{-6}.

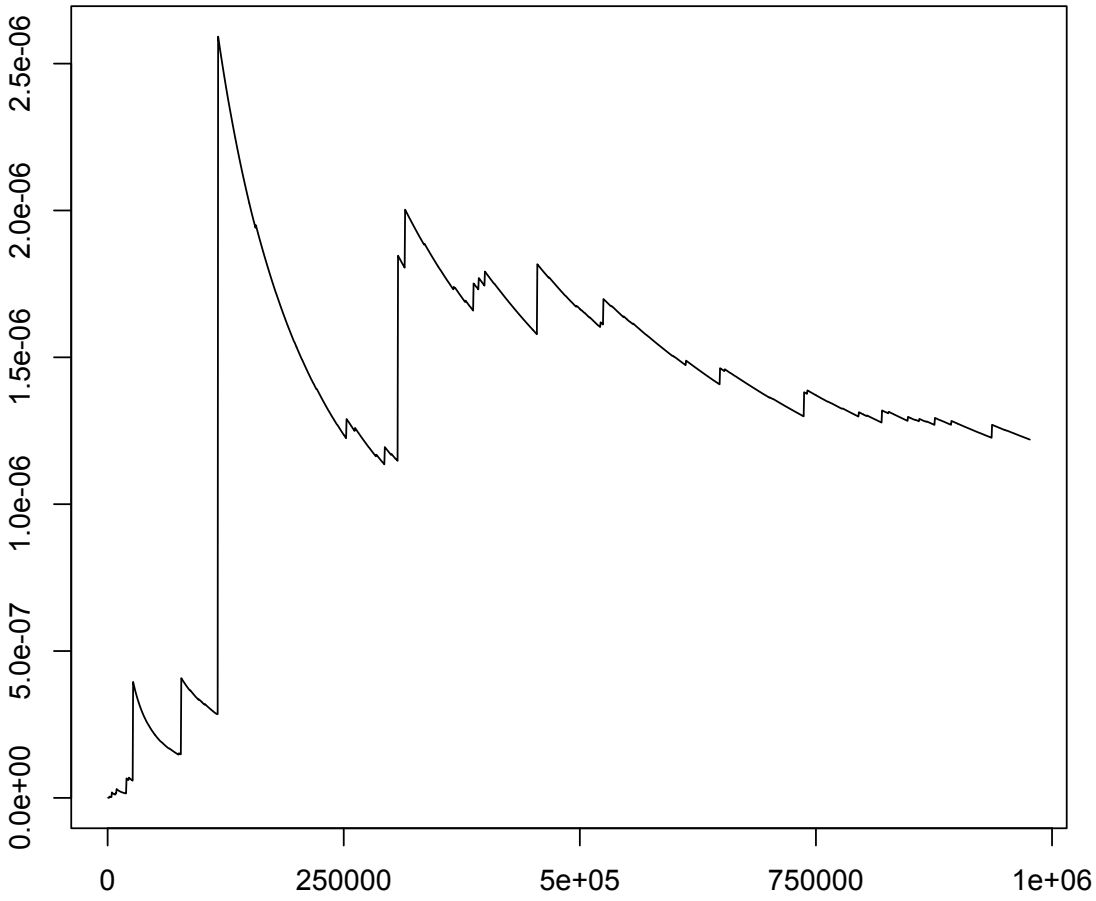

Figure 2.4: **Transition of estimates for the defect rate.** The horizontal axis indicates $\tilde{n}$ and the vertical axis indicates the evaluation in (2.14).

2.8 Summary in this chapter

In this chapter, under the situation where various variables may cause a defect, we have treated a problem to actively estimate the discriminant function which determines the

probability of causing the defect. And then, we have discovered that even the ASVM, the latest active learning method in the nonlinear discriminant analysis, does not work well for the case where the nonlinearity of the discriminant function is strong and the region producing the defect more likely than not is separated. To overcome this difficulty, we have proposed the LSVM which uses a linear discriminant method by consciously forgetting the nonlinearity of the discriminant function at the sampling stage in active learning. In numerical studies, we have simulated the cases where the region is actually separated, and then it has been checked that the LSVM is superior to the ASVM for such cases. Also it has been checked through real data analysis that the error rate for the LSVM is smaller than that for the ASVM and becomes stable quickly. Moreover, we have proposed a method to efficiently estimate the defect rate by use of the importance sampling after obtaining the estimated discriminant function by the LSVM. We have used a single Gaussian distribution for the importance sampling, but we may be able to evaluate it faster by using a multi-modal distribution such as a Gaussian mixture.

The above-mentioned case is natural for the defect rate evaluation problem, and so we can say that our simple active learning method is useful in product manufacturing, that is, valuable from engineering viewpoint. On the other hand, brushing up the method is our important future theme in order to cope with the case of existing more variables. One idea is to make a hybrid-type active learning method by combining the LSVM and the ASVM so that the weak and strong points of the ASVM are respectively overcome and kept.

Chapter 3

Ridge-type Regularization Method for Questionnaire Data Analysis

In the questionnaire study, it is often required to answer to the several evaluation items for the objects should be evaluated. However, to reduce the burden of the evaluator, he/she frequently gives a reply to a part of the objects when those are large amounts. Then the answers that were not evaluated is analyzed by regarding it as a missing data. Although the method such as the collaborative filtering or matrix completion were proposed in order to predict a missing part of such data, we have no evaluation formula for the variance of the predicted value. In this chapter, we consider the model, which can be interpreted as an extension of the GMANOVA model, by using a dummy matrix representing the existence of the missing. We propose a C_p criterion in order to choose the tuning parameter which is included in the Ridge-type estimator. Through simulation studies and real data analysis, the validity of the proposed method is checked.

3.1 Background

In the questionnaire study, n evaluators are often required to the K evaluation items for M of the J objects. For instance, they evaluate on a scale 1 to 5 for each of the evaluation items. Because the evaluator answers only for M objects, replies for the rest of $J - M$ objects are missing, i.e., we can not observe them. Therefore, we have a 3-dimensional array data of size $J \times K \times n$ consisting of the MKn observations and $(J-M)Kn$ missing

values.

For such a data, it is often interested in predicting the missing values based on the observations (for example, recommendation systems of the Amazon or Netflix). Nowadays the methods such as the collaborative filtering or matrix completion are developed to predict the missing part, although it is, in general, impossible to predict them unless we do not assume any conditions for the data structure (see e.g., Jannach et al. 2010; Candès and Recht 2009; Koltchinskii et al. 2011). These methods reconstruct a matrix based on the small observations by assuming the data being low-rank, and are useful since we can use a popular method of convex optimization. However, there is a problem such as the selection of a tuning parameter which is included the model and, in general, it is difficult to evaluate the prediction accuracy because we have no evaluation formula for the variance of the predicted value.

On the other hand, the correspondence analysis has been used in the questionnaire data analysis in order to extract features from the data (e.g., Benzécri 1992). The correspondence analysis, however, is an exploratory method as well as the principal component analysis and is not applicable to the data including the missing values. So we can not use it to analyze our data. As described in Section 5.2, it is possible to construct a parametric model by using a dummy matrix representing the existence of the missing value. The model we consider can formally be interpreted as an extension of the generalized multivariate analysis of variance (GMANOVA) model in Potthoff and Roy (1964) to 3-dimensional array data. We then can express the ordinal least squares estimator explicitly and, moreover, evaluate the average or variance of the estimator. However, there is a problem that the estimator becomes unstable when M or K are large or the multicollinearity is present in the data.

The Ridge-type estimator were often used in order to assure the stability of the estimator (e.g., Hoerl and Kennard 1970). We then need to choose appropriate tuning parameters which is included in the estimator. Computational methods such as the cross

validation (CV: Stone (1974)) are usually used to choose tuning parameters although it requires a considerable number of computational costs when the number of the parameter is large. An information criterion such as C_p (Mallows 1973, 1995) may also be used to choose it. Nagai (2011) has derived an unbiased estimator of a standardized mean squared error for the Ridge-type estimator in the GMANOVA model. However the objective variable of our data is an $(M \times K)$-dimensional matrix, and thus we can not apply his result directly. Moreover, since this kind of data is discrete in general, it is unnatural to assume the normality to a noise.

Although it is very important task to predict the missing part for such a the questionnaire data, our goal is to construct a good model for the future observations in a prediction perspective as a first step. To do this we derive an unbiased estimator of the standardized mean squared error for the model that is defined in Section 3.2. Moreover in Section 3.3, after describing some preparations, we develop a C_p-type information criterion in order to choose tuning parameters which is included in the estimator. The method is shown to be valid through a simulation study in Section 3.4, and then we report the result in which the method is applied to real data in Section 3.5. Finally, some concluding remarks are presented in Section 3.6.

3.2 Setting and assumptions

Let $\mathcal{J} = \{1, 2, , \ldots, J\}$ be an index set of items, and $\mathcal{J}_i = \{j_{i1}, j_{i2}, \ldots, j_{iM}\}$ be a subset of $\mathcal{J}$ arranged in ascending order. We consider the response for i-th subject as an $(M \times K)$-dimensional matrix $Y_i = (y_{ij_mk})_{m=1,2,\ldots,M;k=1,2,\ldots,K}$, and the model

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk},$$

where μ is a general mean, α_j and β_k is a main effect, γ_{jk} is an interaction effect between the subject j and the item j, and ε_{ijk} is an unknown noise. Note that we can not observe the response y_{ijk}, i.e., y_{ijk} is missing, whenever $j \notin \mathcal{J}_i$. Let $\tilde{X}_i$ be a $(M \times J)$-dimensional

matrix whose (m, j)-th element is 1 when $j = j_m$ and 0 otherwise. Then, we can rewrite this model as

$$Y_i = X_i B A + E_i, \tag{3.1}$$

where $X_i = (\mathbf{1}_M, \tilde{X}_i) \in \mathbb{R}^{M\times(J+1)}$, $A = (\mathbf{1}_K, I_K)^{\mathrm{T}} \in \mathbb{R}^{(K+1)\times K}$, and

$$B = \begin{pmatrix} \mu & \beta_1 & \cdots & \beta_K \\ \alpha_1 & \gamma_{11} & \cdots & \gamma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_J & \gamma_{J1} & \cdots & \gamma_{JK} \end{pmatrix} \in \mathbb{R}^{(J+1)\times(K+1)}.$$

For a positive integer d, we will denote $\mathbf{1}_d$ and I_d by a d-dimensional one-vector and $(d\times d)$-dimensional identity matrix. Suppose that $E_i = (\boldsymbol{\varepsilon}_{i1}, \boldsymbol{\varepsilon}_{i2}, \ldots, \boldsymbol{\varepsilon}_{iM}) = (\varepsilon_{ij_mk})_{m=1,2,\ldots,M;k=1,2,\ldots,K}$ are independent random matrices with mean $\mathrm{E}[E_i] = O$ and covariance $\mathrm{V}[E_i] = \Sigma \otimes \Xi$, where Σ and Ξ is an unknown $(K \times K)$-dimensional matrix and a known $(M \times M)$-dimensional matrix, respectively. This means that each columns of E_i have a covariance $\mathrm{E}[\boldsymbol{\varepsilon}_{ij}\boldsymbol{\varepsilon}_{ik}^{\mathrm{T}}] = \sigma_{jk}\Xi$ for $j, k = 1, 2, \ldots, M$. For a random matrix M, we will denote $\mathrm{E}[\mathrm{vec}(M - \mathrm{E}[M])\mathrm{vec}(M - \mathrm{E}[M])^{\mathrm{T}}]$ by $\mathrm{V}[M]$, where $\mathrm{vec}(M)$ is a random vector obtained by stacking its columns into a single vector. Although j_m's are assigned randomly, we think that X_i are deterministic for simplicity. Note that this model includes a so-called GMANOVA model of Potthoff and Roy (1964) in a special case when $M = 1$ and E_i is distributed according to some Gaussian distribution.

To avoid a redundancy of the model, we impose

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = \sum_{j=1}^{J} \gamma_{jk} = \sum_{k=1}^{K} \gamma_{jk} = 0$$

on the parameter as is often used in the ANOVA model. Since

$$\alpha_J = -\sum_{j=1}^{J-1} \alpha_j, \quad \beta_K = -\sum_{k=1}^{K-1} \beta_k, \quad \gamma_{Jk} = -\sum_{j=1}^{J-1} \gamma_{jk}, \qquad \text{and} \qquad \gamma_{jK} = -\sum_{k=1}^{K-1} \gamma_{jk},$$

we can remove this restriction. Indeed by defining $C = (I_{J-1}, -\mathbf{1}_{J-1})^{\mathrm{T}}$, $D = (I_{K-1}, -\mathbf{1}_{K-1})^{\mathrm{T}}$, $\tilde{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, \ldots, \alpha_{J-1})^{\mathrm{T}}$, $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_{K-1})^{\mathrm{T}}$ and $\tilde{\Gamma} = (\gamma_{jk})_{j=1,2,\ldots,J-1;k=1,2,\ldots,K-1}$, B can be

rewritten as

$$B = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & C \end{pmatrix} \begin{pmatrix} \mu & \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \\ \tilde{\boldsymbol{\alpha}} & \tilde{\Gamma} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & D^{\mathrm{T}} \end{pmatrix},$$

and thus we can redefine

$$X_i \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & C \end{pmatrix} \in \mathbb{R}^{M \times J}, \quad \begin{pmatrix} \mu & \tilde{\boldsymbol{\beta}}^{\mathrm{T}} \\ \tilde{\boldsymbol{\alpha}} & \tilde{\Gamma} \end{pmatrix} \in \mathbb{R}^{J \times K}$$

and

$$\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & D^{\mathrm{T}} \end{pmatrix} A = \begin{pmatrix} \mathbf{1}_{K-1}^{\mathrm{T}} & 1 \\ I_{K-1} & -\mathbf{1}_{K-1} \end{pmatrix} \in \mathbb{R}^{K \times K}$$

by X_i, B and A, respectively.

In the following, let us denote $\sum_{i=1}^n X_i^{\mathrm{T}} X_i$ and $\sum_{i=1}^n X_i^{\mathrm{T}} Y_i$ by $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$, respectively. Then an ordinary least square estimator of the model (3.1), that is, an minimizer of $\sum_{i=1}^n \|Y_i - X_i B A\|_F^2$, is given by

$$B^{\dagger} = (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} Y A^{\mathrm{T}} (A A^{\mathrm{T}})^{-1},$$

where $\|\cdot\|_F$ denotes a Frobenius norm, i.e, $\|M\|_F = (\mathrm{tr}(M^{\mathrm{T}}M))^{1/2}$ for a matrix M. It is easy to see that $B^{\dagger}$ is an unbiased and consistent estimator of B. Moreover, if E_i and X_i satisfy that,

- $n^{-1}\sum_{i=1}^n X_i^{\mathrm{T}} X_i$ and $n^{-1}\sum_{i=1}^n X_i^{\mathrm{T}} \Xi X_i$ converge to an positive definite matrix P and Q, respectively.
- $\mathrm{E}[\|E_i\|_F^4]$ exist for any $i = 1, 2, \ldots, n$,

then the Lindeberg central limit theorem and the Cramér-Wold device imply that $\sqrt{n}(B^{\dagger} - B)$ converges in distribution to a Gaussian distribution with mean O and covariance $(AA^{\mathrm{T}})^{-1} A \Sigma A^{\mathrm{T}} (AA^{\mathrm{T}})^{-1} \otimes P^{-1} Q P^{-1}$. In addition, an unbiased estimator of Σ is given by

$$\hat{\Sigma} = \frac{1}{n\mathrm{tr}(\Xi) - \tilde{J}} \sum_{i=1}^n (Y_i - X_i B^{\dagger} A)'(Y_i - X_i B^{\dagger} A), \tag{3.2}$$

where $\tilde{J} = \mathrm{tr}[\{\sum_{i=1}^{n} X_i(X'X)^{-1}X_i\}\Xi]$. The details for deriving (3.2) are given in Appendix 6.2.2.

However, when J or K are large enough, the inverse of $X^{\mathrm{T}}X$ or AA^{T} may not be exist or the variance of the estimator may become unstable, so we consider the Ridge-type estimator given by

$$\hat{B}_{\lambda,\mu} = (X^{\mathrm{T}}X + \lambda I_J)^{-1}X^{\mathrm{T}}YA^{\mathrm{T}}(AA^{\mathrm{T}} + \mu I_K)^{-1}, \tag{3.3}$$

where λ and μ are positive constants, also known as tuning parameter (see, e.g., Hoerl and Kennard 1970; Nagai 2011).

3.3 Deriving the C_p criterion

3.3.1 Preparation

Nagai (2011) derived a C_p criterion for Ridge-type estimator in the GMANOVA model. The difference between his result and ours is that whether there is missing values of the data and that the i-th observation is an $(M \times K)$-dimensional matrix in our case. Moreover, we do not assume that the normality of E_i.

To derive a C_p criterion, we need some preparation with matrix calculation. Let us define

$$H_\mu = A^{\mathrm{T}}(AA^{\mathrm{T}} + \mu I_K)^{-1}A \quad \text{and} \quad G_\lambda = (X^{\mathrm{T}}X + \lambda I_J)^{-1}.$$

Note that by the definition of A,

$$AA^{\mathrm{T}} = \begin{pmatrix} K & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}} \end{pmatrix}.$$

Because

$$\{(1+\mu)I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}}\}^{-1} = \frac{1}{1+\mu}\left(I_{K-1} - \frac{1}{K+\mu}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}}\right)$$

from (6.1), it follows that

$$H_\mu = \frac{1}{(1+\mu)(K+\mu)}\begin{pmatrix} \tilde{H}_\mu & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & K(1+\mu) \end{pmatrix}, \tag{3.4}$$

where $\tilde{H}_\mu = (K+\mu)I_{K-1} + \mu \mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}}$.

Next, we see that

$$X_i = (\mathbf{1}_M, \tilde{X}_i) \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & C^{\mathrm{T}} \end{pmatrix} = (\mathbf{1}_M, \tilde{X}_i C)$$

and thus we have

$$X^{\mathrm{T}}X = \sum_{i=1}^{n} \begin{pmatrix} M & \mathbf{1}_M^{\mathrm{T}}\tilde{X}_i C \\ C^{\mathrm{T}}\tilde{X}_i^{\mathrm{T}}\mathbf{1}_M & C^{\mathrm{T}}\tilde{X}_i^{\mathrm{T}}\tilde{X}_i C \end{pmatrix}.$$

By letting $\boldsymbol{\delta} = \sum_{i=1}^{n} \tilde{X}_i^{\mathrm{T}}\mathbf{1}_M$, δ_j represents the total number of assignment of j-th item and $\Delta = \sum_{i=1}^{n} \tilde{X}_i^{\mathrm{T}}\tilde{X}_i$ is a diagonal matrix whose (j,j)-th element is δ_j. From Lemma 6, we have

$$G_\lambda = \frac{1}{(nM+\lambda)^2} \begin{pmatrix} (nM+\lambda) + \boldsymbol{\delta}^{\mathrm{T}} C \tilde{G}_\lambda^{-1} C^{\mathrm{T}} \boldsymbol{\delta} & (nM+\lambda)\boldsymbol{\delta}^{\mathrm{T}} C \tilde{G}_\lambda^{-1} \\ (nM+\lambda)\tilde{G}_\lambda^{-1} C^{\mathrm{T}} \boldsymbol{\delta} & (nM+\lambda)^2 \tilde{G}_\lambda^{-1} \end{pmatrix}, \tag{3.5}$$

where

$$\tilde{G}_\lambda = C^{\mathrm{T}} \left(\Delta - \frac{1}{nM+\lambda} \boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}} \right) C + \lambda I_{J-1}.$$

Let us define $\tilde{\Delta} = \Delta - (nM+\lambda)^{-1}\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}$. Then, from Lemma 7, we see that

$$\tilde{\Delta}^{-1} = \Delta^{-1} + \frac{1}{\lambda}\mathbf{1}_J\mathbf{1}_J^{\mathrm{T}}, \tag{3.6}$$

since $\Delta^{-1}\boldsymbol{\delta} = \mathbf{1}_J$ and $\boldsymbol{\delta}^{\mathrm{T}}\mathbf{1}_J = nM$. Moreover, by using Lemma 7, we have

$$\tilde{G}_\lambda^{-1} = \frac{1}{\lambda} I_{J-1} - \frac{1}{\lambda^2} C^{\mathrm{T}} \left(\tilde{\Delta}^{-1} + \frac{1}{\lambda} C C^{\mathrm{T}} \right)^{-1} C.$$

Let $\Delta_{J-1} \in \mathbb{R}^{(J-1)\times(J-1)}$ be a diagonal matrix whose (j,j)-th element is δ_j, then it is easy to see that

$$\tilde{\Delta}^{-1} + \frac{1}{\lambda} C C^{\mathrm{T}} = \begin{pmatrix} \Delta^{\dagger} + \lambda^{-1}\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \delta_J^{-1} + \lambda^{-1} J \end{pmatrix}$$

from (3.6), where $\Delta^{\dagger} = \Delta_{J-1}^{-1} + \lambda^{-1} I_{J-1}$ is a diagonal matrix whose diagonal element is given by $\delta_j^{-1} + \lambda^{-1}$ for $j = 1, 2, \ldots, J-1$. Let us define $P = \Delta_{J-1}(\Delta_{J-1} + \lambda I_{J-1})^{-1}$, and note that $\Delta^{-\dagger} = \lambda P$. Then, by using Lemma 7 again, we have

$$(\Delta^{\dagger} + \lambda^{-1}\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}})^{-1} = \Delta^{-\dagger} - \frac{\Delta^{-\dagger}\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}}\Delta^{-\dagger}}{\lambda + \mathbf{1}_{J-1}^{\mathrm{T}}\Delta^{-\dagger}\mathbf{1}_{J-1}} = \lambda \left(P - \frac{P\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}}P}{1 + \mathrm{tr}(P)} \right).$$

In the second equality, we just use $\mathbf{1}_{J-1}^{\mathrm{T}}\Delta_{J-1}^{-\dagger}\mathbf{1}_{J-1} = \lambda\mathrm{tr}(P)$. Finally, we obtain that

$$\tilde{G}_{\lambda}^{-1} = \frac{1}{\lambda}\left\{I_{J-1} - P + \frac{P\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}}P}{1+\mathrm{tr}(P)} - \frac{\delta_J}{\lambda + J\delta_J}\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}}\right\}. \tag{3.7}$$

3.3.2 Main result

Now, we can derive the C_p criterion as an unbiased estimator of a standardized mean squared error (MSE) defined by

$$\sum_{i=1}^{n}\mathrm{E}\Big[\mathrm{vec}(\hat{Y}_i - \mathrm{E}[Y_i])^{\mathrm{T}}(\Sigma\otimes\Xi)^{-1}\mathrm{vec}(\hat{Y}_i - \mathrm{E}[Y_i])\Big\}\Big],$$

where $\hat{Y}_i = X_i\hat{B}_{\lambda,\mu}A$. From Lemma 8, this is equivalent to

$$\sum_{i=1}^{n}\mathrm{E}\Big[\mathrm{tr}\Big\{(\hat{Y}_i - \mathrm{E}[Y_i])^{\mathrm{T}}\Xi^{-1}(\hat{Y}_i - \mathrm{E}[Y_i])\Sigma^{-1}\Big\}\Big].$$

Because $\mathrm{E}[Y_i] = Y_i - E_i$ and

$$\mathrm{E}[\mathrm{tr}\{\Sigma^{-1}(Y_i - \hat{Y}_i)^{\mathrm{T}}E_i\}] = \mathrm{E}[\mathrm{tr}\{\Sigma^{-1}E_i^{\mathrm{T}}\mathrm{E}_i\}] - \mathrm{E}[\mathrm{tr}\{\Sigma^{-1}\hat{Y}_i^{\mathrm{T}}E_i\}],$$

MSE can be rewritten as

$$\begin{aligned}\mathrm{MSE} =& \sum_{i=1}^{n}\mathrm{E}[\{\mathrm{tr}\{(Y_i - \hat{Y}_i)^{\mathrm{T}}\Xi^{-1}(Y_i - \hat{Y}_i)\Sigma^{-1}\}] \\ &- \sum_{i=1}^{n}\mathrm{E}[\mathrm{tr}(E_i^{\mathrm{T}}\Xi^{-1}E_i\Sigma^{-1})] + 2\sum_{i=1}^{n}\mathrm{E}[\mathrm{tr}(\hat{Y}_i^{\mathrm{T}}\Xi^{-1}E_i\Sigma^{-1}\}].\end{aligned} \tag{3.8}$$

By using Lemma 8 and $V[E_i] = \Sigma\otimes\Xi$, the second term of the right-hand side in (3.8) can be reduced to nMK since

$$\begin{aligned}\mathrm{E}[\mathrm{tr}(E_i^{\mathrm{T}}\Xi^{-1}E_i\Sigma^{-1})] =& \mathrm{E}[\mathrm{vec}(E_i)^{\mathrm{T}}(\Sigma^{-1}\otimes\Xi^{-1})\mathrm{vec}(E_i)] \\ =& \mathrm{tr}(I_M\otimes I_K) = \mathrm{tr}(I_M)\mathrm{tr}(I_K) = MK.\end{aligned}$$

Next, we evaluate the third term of the right-hand side in (3.8). By the definition of the model in (3.1), we have

$$\hat{Y}_i = \sum_{j=1}^{n}X_iG_\lambda X_j^{\mathrm{T}}Y_jH_\mu = X_iG_\lambda X^{\mathrm{T}}XBAH_\mu + \sum_{j=1}^{n}X_iG_\lambda X_j^{\mathrm{T}}E_jH_\mu.$$

Because the first term of the right-hand side in this equality is non-stochastic, and E_i are independent, so we see that

$$\begin{aligned}\mathrm{E}[\mathrm{tr}(\hat{Y}_i^{\mathrm{T}}\Xi^{-1}E_i\Sigma^{-1})] &= \sum_{j=1}^{n}\mathrm{E}[\mathrm{vec}(E_j)^{\mathrm{T}}(H_\mu\Sigma^{-1}\otimes X_jG_\lambda X_i^{\mathrm{T}}\Xi^{-1})\mathrm{vec}(E_i)] \\ &= \mathrm{tr}(H_\mu\otimes G_\lambda X_i^{\mathrm{T}}X_i) = \mathrm{tr}(H_\mu)\mathrm{tr}(G_\lambda X_i^{\mathrm{T}}X_i)\end{aligned}$$

from (6.2), and thus the third term of the right-hand side in (3.8) can be reduced to $\mathrm{tr}(H_\mu)\mathrm{tr}(G_\lambda X^{\mathrm{T}}X)$. From(3.4), we have

$$\mathrm{tr}(H_\mu) = \frac{1}{(1+\mu)(K+\mu)}\{\mathrm{tr}(\tilde{H}_\mu) + K(1+\mu)\} = \frac{K^2+3K\mu-2\mu}{(1+\mu)(K+\mu)}. \tag{3.9}$$

On the other hand, by the definition of G_λ, we have $G_\lambda X^{\mathrm{T}}X = I_J - \lambda G_\lambda$ and

$$\lambda\mathrm{tr}(G_\lambda) = \frac{\lambda}{nM+\lambda} + \frac{\lambda\boldsymbol{\delta}^{\mathrm{T}}C\tilde{G}_\lambda^{-1}C^{\mathrm{T}}\boldsymbol{\delta}}{(nM+\lambda)^2} + \lambda\mathrm{tr}(\tilde{G}_\lambda^{-1})$$

from (3.5). Since $\mathrm{tr}(P\mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}}P) = \mathrm{tr}(P^2)$, the third term in the right-hand side can be reduced to

$$J - 1 - \mathrm{tr}(P) + \frac{\mathrm{tr}(P^2)}{1+\mathrm{tr}(P)} - \frac{\delta_J(J-1)}{\lambda+J\delta_J}$$

from (3.7). Moreover, by a simple calculation, we have

$$\begin{aligned}\boldsymbol{\delta}^{\mathrm{T}}C\mathbf{1}_{J-1} &= (nM+\lambda) - (\lambda+J\delta_J), \\ \boldsymbol{\delta}^{\mathrm{T}}CP\mathbf{1}_{J-1} &= (nM+\lambda) - (\lambda+\delta_J)(1+\mathrm{tr}(P)),\end{aligned}$$

and

$$\boldsymbol{\delta}^{\mathrm{T}}C(I_{J-1}-P)C^{\mathrm{T}}\boldsymbol{\delta} = -\lambda(\lambda+\delta_J)^2(nM-J\delta_J)\mathrm{tr}(P) + \delta_J(J-1)(\lambda+\delta_J).$$

Then, it follows that

$$\lambda\boldsymbol{\delta}^{\mathrm{T}}C\tilde{G}_\lambda^{-1}C^{\mathrm{T}}\boldsymbol{\delta} = (nM+\lambda)^2\left(\frac{1}{1+\mathrm{tr}(P)} - \frac{\delta_J}{\lambda+J\delta_J}\right) - \lambda(nM+\lambda),$$

and thus we have

$$\mathrm{tr}(G_\lambda X^{\mathrm{T}}X) = 1 + \mathrm{tr}(P) - \frac{1+\mathrm{tr}(P^2)}{1+\mathrm{tr}(P)} + \frac{J\delta_J}{\lambda+J\delta_J}. \tag{3.10}$$

Combining all the above, we obtain the following theorem:

Theorem 1. An unbiased estimator of MSE in (3.8) is given by

$$\sum_{i=1}^{n} \mathrm{tr}\{(Y_i - \hat{Y}_i)^{\mathrm{T}} \Xi^{-1} (Y_i - \hat{Y}_i) \Sigma^{-1}\} - nMK + 2\mathrm{tr}(G_\lambda X^{\mathrm{T}} X)\mathrm{tr}(H_\mu),$$

where $\mathrm{tr}(G_\lambda X^{\mathrm{T}} X)$ and $\mathrm{tr}(H_\mu)$ can be evaluated as (3.9) and (3.10), respectively.

As a result, we propose the following index as an C_p-type information criterion:

$$C_p = \sum_{i=1}^{n} \mathrm{tr}\{(Y_i - \hat{Y}_i)^{\mathrm{T}} \Xi^{-1} (Y_i - \hat{Y}_i) \hat{\Sigma}^{-1}\} - nMK + 2\mathrm{tr}(G_\lambda X^{\mathrm{T}} X)\mathrm{tr}(H_\mu), \tag{3.11}$$

where $\hat{\Sigma}$ is an unbiased estimator of Σ defined in (3.2). By minimizing the C_p in (3.11), we can obtain the optimal value of the tuning parameter (λ, μ).

3.4 Simulation study

In this section, we conduct some simulation studies to check the performance of the tuning parameter selection based on the C_p in (3.11). The performances for the C_p and CV are compared.

Concretely speaking, we assessed the performances in terms of the following prediction squared error:

$$\mathrm{PSE} = \mathrm{E}\left[\left\{(\tilde{Y}_i - \tilde{X}_i \hat{B}_{\hat{\lambda},\hat{\mu}} A)^{\mathrm{T}} \Xi^{-1} (\tilde{Y}_i - \tilde{X}_i \hat{B}_{\hat{\lambda},\hat{\mu}} A) \hat{\Sigma}^{-1}\right\}\right], \tag{3.12}$$

where $\hat{\lambda}$ and $\hat{\mu}$ are the value of the tuning parameters given by each of the criteria, and $\hat{\Sigma}$ is an unbiased estimator of Σ given by (3.2). The expectation in (3.12) is evaluated using an empirical mean of N $(= 1,000)$ tuples of the test data $\{(\tilde{Y}_i, \tilde{X}_i); i = 1, 2, \ldots, N\}$ and we interpret that the criterion giving the small value of the PSE is good. Moreover, we checked the standard deviation of the difference between the values of PSE given by each criteria. This represents a stability of the difference between the values of PSE given by each criteria, that is, we interpret that the difference is significant when its value is small. We also checked the computation time (sec) to compute the C_p and CV for fixed tuning parameters as a secondary index for the assessment.

The simulation settings were as follows. First, we made $(M \times K)$-dimensional matrices X_i $(i = 1, 2, \ldots, n)$ based on the sampling without replacement from $\{1, 2, \ldots, J\}$ with probability J^{-1}. We then constructed Y_i based on the model in (3.1) for each $i = 1, 2, \ldots, n$, and rounded so the elements of Y_i is in $\{1, 2, \ldots, 5\}$. Next, we used $\Sigma = (0.5^{i-j})_{i,j=1,2,\ldots,K}$ and $\Xi = (1-\rho)I_M + \rho \mathbf{1}_M \mathbf{1}'_M$ with known $\rho \in [0, 1]$. The true parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{J-1})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{J-1})^{\mathrm{T}}$, $\Gamma = (\gamma_{jk})_{j=1,2,\ldots,J-1;k=1,2,\ldots,K-1}$ are drawn from

$$\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, 10^{-3}(I_{J-1} + \mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}})), \qquad \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, 10^{-3}(I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}}))$$
$$\Gamma \sim \mathrm{N}(\mathbf{0}, 10^{-3}(I_{K-1} + \mathbf{1}_{K-1}\mathbf{1}_{K-1}^{\mathrm{T}}) \otimes (I_{J-1} + \mathbf{1}_{J-1}\mathbf{1}_{J-1}^{\mathrm{T}})),$$

and $\mu = 3$. In this case, we can easily evaluate

$$n\mathrm{tr}(\Xi) - \tilde{J} = nM - J \tag{3.13}$$

for an un biased estimator of Σ in (3.2). The details for deriving (3.13) are given in Appendix 6.2.2. n was set to 500 or 1000, and nine cases were considered for three-tuple (J, M, K), and fifty simulations were conducted.

Tables 3.1 and 3.2 show the results for $\rho = 0$ and $\rho = 0.5$, and in each table, the average and standard deviation of the PSE, and the standard deviation of the difference between the values of the PSE given by each criteria ($\sqrt{\mathrm{V}[C_p - \mathrm{CV}]}$) are provided. In each case, we see that both of the average and standard deviation of PSE based on the C_p in (3.11) are smaller than those of CV. Moreover, comparing $n = 500$ and $n = 1,000$ in the same (J, M, K), the value of $\sqrt{\mathrm{V}[C_p - \mathrm{CV}]}$ become small, and thus we can say that the difference between the values of the PSE given by C_p and CV is significant as n increases.

On the other hand, Figure 3.1 show the comparison of the computation time to compute the C_p and CV for fixed tuning parameters. We also fixed M and K on 5 and 4, respectively. In the left, we can see that the computation time to compute CV is increase

(J, M, K)	n	C_p (sd)	CV (sd)	$\sqrt{\mathrm{V}[C_p - \mathrm{CV}]}$
$(30, 5, 2)$	500	11.726 (0.432)	12.764 (0.697)	0.516
	1000	11.370 (0.369)	11.661 (0.480)	0.226
$(60, 5, 2)$	500	13.311 (0.713)	16.189 (1.547)	1.146
	1000	12.136 (0.408)	13.003 (0.703)	0.476
$(90, 5, 2)$	500	14.544 (0.688)	21.219 (2.249)	1.918
	1000	12.819 (0.432)	14.757 (0.835)	0.664
$(30, 5, 4)$	500	23.623 (0.931)	25.612 (1.391)	1.068
	1000	22.411 (0.539)	22.870 (0.694)	0.326
$(60, 5, 4)$	500	25.642 (1.207)	31.381 (3.331)	2.484
	1000	23.919 (0.715)	26.073 (1.317)	1.05
$(90, 5, 4)$	500	26.141 (0.835)	37.661 (4.553)	4.198
	1000	24.376 (0.616)	27.986 (1.371)	1.113
$(30, 10, 4)$	500	43.211 (1.101)	43.525 (1.230)	0.336
	1000	41.958 (0.664)	42.084 (0.663)	0.161
$(60, 10, 4)$	500	46.600 (1.136)	49.628 (1.929)	1.475
	1000	44.857 (0.781)	45.629 (0.831)	0.408
$(90, 10, 4)$	500	48.550 (1.040)	56.096 (3.186)	2.735
	1000	45.542 (0.956)	47.192 (0.991)	0.611

Table 3.1: Comparison between C_p and CV for simulated data with $\rho = 0$.

although that of C_p is not so much change as n or J increase. An enlarged view of computation time of C_p is drawn in the right. Since the difference among each n and J is really small, we can say that the computation of C_p is robust to scale changes. Moreover, we can easily implement because the C_p in (3.11) has the simple form. On the whole, we can conclude that the C_p in (3.11) is better than the CV.

3.5 Real data analysis

In this section, we compare the methods through applying them to real data. Although this data, in fact, consist of three categories, we consider the data is indepen-

(J, M, K)	n	C_p (sd)	CV (sd)	$\sqrt{\mathrm{V}[C_p - \mathrm{CV}]}$
$(30, 5, 2)$	500	12.905 (0.670)	14.282 (1.080)	0.892
	1000	12.294 (0.650)	12.609 (0.774)	0.288
$(60, 5, 2)$	500	15.197 (1.043)	19.896 (2.325)	1.899
	1000	13.760 (0.659)	14.895 (0.752)	0.481
$(90, 5, 2)$	500	16.520 (0.950)	28.466 (4.101)	3.773
	1000	14.553 (0.560)	17.854 (1.386)	1.101
$(30, 5, 4)$	500	24.779 (1.289)	26.680 (1.715)	1.126
	1000	22.189 (0.644)	22.972 (0.832)	0.548
$(60, 5, 4)$	500	27.027 (1.295)	36.329 (4.517)	3.960
	1000	24.047 (0.714)	27.204 (1.710)	1.322
$(90, 5, 4)$	500	28.378 (1.413)	44.593 (5.329)	4.685
	1000	25.965 (0.635)	33.100 (1.840)	1.698
$(30, 10, 4)$	500	45.319 (1.232)	46.186 (1.577)	0.725
	1000	42.442 (1.009)	42.490 (0.917)	0.491
$(60, 10, 4)$	500	52.135 (3.347)	56.454 (3.215)	3.110
	1000	46.593 (1.320)	48.044 (1.696)	0.841
$(90, 10, 4)$	500	55.531 (4.301)	65.578 (4.003)	3.381
	1000	48.748 (1.115)	50.934 (1.566)	1.337

Table 3.2: Comparison between C_p and CV for simulated data with $\rho = 0.5$.

dent for each categories for simplicity. For each categories, we know that $(n, J) = (1884, 60),\ (1364, 21),\ (1425, 44),\ K = 4$ and $M = 5$.

We used 1,200 data sampled at random as a training data for each categories, and the rest of the data as a test data. Table 3.2 shows the PSE in (3.12) evaluated by using the test data after selecting the tuning parameter based on C_p and CV for each categories and $\rho = 0$ and 0.5. As mentioned in Section 3.4, the criterion giving the small value of the PSE is good, and we can say that the tuning parameter selection based on C_p is superior to that of CV. Looking at the result of C_p, we see that the value of PSE with $\rho = 0.5$ is small for categories 1 and 3, and that the value of PSE with $\rho = 0$ is small for category

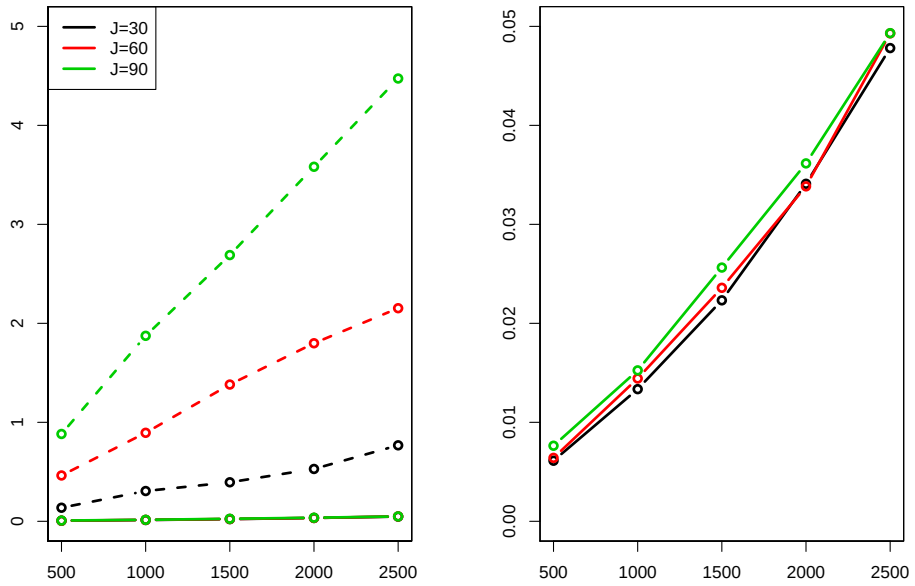

Figure 3.1: **Transition of computation time.** In each figures, the horizontal axis indicates the sample size n and the vertical axis indicates the average of computation time (sec) for fixed tuning parameters. In the left, the solid and the dashed line represents the average of computation time via C_p and CV, respectively. An extended figure of the average of computation time bia C_p is drawn in the right.

2. Thus it is suggested that the correlation among the objects in category 2 is smaller than those of other categories. In addition, while it may seem that there is no significance of the result between C_p and CV, this is probably due to the size of test data is small compared to categories 1 and 3.

3.6 Summary in this chapter

In this chapter, we have considered the model obtained from the questionnaire study and proposed the C_p-type information criterion to choose the tuning parameters which is included in the estimator. By using a dummy matrix representing the existence of the

ρ	category	C_p	CV
0	1	21.433	22.239
	2	20.323	20.364
	3	19.989	20.177
0.5	1	19.861	20.390
	2	20.523	20.566
	3	19.000	19.221

Table 3.3: Comparison between C_p and CV for real data.

missing, the model that is interpreted as a extension of the GMANOVA model to a 3-dimensional array data has constructed at first. Although the model we have considered is somewhat special in a sense that it has the particular structure for X_i and A, it is expected in the future that the similar results may be obtained even for more complex models since we have relaxed the assumption of the normality of the noise. For the simplicity, though we have assumed that X_i is non-random and that Ξ is known, we also need to estimate Ξ properly to make more flexible model. These are the future works.

The result obtained in this chapter has been based on the main effect of the object and evaluation items, and the interaction effect between these. It is well known that the predicting the missing part is important when we construct a recommendation systems in the recent questionnaire study and WEB survey. However as mentioned in the beginning, it is in general difficult to evaluate the prediction accuracy for the methods such as the collaborative filtering or matrix completion. To overcome this problem, by extending the model in this chapter to the model which contains the random effect for the evaluators, it might be possible to predict the missing part and to do a statistical inference for such a data.

We have confirmed the usefulness of C_p given in Section 3.3 through the simulation study and real data analysis. The proposed method has have high prediction accuracy and low computational costs compared to the CV, because this criterion could be expressed by

a simple form and obtained explicitly. Since the real data consists of the three categories, we need to formulate a model as a 4-dimensional array data by considering the correlation among the categories.

Chapter 4

AIC for Non-concave Penalized Likelihood Method

Non-concave penalized maximum likelihood methods, such as the Bridge, the SCAD, and the MCP, are widely used because they not only do parameter estimation and variable selection simultaneously but also have a high efficiency as compared to the Lasso. They include a tuning parameter which controls a penalty level, and several information criteria have been developed for selecting it. While these criteria assure the model selection consistency and so have a high value, it is a severe problem that there are no appropriate rules to choose the one from a class of information criteria satisfying such a preferred asymptotic property. In this chapter, we derive an information criterion based on the original definition of the AIC by considering the minimization of the prediction error rather than the model selection consistency. Concretely speaking, we derive a function of the score statistic which is asymptotically equivalent to the non-concave penalized maximum likelihood estimator, and then we provide an asymptotically unbiased estimator of the Kullback-Leibler divergence between the true distribution and the estimated distribution based on the function. Furthermore, through simulation studies, we check that the performance of the proposed information criterion gives almost the same as or better than that of the cross-validation.

4.1 Introduction

The Lasso (Tibshirani 1996) is a regularization method which imposes an ℓ_1 penalty term $\lambda\|\boldsymbol{\beta}\|_1$ on an estimating function with respect to an unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$, where λ (> 0) is a tuning parameter which controls a penalty level. The Lasso can simultaneously do estimation and variable selection by employing the non-differentiability of the penalty term at the origin. Concretely speaking, if $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \hat{\beta}_{\lambda,2}, \ldots, \hat{\beta}_{\lambda,p})^{\mathrm{T}}$ is the estimator based on the Lasso, several of its components will be shrunk to exactly 0 when λ is not close to 0. However, the efficiency of the parameter estimation based on the Lasso is not necessarily large, because the Lasso shrinks the estimator to the zero vector too strongly. To avoid such a problem, it has been proposed to use a penalty term that does not shrink the estimator with the large value. Typical examples of such regularization methods are the Bridge (Frank and Friedman 1993), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001), and the minimax concave penalty (MCP; Zhang 2010). Whereas the Bridge uses an ℓ_q penalty term $(0 < q < 1)$, the SCAD and the MCP use penalty terms which can be approximated by an ℓ_1 penalty term in the neighborhood of the origin and which we hereafter call an ℓ_1-type. Although it is difficult to obtain their estimates since their penalties are non-convex, there are several algorithms which assure the convergence to a local optimal solution such as the coordinate descent method and the gradient descent method.

On the other hand, we have to choose the proper value of the tuning parameter λ which exists in the above regularization methods, and this is an important task for the appropriate model selection. One of the simplest methods for selecting λ is to use the cross-validation (CV; Stone 1974). A stability selection method (Meinshausen and Bühlmann 2010) based on subsampling in order to avoid problems caused by selecting a model based on only one value of λ is attractive, but it requires a considerable number of computational costs like the CV. For an analytical approach without such a problem,

information criteria have been developed rapidly (Yuan and Lin 2007; Wang et al. 2007, 2009; Zhang et al. 2010; Fan and Tang 2013). Letting $\ell(\cdot)$ be the log-likelihood function and $\hat{\boldsymbol{\beta}}_\lambda$ be the estimator of $\boldsymbol{\beta}$ obtained by the above regularization methods, their information criteria take the form of $-2\ell(\hat{\boldsymbol{\beta}}_\lambda) + \kappa_n \|\hat{\boldsymbol{\beta}}_\lambda\|_0$. Then, for some sequence κ_n that depends on at least the sample size n, the model selection consistency is at least assured. For example, the information criterion with $\kappa_n = \log n$ is proposed as the BIC. This approach includes the results for the case in which the dimension of the parameter vector p goes to infinity, and so it has a high value. However, the choice of the tuning parameter remains somewhat arbitrary. That is, there is a class of κ_n assuring a preferred asymptotic property such as the model selection consistency, but there are no appropriate rules to choose the one from the class. For example, since the BIC described above is not derived from Bayes factor, there is no reason to use $\kappa_n = \log n$ instead of $\kappa_n = 2 \log n$. This is a severe problem because data analysts can choose κ_n arbitrarily and do model selection as they want.

The information criterion without such a problem about the arbitrariness is proposed by Efron et al. (2004) or Zou et al. (2007) in a Gaussian linear regression setting or by Ninomiya and Kawano (2014) in a generalized linear regression setting. Concretely speaking, based on the original definition of the C_p or the AIC, an unbiased estimator of the mean squared error or an asymptotically unbiased estimator of a Kullback-Leibler divergence is derived. However, these information criteria are basically only for the Lasso. In addition, the asymptotic setting used in Ninomiya and Kawano (2014) does not assure even the estimation consistency.

Our goal in this chapter is to derive an information criterion based on the original definition of the AIC under an asymptotic setting that assures the estimation consistency for the regularization methods using a non-concave penalty including the Bridge, the SCAD, and the MCP. To achieve it, the result in Hjort and Pollard (1993) is slightly extended to derive an asymptotic property for the estimator. Then, for the Kullback-

Leibler divergence, we construct an asymptotically unbiased estimator by evaluating the asymptotic bias between the divergence and the log-likelihood in which the estimator is plugged.

The rest of the chapter is as follows. In Section 4.2, the generalized linear model and the regularization method we treat are introduced, and then some assumptions for our asymptotic theory are described. In Section 4.3, we discuss an asymptotic property of the estimator obtained from the regularization method, and in Section 4.4, we use the asymptotic property to evaluate an asymptotic bias, which is needed to derive the AIC. The validity of the proposed information criterion is demonstrated for several models through some simulation studies in Section 4.5, and some concluding remarks and future works are presented in Section 4.6. Several proofs are relegated to the appendix.

4.2 Setting and assumptions for asymptotics

Let us consider a natural exponential family with a natural parameter $\boldsymbol{\theta}$ in Θ ($\subset \mathbb{R}^r$) for an r-dimensional random variable $\boldsymbol{y}$, whose density is

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\boldsymbol{y})\right\}$$

with respect to a σ-finite measure. We assume that Θ is the natural parameter space, that is, $\boldsymbol{\theta}$ in Θ satisfies $0 < \int \exp\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\theta} + b(\boldsymbol{y})\}d\boldsymbol{y} < \infty$. Then all the derivatives of $a(\boldsymbol{\theta})$ and all the moments of $\boldsymbol{y}$ exist in the interior Θ^{int} of Θ, and, in particular, $\mathrm{E}[\boldsymbol{y}] = a'(\boldsymbol{\theta})$ and $\mathrm{V}[\boldsymbol{y}] = a''(\boldsymbol{\theta})$. For a function $c(\boldsymbol{\eta})$, we will denote $\partial c(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and $\partial^2 c(\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}$ by $c'(\boldsymbol{\eta})$ and $c''(\boldsymbol{\eta})$, respectively. We also assume that $\mathrm{V}[\boldsymbol{y}] = a''(\boldsymbol{\theta})$ is positive definite and so $-\log f(\boldsymbol{y};\boldsymbol{\theta})$ is a strictly convex function with respect to $\boldsymbol{\theta}$.

Let $(\boldsymbol{y}_i, \boldsymbol{X}_i)$ be the i-th set of responses and regressors ($i = 1, 2, \ldots, n$); we assume that $\boldsymbol{y}_i$ are independent r-dimensional random vectors and $\boldsymbol{X}_i$ in $\mathcal{X}$ ($\subset \mathbb{R}^{r\times p}$) are $(r \times p)$-matrices of known constants. We will consider generalized linear models with natural link functions for such data (see McCullagh and Nelder 1983), that is, we consider a class of

density functions $\{f(\boldsymbol{y}; \boldsymbol{X}\boldsymbol{\beta});\ \boldsymbol{\beta} \in \mathcal{B}\}$ for $\boldsymbol{y}_i$ and so the log-likelihood function of $\boldsymbol{y}_i$ is given by

$$g_i(\boldsymbol{\beta}) = \boldsymbol{y}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{\beta} - a(\boldsymbol{X}_i \boldsymbol{\beta}) + b(\boldsymbol{y}_i),$$

where $\boldsymbol{\beta}$ is a p-dimensional coefficient vector and $\mathcal{B}$ ($\subset \mathbb{R}^p$) is an open convex set. To develop an asymptotic theory for this model, we assume two conditions about the behavior of $\{\boldsymbol{X}_i\}$ as follows:

(C1) $\mathcal{X}$ is a compact set with $\boldsymbol{X}\boldsymbol{\beta} \in \Theta^{\mathrm{int}}$ for all $\boldsymbol{X} \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathcal{B}$.

(C2) There exists an invariant distribution μ on $\mathcal{X}$. In particular, $n^{-1} \sum_{i=1}^n \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{X}_i$ converges to a positive definite matrix $\boldsymbol{J}(\boldsymbol{\beta}) \equiv \int_{\mathcal{X}} \boldsymbol{X}^{\mathrm{T}} a''(\boldsymbol{X}\boldsymbol{\beta}) \boldsymbol{X} \mu(d\boldsymbol{X})$.

Under the above setting, we have the following lemma.

Lemma 1. Let $\boldsymbol{\beta}^*$ be the true value of $\boldsymbol{\beta}$. Then under the conditions (C1) and (C2), we obtain the following expressions:

(R1) There exists a convex and differentiable function $h(\boldsymbol{\beta})$ such that $n^{-1} \sum_{i=1}^n \{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} \xrightarrow{\mathrm{p}} h(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$.

(R2) $\boldsymbol{J}_n(\boldsymbol{\beta}) \equiv -n^{-1} \sum_{i=1}^n g_i''(\boldsymbol{\beta})$ converges to $\boldsymbol{J}(\boldsymbol{\beta})$.

(R3) $\boldsymbol{s}_n \equiv n^{-1/2} \sum_{i=1}^n g_i'(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} \boldsymbol{s} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\beta}^*))$.

Note that $h(\boldsymbol{\beta})$ can be written as

$$\int [a'(\boldsymbol{X}\boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - \{a(\boldsymbol{X}\boldsymbol{\beta}^*) - a(\boldsymbol{X}\boldsymbol{\beta})\}] \mu(d\boldsymbol{X})$$

explicitly since we assume (C2), and so we can prove (R1) without using the technique in convex analysis (Rockafellar 1970). (R2) is a direct consequence of (C2) since $g_i''(\boldsymbol{\beta}) = -\boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{X}_i$. The proof of (R3) is in Section 6.2.3.

Let us consider a non-concave penalized maximum likelihood estimator

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n^{1/2} \sum_{j=1}^{p} p_\lambda(\beta_j) \right\}, \tag{4.1}$$

where λ (> 0) is a tuning parameter and $p_\lambda(\beta_j)$ is a penalty term with respect to β_j, which is not necessarily convex. Let us assume that $p_\lambda(\cdot)$ satisfies the following conditions, and hereafter we call it an ℓ_q-type:

(P1) $p_\lambda(\beta)$ is not differentiable only at the origin, symmetry with respect to $\beta = 0$, and monotone non-decreasing with respect to $|\beta|$.

(P2) $\lim_{\beta\to 0} p_\lambda(\beta)/|\beta|^q = \lambda$ for any $0 < q \leq 1$.

Such a penalty term for the Bridge, the SCAD, and the MCP are given by

$$p_\lambda^{\text{Bridge}}(\beta) = \lambda|\beta|^q,$$
$$p_\lambda^{\text{SCAD}}(\beta) = \lambda|\beta|1_{\{|\beta|\leq(r+1)\lambda\}} - (|\beta|-\lambda)^2/(2r)1_{\{\lambda<|\beta|\leq(r+1)\lambda\}} + \lambda^2(1+r/2)1_{\{|\beta|>(r+1)\lambda\}},$$

and

$$p_\lambda^{\text{MCP}}(\beta) = r\lambda^2/2 - (r\lambda - |\beta|)^2/(2r)1_{\{|\beta|\leq r\lambda\}},$$

respectively, where $0 < q \leq 1$ and $r > 1$. The Bridge penalty is the Lasso penalty itself when $q = 1$, and has a property that the derivative at the origin diverges when $0 < q < 1$. For the SCAD and the MCP penalties, the condition (P2) about the behavior at the neighborhood of the origin is satisfied by setting $q = 1$ just like the Lasso penalty. Thus, it is easy to imagine that a lot of penalties including these three kinds satisfy these conditions. Note that by using such penalties, several components of $\hat{\boldsymbol{\beta}}_\lambda$ tend to exactly 0 because of the non-differentiability at the origin. Also note that $p_\lambda(\cdot)$ is assumed not to depend on a subscript j of the parameter for simplicity, but this is not essential. While Ninomiya and Kawano (2014) put n on the penalty term, we put $n^{1/2}$ on it in this study. From this, we can prove the estimation consistency, which is described in the next section.

Moreover, we can prove the weak convergence of $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$ although the asymptotic distribution is not normal in general.

4.3 Asymptotic behavior

4.3.1 Preparation

Although the objective function in (4.1) is no longer convex because of the non-convexity of $p_\lambda(\cdot)$, the consistency of $\hat{\boldsymbol{\beta}}_\lambda$ can be derived by a similar argument to in Knight and Fu (2000). First, it holds the following lemma.

Lemma 2. $\hat{\boldsymbol{\beta}}_\lambda$ is a consistent estimator of $\boldsymbol{\beta}^*$, that is, $\hat{\boldsymbol{\beta}}_\lambda \xrightarrow{\mathrm{p}} \boldsymbol{\beta}^*$ under the conditions (C1), (C2), (P1), and (P2).

The proof of this lemma is in Section 6.2.3. Hereafter, we denote $\boldsymbol{J}(\boldsymbol{\beta}^*)$ by $\boldsymbol{J}$ as long as there is no confusion. In addition, we denote $\{j;\ \beta_j^* = 0\}$ and $\{j;\ \beta_j^* \neq 0\}$ by $\mathcal{J}^{(1)}$ and $\mathcal{J}^{(2)}$, respectively. Moreover, the vector $(u_j)_{j\in\mathcal{J}^{(k)}}$ and the matrix $(\boldsymbol{J}_{ij})_{i\in\mathcal{J}^{(k)},j\in\mathcal{J}^{(l)}}$ are denoted by $\boldsymbol{u}^{(k)}$ and $\boldsymbol{J}^{(kl)}$, respectively, and we sometimes express, for example, $\boldsymbol{u}$ by $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$.

To develop the asymptotic property of the penalized maximum likelihood estimator in (4.1), which will be used to derive an information criterion, we need to make a small generalization of the result in Hjort and Pollard (1993), as follows:

Lemma 3. Suppose that $\eta_n(\boldsymbol{u})$ is a strictly convex random function that is approximated by $\tilde{\eta}_n(\boldsymbol{u})$. Let $\boldsymbol{u}^\dagger$ be a subvector of $\boldsymbol{u}$, and let $\phi(\boldsymbol{u})$ and $\psi(\boldsymbol{u}^\dagger)$ be continuous functions such that $\phi_n(\boldsymbol{u})$ and $\psi_n(\boldsymbol{u}^\dagger)$ converge to $\phi(\boldsymbol{u})$ and $\psi(\boldsymbol{u}^\dagger)$ uniformly over $\boldsymbol{u}$ and $\boldsymbol{u}^\dagger$ in any compact set, respectively, and assume that $\phi(\boldsymbol{u})$ is convex and $\psi(\boldsymbol{0}) = 0$. In addition, for

$$\nu_n(\boldsymbol{u}) = \eta_n(\boldsymbol{u}) + \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger) \quad \text{and} \quad \tilde{\nu}_n(\boldsymbol{u}) = \tilde{\eta}_n(\boldsymbol{u}) + \phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger),$$

let $\boldsymbol{u}_n$ and $\tilde{\boldsymbol{u}}_n$ be the argmin of $\nu_n(\boldsymbol{u})$ and $\tilde{\nu}_n(\boldsymbol{u})$, respectively, and assume that $\tilde{\boldsymbol{u}}_n$ is unique and $\tilde{\boldsymbol{u}}_n^\dagger = \boldsymbol{0}$. Then, for any $\varepsilon\ (>0)$, $\delta\ (>0)$ and $\xi\ (>\delta)$, there exists $\gamma\ (>0)$ such

that

$$\mathrm{P}(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \delta) \leq \mathrm{P}\left(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)\right) + \mathrm{P}(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \xi) + \mathrm{P}(|\boldsymbol{u}_n^\dagger| > \gamma), \quad (4.2)$$

where

$$\Delta_n(\delta) = \sup_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|\leq\delta} |\nu_n(\boldsymbol{u}) - \tilde{\nu}_n(\boldsymbol{u})| \quad \text{and} \quad \Upsilon_n(\delta) = \inf_{|\boldsymbol{u}-\tilde{\boldsymbol{u}}_n|=\delta} \tilde{\nu}_n(\boldsymbol{u}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n). \quad (4.3)$$

Hjort and Pollard (1993) derived an inequality $\mathrm{P}(|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \geq \delta) \leq \mathrm{P}\left(2\Delta_n(\delta) \geq \Upsilon_n(\delta)\right)$; they assumed that $\nu_n(\boldsymbol{u})$ is convex. Although $\phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ is non-convex (hence $\nu_n(\boldsymbol{u})$ is too), we will use the fact that $\phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ converge to $\phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$ over $\mathcal{U} \equiv \{\boldsymbol{u};\ |\boldsymbol{u}^\dagger| \leq \gamma,\ \delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi\}$. In fact, if n is sufficiently large, the inequality satisfied by the convex function is approximately satisfied for $\phi_n(\boldsymbol{u})$; that is, we have

$$(1 - \delta/l)\,\phi_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\phi_n(\boldsymbol{u}) - \phi_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) > -\varepsilon/2 \quad (4.4)$$

in $\mathcal{U}$. Here, $\boldsymbol{w}$ is a unit vector such that $\boldsymbol{u} = \tilde{\boldsymbol{u}}_n + l\boldsymbol{w}$, and l is in $[\delta, \xi]$, since $\delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi$. Moreover, if γ is sufficiently small and n is sufficiently large, since $\psi(\tilde{\boldsymbol{u}}_n^\dagger) = 0$, we have

$$(1 - \delta/l)\,\psi_n(\tilde{\boldsymbol{u}}_n^\dagger) + (\delta/l)\psi_n(\boldsymbol{u}^\dagger) - \psi_n(\tilde{\boldsymbol{u}}_n^\dagger + \delta\boldsymbol{w}^\dagger) > -\varepsilon/2 \quad (4.5)$$

in $\mathcal{U}$. Hence, we can show that

$$\mathrm{P}(|\boldsymbol{u}_n^\dagger| \leq \gamma,\ \delta \leq |\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \leq \xi) \leq \mathrm{P}(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)) \quad (4.6)$$

in the same way as in Hjort and Pollard (1993), from which we obtain the above lemma. See Section 6.2.3 for the details.

4.3.2 Limiting distribution

We use Lemma 3 to derive the asymptotic property of the penalized maximum likelihood estimator in (4.1). Because the asymptotic property depends on the value of q, we will develop our argument by setting $0 < q < 1$. Furthermore, we will use $\tilde{q} = 1/(2q)$ for the sake of simplicity.

Let us define a strictly convex random function,

$$\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \sum_{i=1}^{n} \left\{ g_i(\boldsymbol{\beta}^{*(1)}, \boldsymbol{\beta}^{*(2)}) - g_i(n^{-\tilde{q}}\boldsymbol{u}^{(1)}, n^{-1/2}\boldsymbol{u}^{(2)} + \boldsymbol{\beta}^{*(2)}) \right\} \tag{4.7}$$

and

$$\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = -\boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{s}_n^{(2)} + \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}\boldsymbol{u}^{(2)}/2, \tag{4.8}$$

where $\boldsymbol{s}_n^{(2)} = n^{-1/2}\sum_{i=1}^{n} g'^{(2)}(\boldsymbol{\beta}^*)$. Using the Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{0})$, $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ can be expressed as

$$-\sum_{i=1}^{n}\{n^{-\tilde{q}}\boldsymbol{u}^{(1)\mathrm{T}}g_i'^{(1)}(\boldsymbol{\beta}^*) + n^{-1/2}\boldsymbol{u}^{(2)\mathrm{T}}g_i'^{(2)}(\boldsymbol{\beta}^*)\}$$
$$-\sum_{i=1}^{n}\{n^{-2\tilde{q}}\boldsymbol{u}^{(1)\mathrm{T}}g_i''^{(11)}(\boldsymbol{\beta}^*)\boldsymbol{u}^{(1)}/2 + n^{-\tilde{q}-1/2}\boldsymbol{u}^{(1)\mathrm{T}}g_i''^{(12)}(\boldsymbol{\beta}^*)\boldsymbol{u}^{(2)} + n^{-1}\boldsymbol{u}^{(2)\mathrm{T}}g_i''^{(22)}(\boldsymbol{\beta}^*)\boldsymbol{u}^{(2)}/2\}$$

plus $\mathrm{o_p}(1)$. Note that the term $-n^{-1}\sum_{i=1}^{n}\boldsymbol{u}^{(2)\mathrm{T}}g_i''^{(22)}(\boldsymbol{\beta}^*)\boldsymbol{u}^{(2)}$ converges to $\boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{J}\boldsymbol{u}^{(2)}$ from (R2), and the terms including $\boldsymbol{u}^{(1)}$ reduce to $\mathrm{o_p}(1)$. Then we see that $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ is asymptotically equivalent to $\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$. Next, letting $\boldsymbol{u}^{\dagger}$ be $\boldsymbol{u}^{(1)}$ and letting

$$\phi_n(\boldsymbol{u}) = n^{1/2}\sum_{j\in\mathcal{J}^{(2)}}\left\{p_\lambda\left(\frac{u_j}{n^{1/2}} + \beta_j^*\right) - p_\lambda(\beta_j^*)\right\} \tag{4.9}$$

and

$$\psi_n(\boldsymbol{u}^{\dagger}) = n^{1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda\left(\frac{u_j}{n^{\tilde{q}}}\right), \tag{4.10}$$

we see from (P2) that $\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ converges to a convex function

$$\phi(\boldsymbol{u}) = \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{p}_\lambda'^{(2)} \quad \text{and} \quad \psi(\boldsymbol{u}^{\dagger}) = \lambda\|\boldsymbol{u}^{(1)}\|_q^q, \tag{4.11}$$

uniformly over $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ in a compact set, respectively, where $\boldsymbol{p}_\lambda'^{(2)} = (p_\lambda'(\beta_j^*))_{j\in\mathcal{J}^{(2)}}$. In addition, letting $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^{\dagger})$ and $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^{\dagger})$, we see that the argmins of $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ and $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ are given by

$$(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}) = (n^{\tilde{q}}\hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})) \quad \text{and} \quad (\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) = (\boldsymbol{0}, \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})),$$

respectively. Using Lemma 3 together with the above preliminaries, we find that, for any $\varepsilon\ (>0)$, $\delta\ (>0)$ and $\xi\ (>\delta)$, there exists $\gamma\ (>0)$ such that

$$\begin{aligned}
&\mathrm{P}(|(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| \geq \delta) \\
&\leq \mathrm{P}(2\Delta_n(\delta) + \varepsilon \geq \Upsilon_n(\delta)) + \mathrm{P}(|(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})| \geq \xi) + \mathrm{P}(|\boldsymbol{u}_n^{(1)}| > \gamma), \qquad (4.12)
\end{aligned}$$

where $\Delta_n(\delta)$ and $\Upsilon_n(\delta)$ are the functions defined in (4.3). The triangle inequality, the convexity of $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{s}_n^{(2)}$ and the uniformly convergence of $\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ imply

$$\begin{aligned}
\Delta_n(\delta) \leq & \sup_{\|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})\| \leq \delta} |\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{s}_n^{(2)} - \boldsymbol{u}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)}\boldsymbol{u}^{(2)}/2| \\
& + \sup_{\|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})\| \leq \delta} |\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \phi(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})| \\
& \xrightarrow{\mathrm{p}} 0. \qquad (4.13)
\end{aligned}$$

Let $\rho\ (>0)$ be half the smallest eigenvalue of $\boldsymbol{J}^{(22)}$. Then, we have

$$\Upsilon_n(\delta) = \inf_{\|(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})\| = \delta} \left\{\lambda\|\boldsymbol{u}^{(1)}\|_q^q + (\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{u}^{(2)} - \tilde{\boldsymbol{u}}_n^{(2)})/2\right\} \geq \min\{\lambda\delta^q, \rho\delta^2\} \qquad (4.14)$$

by a simple calculation. From (4.13) and (4.14), the first term on the right-hand side in (4.12) is shown to converge to 0. In addition, we can generalize the result in Radchenko (2005) with respect to the model and the penalty term; thus, for any $\gamma\ (>0)$, we have

$$\mathrm{P}(|\boldsymbol{u}_n^{(1)}| \leq \gamma) \to 1 \quad \text{and} \quad |\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| = \mathrm{O_p}(1). \qquad (4.15)$$

See Section 6.2.3 for the proof of (4.15). From this, by considering a sufficiently large ξ and a sufficiently large n, the second and third terms on the right-hand side in (4.12) can be made arbitrarily small. Thus we conclude that

$$\boldsymbol{u}_n^{(1)} = \mathrm{o_p}(1) \quad \text{and} \quad \boldsymbol{u}_n^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + \mathrm{o_p}(1).$$

Theorem 2. Let $\boldsymbol{p}_\lambda'^{(2)} = (p_\lambda'(\beta_j^*))_{j \in \mathcal{J}^{(2)}}$, $\boldsymbol{J}^{(1|2)} = \boldsymbol{J}^{(11)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}$, $\boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})$ and

$$\hat{\boldsymbol{u}}_n^{(1)} = \operatorname*{argmin}_{\boldsymbol{u}^{(1)}} \left\{\boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\|\boldsymbol{u}^{(1)}\|_1\right\}. \qquad (4.16)$$

Under the conditions (C1), (C2), (P1), and (P2), we have

$$n^{1/(2q)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathrm{o_p}(1) \quad \text{and} \quad n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}) + \mathrm{o_p}(1)$$

when $0 < q < 1$, and we have

$$n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + \mathrm{o_p}(1) \tag{4.17}$$

and

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = -\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)} + \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}) + \mathrm{o_p}(1) \tag{4.18}$$

when $q = 1$.

In the case of $q = 1$, we can also obtain the result by almost the same way as in the case of $0 < q < 1$ (see Section 6.2.3 for the detail). From Theorem 2, the estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (4.1) is shown to converge in distribution to some function of a Gaussian distributed random variable. When $0 < q < 1$, we can immediately see that it is 0 or the Gaussian distributed random variable itself, and this simple fact is useful to derive an information criterion explicitly and reduce the computational costs for model selection. On the other hand, when $q = 1$, we can show the weak convergence since the convex objective function in (4.16) converges uniformly from the convexity lemma in Hjort and Pollard (1993).

Corollary 1. Let $\boldsymbol{s}^{(2)}$ be a Gaussian distributed random variable with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{J}^{(22)}$ and

$$\hat{\boldsymbol{u}}^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\mathrm{argmin}} \left\{ \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{\tau}_\lambda(\boldsymbol{s}) + \lambda\|\boldsymbol{u}^{(1)}\|_1 \right\}. \tag{4.19}$$

Then, under the same conditions as in Theorem 2, we have

$$n^{1/(2q)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} \xrightarrow{\mathrm{d}} 0 \quad \text{and} \quad n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) \xrightarrow{\mathrm{d}} \boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})$$

when $0 < q < 1$, and we have

$$n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} \xrightarrow{\mathrm{d}} \hat{\boldsymbol{u}}^{(1)} \quad \text{and} \quad n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) \xrightarrow{\mathrm{d}} -\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}^{(1)} + \boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})$$

when $q = 1$.

In the case of $q = 1$, we need to solve the minimization problem in (4.19) for evaluating the AIC later, but it is easy because the objective function is convex with respect to $\boldsymbol{u}^{(1)}$ and so we can use the existing convex optimization techniques. It is well-known that the proximal gradient method (Rockafellar 1976; Beck and Teboulle 2009) is effective for such a minimization problem when the objective function is the sum of a differentiable function and a non-differentiable function. We use, however, the coordinate descent method (Mazumder et al. 2011) because the objective function can be minimized explicitly for each variable. Actually, when we fix all the elements of $\hat{\boldsymbol{u}}$ except for the j-th one, $\hat{u}_j^{(1)}$ is given by

$$\hat{u}_j^{(1)} = \frac{1}{\boldsymbol{J}_{jj}^{(1|2)}} \mathrm{sgn}\left(\tau_j - \sum_{k \neq j} \boldsymbol{J}_{jk}^{(1|2)} \hat{u}_k^{(1)}\right) \max\left\{\left|\tau_j - \sum_{k \neq j} \boldsymbol{J}_{jk}^{(1|2)} \hat{u}_k^{(1)}\right| - \lambda, 0\right\}.$$

Then, for $(t+1)$-th steps in the algorithm, we have only to update the value of $u_j^{(t)}$ as

$$u_j^{(t+1)} = \underset{u}{\mathrm{argmin}}\, h(u_1^{(t+1)}, u_2^{(t+1)}, \ldots, u_{j-1}^{(t+1)}, u, u_{j+1}^{(t)}, u_{j+2}^{(t)}, \ldots, u_{|\mathcal{J}^{(1)}|}^{(t)}),$$

for $j = 1, 2, \ldots, |\mathcal{J}^{(1)}|$, and we repeat this until $\|\boldsymbol{u}^{(t+1)} - \boldsymbol{u}^{(t)}\|$ converges. Note that the optimal value $\hat{u}_j^{(1)}$ satisfies that $\hat{u}_j^{(1)} = 0$ if $|(\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}} + \boldsymbol{\tau}_\lambda(\boldsymbol{s}))_j| \leq \lambda$ and $(\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}} + \boldsymbol{\tau}_\lambda(\boldsymbol{s}))_j = -\lambda \mathrm{sgn}(\hat{u}_j^{(1)})$ otherwise.

4.4 Information criterion

Model selection by the AIC is approached by trying to reduce twice the Kullback-Leibler divergence (Kullback and Leibler 1951) between the true distribution and the estimated distribution,

$$2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\boldsymbol{\beta}^*)\right] - 2\tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]$$

in the prediction perspective, where $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$ is a copy of $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$, in other words, $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$ has the same distribution as $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ and is independent of $(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$. In addition, $\tilde{g}_i(\boldsymbol{\beta})$ and $\tilde{\mathrm{E}}$ denote a log-likelihood function based on $\tilde{\boldsymbol{y}}_i$, that

is, $\log f(\tilde{\boldsymbol{y}}_i; \boldsymbol{X}_i\boldsymbol{\beta})$, and the expectation with respect to only $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$, respectively. Because the first term is a constant, i.e., does not depend on the model selection, we need to consider only the second term, and then the AIC is defined as an asymptotically biased estimator for it (Akaike 1973). In our setting, a simple estimator of the second term is $-2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda)$, but it is known that this underestimates the second term. We then consider minimizing the bias correction,

$$-2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2\mathrm{E}\left[\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]\right], \tag{4.20}$$

in AIC-type information criteria (see Konishi and Kitagawa 2008). Because the expectation in (4.20), the bias term, depends on the true distribution, it cannot be given explicitly in general, and so we evaluate it asymptotically by the same way as was done for the AIC.

For the Lasso, Efron et al. (2004) and Zou et al. (2007) developed the C_p-type information criterion as an unbiased estimator of the prediction squared error in a Gaussian linear regression setting, in other word, a finite correction of the AIC (Sugiura 1978) in a Gaussian linear setting with a known variance. For the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,p})$, it can be expressed as

$$\sum_{i=1}^{n}\{(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_\lambda)^{\mathrm{T}}\mathrm{V}[\boldsymbol{y}_i]^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_\lambda) + \log|2\pi\mathrm{V}[\boldsymbol{y}_i]|\} + 2|\{j;\ \hat{\beta}_{\lambda,j} \neq 0\}|,$$

where the index set $\{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ is called an active set. Unfortunately, since Stein's unbiased risk estimation theory (Stein 1981) is used for deriving this criterion, it was difficult to extend the result to other models. In that situation, Ninomiya and Kawano (2014) relied on the statistical asymptotic theory and extended the result to generalized linear models based on the asymptotic distribution of the Lasso estimator. The Lasso estimator in their paper is defined by

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\operatorname{argmin}}\left\{-\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n\lambda\|\boldsymbol{\beta}\|_1\right\},$$

but as was mentioned in the previous Section, the estimation consistency is not assured because the order of the penalty term is $\mathrm{O}(n)$. In this study, under the setting the

estimation consistency holds as was shown in Lemma 2, we derive an information criterion for not only the Lasso but also the non-concave penalized likelihood method.

The bias term in (4.20) can be rewritten as the expectation of

$$\sum_{i=1}^{n}\{g_i(\hat{\boldsymbol{\beta}}_\lambda) - g_i(\boldsymbol{\beta}^*)\} - \sum_{i=1}^{n}\{\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{g}_i(\boldsymbol{\beta}^*)\}, \tag{4.21}$$

and so we derive an AIC by evaluating $\mathrm{E}[z^{\mathrm{limit}}]$, where z^{limit} is the limit to which (4.21) converges in distribution. We call $\mathrm{E}[z^{\mathrm{limit}}]$ an asymptotic bias. Here we develop an argument by setting $0 < q < 1$.

Using the Taylor's theorem, the first term in (4.21) can be expressed as

$$(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}} \sum_{i=1}^{n} g_i'(\boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}} \sum_{i=1}^{n} g_i''(\boldsymbol{\beta}^\dagger)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2, \tag{4.22}$$

where $\boldsymbol{\beta}^\dagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. Note that $-n^{-1}\sum_{i=1}^{n} g_i''(\boldsymbol{\beta}^\dagger)$ converges in probability to $\boldsymbol{J}$ from (R2) and Lemma 2. Now we apply the result in Theorem 2. First, the terms including $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ reduce to $\mathrm{o_p}(1)$ because $n^{1/(2q)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathrm{o_p}(1)$, and next $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^*)$ is asymptotically equivalent to $\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})$. Then (4.22) can be expressed as

$$\boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)}) - (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda'^{(2)})/2 + \mathrm{o_p}(1),$$

and moreover we see that this converges in distribution to

$$\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}) - (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})/2$$

from (R3). Similarly, the second term in (4.21) can be expressed as

$$(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}} \sum_{i=1}^{n} \tilde{g}_i'(\boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}} \sum_{i=1}^{n} \tilde{g}_i''(\boldsymbol{\beta}^\ddagger)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2 \tag{4.23}$$

using the Taylor's theorem, where $\boldsymbol{\beta}^\ddagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$, and moreover we see that this converges in distribution to

$$\tilde{\boldsymbol{s}}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}) - (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})/2$$

by applying Theorem 2 and (R3), where $\tilde{\boldsymbol{s}}^{(2)}$ is a copy of $\boldsymbol{s}^{(2)}$. Hence we have

$$z^{\text{limit}} = \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}) - \tilde{\boldsymbol{s}}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)}).$$

Because $\boldsymbol{s}^{(2)}$ and $\tilde{\boldsymbol{s}}^{(2)}$ are independently distributed according to $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$, the asymptotic bias reduces to

$$\mathrm{E}[z^{\text{limit}}] = \mathrm{E}[\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda'^{(2)})],$$

and then we obtain the following theorem.

Theorem 3. Under the same conditions as in Theorem 2, we have

$$\mathrm{E}[z^{\text{limit}}] = |\mathcal{J}^{(2)}|$$

when $0 < q < 1$, and we have

$$\mathrm{E}[z^{\text{limit}}] = |\mathcal{J}^{(2)}| + K \tag{4.24}$$

when $q = 1$, where $K = \mathrm{E}\left[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}\right]$, $\boldsymbol{s}^{(1|2)} = \boldsymbol{s}^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}$ and $\hat{\boldsymbol{u}}^{(1)}$ is the random vector defined in (4.19).

In the case of $q = 1$, we can also obtain the result by almost the same way as in the case of $0 < q < 1$ (see Section 6.2.3 for the detail). Because the asymptotic bias derived in Theorem 3 depends on an unknown value $\boldsymbol{\beta}^*$, we need to evaluate it. Here we use that $\hat{\boldsymbol{\beta}}_\lambda$ is a consistent estimator of $\boldsymbol{\beta}^*$ from Lemma 2 and that $\boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda) = n^{-1}\sum_{i=1}^n \boldsymbol{X}^{\mathrm{T}} a''(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{X}$ converges in probability to $\boldsymbol{J}$. Concretely speaking, we replace $\mathcal{J}^{(2)}$ by the active set $\hat{\mathcal{J}}^{(2)} = \{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ and K by its empirical mean $\hat{K}$ obtained by generating samples from $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda))$. As a result, we propose the following index as an AIC for the non-concave penalized maximum likelihood method:

$$\mathrm{AIC}_\lambda^{\ell_q\text{-type}} = \begin{cases} -2\displaystyle\sum_{i=1}^n g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}^{(2)}| & (0 < q < 1) \\ -2\displaystyle\sum_{i=1}^n g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}^{(2)}| + 2\hat{K} & (q = 1) \end{cases}. \tag{4.25}$$

When $0 < q < 1$, we can see that the bias term of the information criterion in Efron et al. (2004) or Zou et al. (2007) can be used not only for Gaussian linear regression settings but also for generalized linear settings. Thus, by minimizing the AIC in (4.25), we can obtain the optimal value of the tuning parameter λ.

4.5 Simulation study

In this Section, we conduct some simulation studies to check the performance of the tuning parameter selection based on the AIC in (4.25). Concretely speaking, we consider a linear regression setting (Linear) and a Logistic regression setting (Logistic), and the performances for the AIC and the CV are compared. As regularization methods, we consider the Bridge ($q = 0.2$), the SCAD, and the MCP.

We assessed the performance in terms of the second term of the Kullback-Leibler divergence:

$$\mathrm{KL} = -\tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})\right], \tag{4.26}$$

where $\hat{\lambda}$ is the value of the tuning parameter given by each of the criteria, and we evaluated the expectation using an empirical mean of 500 samples. We interpret that the criterion giving the small value of the KL is good. Although the original aim of the AIC is to minimize the KL, as a secondary index for the assessment, we also determined the number of false positives and false negatives:

$$\mathrm{FP} = |\{j;\ \hat{\beta}_j \neq 0 \wedge \beta_j^* = 0\}| \quad \text{and} \quad \mathrm{FN} = |\{j;\ \hat{\beta}_j = 0 \wedge \beta_j^* \neq 0\}|, \tag{4.27}$$

for each of the criteria.

The AIC we use is that corresponding to the case $0 < q < 1$ in (4.25) for the Bridge, and that corresponding to the case $q = 1$ in (4.25) for the SCAD and the MCP. Note that for a linear or a logistic regression setting, the log-likelihood function $g_i(\boldsymbol{\beta})$ is expressed as

$$y_i \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{\beta} - y_i^2 \quad \text{or} \quad y_i \boldsymbol{X}_i \boldsymbol{\beta} - \log\{1 + \exp(\boldsymbol{X}_i \boldsymbol{\beta})\},$$

respectively, and, $\boldsymbol{J}_n(\boldsymbol{\beta})$ necessary for evaluating $\hat{K}$ can be expressed as

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i \quad \text{or} \quad n^{-1}\sum_{i=1}^{n}\frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})\}^2}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{X}_i,$$

respectively.

The simulation settings were as follows. As the p-dimensional regressors $\boldsymbol{X}_i$, $(i = 1, 2, \ldots, n)$, we used vectors obtained from the multivariate Gaussian distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the $\boldsymbol{\Sigma}$ is $(p \times p)$-covariance matrix whose (i, j)-th element was set to $0.5^{|i-j|}$. The true coefficient vector $\boldsymbol{\beta}^*$ was

$$\boldsymbol{\beta}^* = (\beta_1^*\mathbf{1}_k^{\mathrm{T}}, \beta_2^*\mathbf{1}_k^{\mathrm{T}}, \mathbf{0}_{p-2k}^{\mathrm{T}})^{\mathrm{T}},$$

where $\mathbf{1}_k$ and $\mathbf{0}_{p-2k}$ denote the k-dimensional one-vector and the $(p-2k)$-dimensional zero-vector, respectively. In addition, (β_1^*, β_2^*) was set to (0.1,0.5) or (0.2,1) for a linear regression setting, and (0.5,1.5) or (1,2) for a logistic regression setting, and seven cases were considered for the three-tuple (p, k, n) as follows: (8,2,50), (8,2,100), (8,2,150), (8,1,100), (8,3,100), (12,3,100), (16,4,100). We used the local quadratic approximation in Fan and Li (2001) for the parameter estimation, and fifty simulations were conducted.

Tables 4.1, 4.2, and 4.3 show the results for the Bridge, the SCAD, and the MCP, respectively, and in each table, the averages and standard deviations of the KL, and the averages of the FP and FN are provided for the linear and the logistic regression settings. In Table 4.1, let us look at the main index KL. While the CV gives a smaller value of the KL than the AIC in about half of cases, the differences between two values are small. On the other hand, in the case where the AIC gives a smaller value of the KL than the CV, the differences tend to be large. Next, let us look at the sub indices FP and FN. In Logistic setting, the values of the FP are almost 0 and the values of the FN are considerably large. That is, we can say that the CV causes an imbalance. As long as there is no special hate for the FP, it will be natural to use the AIC. In Tables 4.2 and 4.3, the AIC and the CV give almost the same values of the KL in Linear setting. On the other hand, in

Logistic setting, the AIC is clearly superior to the CV in many cases. On the whole, we can conclude that the AIC in (4.25) is better than the CV.

4.6 Summary in this chapter

While Ninomiya and Kawano (2014) derived the information criterion for the Lasso in generalized linear models based on the original definition of the AIC which is an asymptotically unbiased estimator of the Kullback-Leibler divergence, they used an asymptotic setting where the estimation consistency is not assured. In addition, the Lasso itself has a problem that the efficiency is not necessarily large because the Lasso shrinks the estimator to the zero vector too strongly. Considering these, we have derived an information criterion for non-concave penalized maximum likelihood methods including the Bridge, the SCAD, and the MCP, which are known to be more efficient than the Lasso, based on the original definition of the AIC under a setting the estimation consistency is assured. The AIC in (4.25) can be regarded as the only criterion for such non-concave penalized maximum likelihood methods that has the same roots as those of the classic information criteria. The bias term of the AIC in (4.25) is determined including its coefficient. Therefore, unlike the argument that gives a class of the information criteria which assures the model selection consistency, we can do the model selection without any arbitrariness.

It has been shown through simulation studies that the performance of the AIC in (4.25) is almost the same as or better than that of the CV. Especially in the Bridge-type regularization method, the AIC is clearly better than the CV also in terms of computational costs because of the simple expression for the AIC. This fact will be significant in handling large-scale data.

Although the number of tuning parameters to be selected is only one, we can extend our result to regularization methods which have several tuning parameters such as the SELO (Dicker et al. 2012). In addition, although we treat the natural link function for our generalized linear models, it is possible to treat different link functions under some regular-

			Case 1			Case 2		
Model	(p,k,n)		KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8,2,50)	CV	0.676 (0.019)	0.30	1.58	0.645 (0.026)	0.30	1.29
		AIC	0.679 (0.018)	0.09	1.77	0.649 (0.022)	0.11	1.55
	(8,2,100)	CV	0.670 (0.016)	0.31	1.31	0.631 (0.018)	0.28	1.05
		AIC	0.672 (0.015)	0.05	1.61	0.634 (0.018)	0.07	1.27
	(8,2,150)	CV	0.666 (0.014)	0.32	1.24	0.632 (0.012)	0.40	0.86
		AIC	0.666 (0.013)	0.10	1.45	0.636 (0.014)	0.04	1.17
	(8,1,100)	CV	0.687 (0.008)	0.46	0.75	0.658 (0.017)	0.75	0.45
		AIC	0.687 (0.009)	0.12	0.81	0.658 (0.016)	0.13	0.54
	(8,3,100)	CV	0.655 (0.014)	0.24	1.86	0.615 (0.020)	0.24	1.40
		AIC	0.659 (0.012)	0.03	2.34	0.626 (0.019)	0.04	2.19
	(12,3,100)	CV	0.662 (0.014)	0.47	1.91	0.617 (0.021)	0.46	1.64
		AIC	0.665 (0.014)	0.15	2.38	0.624 (0.018)	0.06	2.17
	(16,4,100)	CV	0.652 (0.021)	0.41	3.03	0.610 (0.024)	0.69	2.47
		AIC	0.652 (0.017)	0.12	3.28	0.618 (0.021)	0.12	2.98
Logistic	(8,2,50)	CV	0.462 (0.061)	0.01	1.28	0.406 (0.070)	0.04	1.21
		AIC	0.473 (0.153)	0.33	0.69	0.417 (0.129)	0.40	0.40
	(8,2,100)	CV	0.419 (0.044)	0.01	1.04	0.348 (0.047)	0.00	0.92
		AIC	0.398 (0.050)	0.31	0.43	0.307 (0.035)	0.50	0.19
	(8,2,150)	CV	0.394 (0.024)	0.00	0.94	0.307 (0.033)	0.01	0.67
		AIC	0.376 (0.018)	0.43	0.33	0.271 (0.018)	0.41	0.11
	(8,1,100)	CV	0.495 (0.029)	0.00	0.42	0.411 (0.021)	0.00	0.22
		AIC	0.513 (0.033)	0.61	0.21	0.423 (0.035)	0.63	0.02
	(8,3,100)	CV	0.408 (0.047)	0.00	1.92	0.348 (0.053)	0.00	1.74
		AIC	0.346 (0.042)	0.22	0.78	0.272 (0.087)	0.35	0.32
	(12,3,100)	CV	0.384 (0.031)	0.01	1.82	0.376 (0.056)	0.00	1.68
		AIC	0.397 (0.134)	0.75	0.58	0.346 (0.112)	0.73	0.35
	(16,4,100)	CV	0.392 (0.048)	0.01	2.72	0.407 (0.045)	0.00	2.66
		AIC	0.414 (0.122)	1.19	1.05	0.379 (0.137)	1.17	0.60

Table 4.1: Comparison between the CV and the AIC in (4.25) for the Bridge penalty. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is respectively (0.1,0.5) or (0.2,1) in the linear regression setting, and respectively (0.5,1.5) or (1,2) in the logistic regression setting.

			Case 1			Case 2		
Model	(p,k,n)		KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8,2,50)	CV	0.557 (0.050)	0.69	0.49	0.563 (0.039)	0.87	0.20
		AIC	0.566 (0.055)	0.60	0.59	0.582 (0.056)	0.95	0.20
	(8,2,100)	CV	0.521 (0.020)	1.01	0.27	0.518 (0.031)	0.93	0.11
		AIC	0.524 (0.025)	0.92	0.28	0.519 (0.028)	0.91	0.15
	(8,2,150)	CV	0.531 (0.013)	0.76	0.24	0.567 (0.012)	1.05	0.03
		AIC	0.534 (0.015)	0.70	0.26	0.569 (0.013)	0.89	0.03
	(8,1,100)	CV	0.526 (0.021)	1.24	0.19	0.500 (0.020)	1.26	0.06
		AIC	0.526 (0.025)	1.05	0.24	0.503 (0.023)	1.13	0.06
	(8,3,100)	CV	0.491 (0.020)	0.49	0.41	0.555 (0.025)	0.59	0.17
		AIC	0.492 (0.021)	0.43	0.51	0.555 (0.027)	0.48	0.22
	(12,3,100)	CV	0.504 (0.020)	1.16	0.37	0.556 (0.023)	1.33	0.15
		AIC	0.509 (0.028)	1.15	0.38	0.561 (0.026)	1.23	0.16
	(16,4,100)	CV	0.550 (0.030)	1.54	0.66	0.565 (0.029)	1.80	0.15
		AIC	0.557 (0.035)	1.39	0.66	0.573 (0.031)	1.44	0.24
Logistic	(8,2,50)	CV	0.506 (0.032)	0.04	0.82	0.493 (0.023)	0.06	0.59
		AIC	0.477 (0.117)	0.76	0.56	0.511 (0.184)	0.48	0.54
	(8,2,100)	CV	0.476 (0.017)	0.07	0.69	0.426 (0.018)	0.04	0.20
		AIC	0.446 (0.059)	0.78	0.41	0.321 (0.037)	0.52	0.25
	(8,2,150)	CV	0.451 (0.015)	0.05	0.41	0.394 (0.015)	0.06	0.13
		AIC	0.411 (0.021)	1.09	0.18	0.301 (0.025)	0.95	0.08
	(8,1,100)	CV	0.541 (0.017)	0.15	0.14	0.454 (0.024)	0.07	0.06
		AIC	0.542 (0.036)	1.40	0.09	0.406 (0.029)	1.01	0.04
	(8,3,100)	CV	0.431 (0.017)	0.05	1.09	0.423 (0.015)	0.05	0.54
		AIC	0.339 (0.043)	0.38	0.66	0.314 (0.056)	0.19	0.55
	(12,3,100)	CV	0.449 (0.014)	0.03	0.95	0.420 (0.015)	0.03	0.53
		AIC	0.413 (0.093)	1.44	0.46	0.349 (0.086)	0.86	0.59
	(16,4,100)	CV	0.436 (0.013)	0.08	1.50	0.423 (0.018)	0.06	1.19
		AIC	0.438 (0.115)	1.52	0.99	0.356 (0.080)	0.87	1.11

Table 4.2: Comparison between the CV and the AIC in (4.25) for the SCAD penalty. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is respectively (0.1,0.5) or (0.2,1) in the linear regression setting, and respectively (0.5,1.5) or (1,2) in the logistic regression setting.

			Case 1			Case 2		
Model	(p,k,n)		KL (sd)	FP	FN	KL (sd)	FP	FN
Linear	(8,2,50)	CV	0.545 (0.047)	0.82	0.42	0.556 (0.046)	0.79	0.23
		AIC	0.545 (0.047)	0.67	0.49	0.557 (0.046)	0.71	0.29
	(8,2,100)	CV	0.558 (0.020)	0.79	0.38	0.527 (0.023)	0.86	0.13
		AIC	0.560 (0.026)	0.64	0.39	0.530 (0.027)	0.92	0.13
	(8,2,150)	CV	0.520 (0.017)	0.91	0.31	0.518 (0.015)	0.94	0.10
		AIC	0.521 (0.018)	0.71	0.38	0.519 (0.015)	0.84	0.11
	(8,1,100)	CV	0.502 (0.015)	1.02	0.25	0.539 (0.023)	1.03	0.15
		AIC	0.503 (0.018)	0.88	0.27	0.540 (0.024)	0.99	0.14
	(8,3,100)	CV	0.553 (0.021)	0.33	0.53	0.508 (0.028)	0.62	0.10
		AIC	0.556 (0.023)	0.30	0.61	0.510 (0.029)	0.49	0.16
	(12,3,100)	CV	0.523 (0.023)	1.24	0.57	0.578 (0.030)	1.45	0.17
		AIC	0.525 (0.024)	1.02	0.57	0.582 (0.028)	1.39	0.19
	(16,4,100)	CV	0.530 (0.029)	1.72	0.72	0.563 (0.035)	1.73	0.28
		AIC	0.532 (0.031)	1.45	0.72	0.565 (0.036)	1.53	0.34
Logistic	(8,2,50)	CV	0.493 (0.037)	0.04	1.04	0.453 (0.035)	0.06	0.81
		AIC	0.514 (0.159)	0.59	0.59	0.383 (0.090)	0.41	0.52
	(8,2,100)	CV	0.447 (0.023)	0.02	0.65	0.397 (0.025)	0.02	0.47
		AIC	0.418 (0.043)	0.79	0.29	0.323 (0.029)	0.54	0.21
	(8,2,150)	CV	0.423 (0.017)	0.04	0.54	0.367 (0.019)	0.01	0.17
		AIC	0.390 (0.019)	0.88	0.21	0.308 (0.020)	0.94	0.09
	(8,1,100)	CV	0.529 (0.020)	0.10	0.23	0.448 (0.021)	0.13	0.08
		AIC	0.530 (0.036)	0.83	0.17	0.429 (0.027)	1.06	0.06
	(8,3,100)	CV	0.429 (0.020)	0.01	1.09	0.409 (0.031)	0.02	0.97
		AIC	0.362 (0.056)	0.33	0.70	0.312 (0.075)	0.16	0.73
	(12,3,100)	CV	0.423 (0.027)	0.01	1.10	0.401 (0.017)	0.01	0.99
		AIC	0.389 (0.070)	1.02	0.66	0.352 (0.075)	0.91	0.65
	(16,4,100)	CV	0.426 (0.022)	0.02	1.92	0.411 (0.017)	0.02	1.54
		AIC	0.440 (0.136)	1.79	0.94	0.345 (0.107)	1.31	1.02

Table 4.3: Comparison between the CV and the AIC in (4.25) for the MCP penalty. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is respectively (0.1,0.5) or (0.2,1) in the linear regression setting, and respectively (0.5,1.5) or (1,2) in the logistic regression setting.

ity conditions. In this study, we have derived the AIC based on the statistical asymptotic theory for which the dimension of the parameter vector is fixed and the sample size diverges. On the other hand, recently it becomes important to analyze high-dimensional data such that the dimension of the parameter vector is comparable to the sample size. Also for such high-dimensional data, we can expect the AIC-type information criterion works well from the view point of the efficiency. Actually, Zhang et al. (2010) shows that, in the framework where the dimension of the parameter vector is increased along with the sample size, a formal AIC has the asymptotic loss efficiency under a sparse setting with particular conditions. It is an important future work in terms of both theory and practice to show a similar asymptotic property for the proposed information criterion.

Chapter 5

Some Extension of Chapter 4

5.1 Back ground

In the previous chapter, we have considered a constant tuning parameter and have derived an information criterion. However, the asymptotic distribution of that model has an asymptotic bias as in Theorem 2. As was shown in Theorem 3, it is inconvenient to compute the AIC since we need to evaluate the expectation when $q = 1$, i.e., $K = \mathrm{E}[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}]$. Then, we are interested in when K is vanished. One of the answer of this question is that when $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0})$ converges to 1, and this is a part of oracle property in Fan and Li (2001). The original definition of an oracle property consists of a sparsity and an asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_\lambda^{(1)}, \hat{\boldsymbol{\beta}}_\lambda^{(2)})$ obtained from the regularization method, that is,

(sparsity): $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1$ and

(asymptotic normality): $\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{J}^{(22)})$.

Instead of the sparsity, a variable selection consistency, i.e., $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} \neq \mathbf{0})$ converges to 1, is often mentioned in the oracle property.

When $0 < q < 1$, we have proved a sparsity of the estimator by generalizing the result in Radchenko (2005) (see (6.9) in Section 6.2.3), although the asymptotic distribution has an asymptotic bias as in Theorem 2. In addition, a sparsistency is not hold when $q = 1$ since we have not considered an important property of a penalty assuring the sparsity, i.e.,

uniform boundedness. By considering the oracle property of the estimator obtained form the regularization method, it is expected that the information criterion can be represented by a simple form, especially in $q = 1$.

Our main focus in this chapter is to consider an asymptotic property of the estimator obtained from the regularization method especially in the case of ℓ_1-type penalty although we consider only the Bridge for ℓ_q-type penalty. Moreover we will derive an information criterion based on such estimators by the same argument as in Chapter 4. To do this, we assume that the tuning parameter converges to 0 with proper rate as n goes to infinity. Unfortunately, to assure the above mentioned properties, the rate of decay of λ $(= \lambda_n)$ is deferent in $0 < q < 1$ and $q = 1$. In fact, $n^{-1/2}\lambda_n \to 0$ and $n^{-1/2}\lambda_n \to \infty$ are required for ℓ_q-type penalty $(0 < q < 1)$ and ℓ_1-type penalty, respectively. Therefore, we need to divide into two cases, i.e., whether $n^{-1/2}\lambda_n$ converges or not.

5.2 Setting and assumptions for asymptotics

Let us consider the same model as in Chapter 4 and a non-concave penalized maximum likelihood estimator

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\operatorname{argmin}} \left\{ -\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_n}(\beta_j) \right\}, \tag{5.1}$$

where for λ (> 0), $\lambda_n = n^{(q_0-2)/2}\lambda$ is a tuning parameter depending on n and $p_{\lambda_n}(\beta_j)$ is a penalty term with respect to β_j, which is not necessarily convex. We assume that the conditions (C1) and (C2) are still hold. Unlike Chapter 4, we put n on the penalty term and λ depends on the sample size. In addition, we assume that q_0 satisfy $q < q_0 \leq 1$ when $0 < q < 1$ and $1 \leq q_0 < 2$ when $q = 1$. In this case, λ_n converges to 0 regardless the value of p since $q_0 < 2$.

To develop the asymptotic property of the estimator obtained from (5.1), we need to set several conditions for ℓ_1-type penalty:

(Q1) $p_{\lambda_n}(\beta)$ is not differentiable only at the origin, symmetry with respect to $\beta = 0$, and

monotone non-decreasing with respect to $|\beta|$.

(Q2) $\lim_{\beta \to 0} p_{\lambda_n}(\beta)/|\beta| = \lambda_n$.

(Q3) $\lim_{n \to \infty} p_{\lambda_n}(\beta) = 0$ for any β.

(Q4) There exist τ (> 0) such that $p'_{\lambda_n}(\beta) = 0$ for any $|\beta| \geq \tau \lambda_n$.

(Q5) $\lim_{n \to \infty} p''_{\lambda_n}(\beta) = 0$ for any β $(\neq 0)$.

The conditions (Q1) and (Q2) are almost the same conditions as in Chapter 4, and (Q2) imples $p_{\lambda_n}(0) = 0$ and

$$\lim_{\beta \to 0} p'_{\lambda_n}(\beta)/\mathrm{sgn}(\beta) = \lambda_n \tag{5.2}$$

Moreover, an estimation consistency is assured from the condition (Q3). The condition (Q4) guarantees the asymptotic unbiasedness of the estimator in (5.1). Note that these conditions are still hold for the SCAD, and the MCP although the Lasso does not satisfy the condition (Q4) since $d|\beta|/d\beta = \mathrm{sgn}(\beta)$ for any β $(\neq 0)$. Hereafter, we also denote the Bridge penalty, that is, $\lambda_n |\beta|^q$ for $0 < q < 1$, by $p_{\lambda_n}(\beta)$ so long as there is no confusion.

Because the penalty term in (5.1) converges to 0, we can immediately see that the following lemma holds by the same argument of Lemma 2:

Lemma 4. $\hat{\boldsymbol{\beta}}_\lambda$ is a consistent estimator of $\boldsymbol{\beta}^*$ under the conditions (C1), (C2), and (Q1)–(Q3).

5.3 Asymptotic behavior

5.3.1 Preparation

To establish the oracle property of the estimator in (5.1), we first show the sparsity. As mentioned in the above, the sparsity for the Bridge estimator with $q_0 = 1$ has been shown in Chapter 4, and we can show, by the same argument as in the proof of (6.9) in Section

6.2.3, that it still hold when $q < q_0 < 1$ since $n^{(2-q)/2}\lambda_n \to \infty$. Now, we only show the sparsity for ℓ_1-type penalty. Let us define a random function as follows:

$$\mathbb{G}_n(\boldsymbol{u}) = n^{-1}\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta}^* + \boldsymbol{u})\} + \sum_{j=1}^{p}\{p_{\lambda_n}(\beta_j^* + u_j) - p_{\lambda_n}(\beta_j^*)\}$$

Note that the argmin of $\mathbb{G}_n(\boldsymbol{u})$ is given by $\hat{\boldsymbol{u}} = \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*$. By the same argument in the proof of (6.9) in Section 6.2.3, we see that

$$0 \geq \mathbb{G}_n(\hat{\boldsymbol{u}}) \geq n^{-1/2}\boldsymbol{s}_n^{\mathrm{T}}\hat{\boldsymbol{u}} + \hat{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)\hat{\boldsymbol{u}}/2 + \sum_{j\in\mathcal{J}^{(2)}} p'_{\lambda_n}(\beta_j^*)\hat{u}_j\{1 + \mathrm{o_p}(1)\},$$

where $\boldsymbol{\beta}^\dagger$ is a segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. For ℓ_1-type penalty, it follows that the third term on the right hand side in this inequality equals to 0 for sufficiently large n from (Q4), and thus $\hat{\boldsymbol{u}} = \mathrm{O_p}(n^{-1/2})$.

Moreover, by the same argument as in the proof of (6.9) in Section 6.2.3 again, we have

$$\|\hat{\boldsymbol{u}}^{(1)}\|^2 + \lambda_n\|\hat{\boldsymbol{u}}^{(1)}\|_1(1 + \mathrm{o}(1)) \leq \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{u}}^{(1)}\|).$$

Then, this implies that

$$\|n^{1/2}\hat{\boldsymbol{u}}^{(1)}\|^2 + n^{(q_0-1)/2}\lambda\|n^{1/2}\hat{\boldsymbol{u}}^{(1)}\|_1(1 + \mathrm{o}(1)) \leq \mathrm{O_p}(\|n^{1/2}\hat{\boldsymbol{u}}^{(1)}\|),$$

and thus the sparsity for the ℓ_1-type regularized estimator with $1 < q_0 < 2$ follows from $n^{1/2}\hat{\boldsymbol{u}}^{(1)} = \mathrm{O_p}(1)$ and $n^{(q_0-1)/2} \to \infty$. Note that the sparsity is not ensured when $q_0 = 1$ because the above inequality does not contradict.

Lemma 5 (Sparsity)**.** Under the conditions (C1), (C2), the Bridge estimator has the sparsity. Moreover, the ℓ_1-type regularized estimator has the sparsity only if $1 < q_0 < 2$ under the conditions (Q1)–(Q4).

5.3.2 Asymptotic distribution

By using Lemma 5, we can derive an asymptotic distribution of the estimator defined in (5.1). In the case of $q_0 = 1$, the result is same as in Theorem 3 as shown bellow, and

thus we only show the case of $q_0 \neq 1$ here. Because we can use Lemma 5, the asymptotic distribution of the estimator can be established somewhat easier than Chapter 4, that is, we do not use Lemma 3 in this chapter except for the case of the ℓ_1-type regularized estimator with $q_0 = 1$.

Asymptotic distribution of the Bridge estimator

Let us denote the objective function in (5.1) by $\mathbb{H}_n(\boldsymbol{\beta})$, that is,

$$\mathbb{H}_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) + n\sum_{j=1}^{n} p_{\lambda_n}(\beta_j). \tag{5.3}$$

Note that $p_{\lambda_n}(\beta_j) = \lambda_n|\beta|^q$ in this case. From Lemma 4 and 5, we see that $\hat{\boldsymbol{\beta}}_\lambda$ is $n^{1/2}$-consistent and that $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ stays away from 0 for sufficiently large n. Thus, it satisfies the likelihood equation:

$$\left.\frac{\partial \mathbb{H}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = -\sum_{i=1}^{n} g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda) + n\lambda_n \boldsymbol{p}'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \mathbf{0} \tag{5.4}$$

with probability converging to 1, where $\boldsymbol{p}'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = (q\mathrm{sgn}(\hat{\beta}_{\lambda,j})|\hat{\beta}_{\lambda,j}|^{q-1})_{j\in\mathcal{J}^{(2)}}$. By using Lemma 4 again, we see that $\boldsymbol{p}'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \boldsymbol{p}'^{(2)} + \mathrm{o_p}(1)$, where $\boldsymbol{p}'^{(2)} = \boldsymbol{p}'(\boldsymbol{\beta}^{*(2)})$. Moreover, from the Taylor' theorem, we have

$$g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda) = g'^{(2)}(\boldsymbol{\beta}^*) + g_i''^{(21)}(\boldsymbol{\beta}^*)\hat{\boldsymbol{\beta}}_\lambda^{(1)} + g_i''^{(22)}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + \mathrm{o_p}(1), \tag{5.5}$$

and thus (5.4) reduces to

$$-n^{1/2}\boldsymbol{s}_n^{(2)} + n\boldsymbol{J}_n^{(21)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} + n\boldsymbol{J}_n^{(22)}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + \mathrm{o_p}(1) + n\lambda_n(\boldsymbol{p}'^{(2)} + \mathrm{o_p}(1)) = \mathbf{0}.$$

Hence, (R2), Lemma 5 and $n^{1/2}\lambda_n = \mathrm{o}(1)$ imply

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1).$$

As in the case of $q_0 = 1$, the first term of the right-hand side in the above equation become $\boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \lambda\boldsymbol{p}'^{(2)})$ because $n^{1/2}\lambda_n = \lambda$ and this is the same result as in Theorem 3.

Theorem 4 (Asymptotic distribution of the Bridge estimator)**.** Under the conditions (C1) and (C2), we have $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1$. Moreover, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$$

when $q < q_0 < 1$, and we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \lambda \boldsymbol{p}'^{(2)}) + \mathrm{o_p}(1)$$

when $q_0 = 1$.

Asymptotic distribution of the ℓ_1-type regularized estimator

Let us consider (5.3) for ℓ_1-type penalty.

$$\left.\frac{\partial \mathbb{H}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = -\sum_{i=1}^{n} g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda) + n\boldsymbol{p}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \mathbf{0} \tag{5.6}$$

with probability converging to 1, where $\boldsymbol{p}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = (p_{\lambda_n}'(\hat{\beta}_{\lambda,j}))_{j\in\mathcal{J}^{(2)}}$. Using the Taylor' theorem, we have

$$\boldsymbol{p}_n'(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \boldsymbol{p}_n'(\boldsymbol{\beta}^{*(2)}) + p_n''(\boldsymbol{\beta}^\ddagger)(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}),$$

where $\boldsymbol{\beta}^\ddagger$ is a segment from $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ to $\boldsymbol{\beta}^{*(2)}$. Because $\hat{\boldsymbol{\beta}}_\lambda$ is $n^{1/2}$-consistent, we see that $\boldsymbol{p}_n'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda^{(2)}) = \mathrm{o_p}(n^{-1/2})$ for sufficiently large n from (Q4) and (Q5). Moreover, $g_i'^{(2)}(\hat{\boldsymbol{\beta}}_\lambda)$ can be expressed as (5.5) by the Taylor expansion, and thus (5.6) reduces to

$$-n^{1/2}\boldsymbol{s}_n^{(2)} + n\boldsymbol{J}_n^{(21)}\hat{\boldsymbol{\beta}}_\lambda^{(1)} + n\boldsymbol{J}_n^{(22)}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + \mathrm{o_p}(1) + \mathrm{o_p}(n^{1/2}) = \mathbf{0}.$$

Hence, we obtain

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1).$$

Theorem 5 (Asymptotic distribution of the ℓ_1-type regularized estimator)**.** Let $\boldsymbol{J}^{(1|2)} = \boldsymbol{J}^{(11)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}$, $\boldsymbol{s}_n^{(1|2)} = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}$ and

$$\hat{\boldsymbol{u}}_n^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\mathrm{argmin}}\{\boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{s}_n^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1\}. \tag{5.7}$$

Under the conditions (C1), (C2), and (Q1)–(Q5), we have $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1$ and

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$$

when $1 < q_0 < 2$. Moreover, we have

$$n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + \mathrm{o_p}(1) \tag{5.8}$$

and

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)}) + \mathrm{o_p}(1) \tag{5.9}$$

when $q_0 = 1$.

Unlike the case of $1 < q_0 < 2$, we need Lemma 3 to prove (5.8) and (5.9). This is because that we can not show the sparsity when $q_0 = 1$ and that the penalty term is not differentiable at the origin. The difference between Theorem 2 and 5 occurs from the condition (Q4), that is, (5.9) does not includes the term about $p'_{\lambda_n}(\beta)$. The proof of (5.8) and (5.9) are given in Section 6.2.4.

5.3.3 Variable selection consistency

It is easy to imagine that the variable selection consistency does not hold when $q_0 = 1$ from Lemma 5 and Theorem 5, that is, $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})$ has an asymptotic bias, so we consider the case of $q_0 \neq 1$ for the ℓ_1-type regularized estimator in this section. The variable selection consistency, $\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} \neq \mathbf{0}) \to 1$, is equivalent to $\mathrm{P}(\hat{\mathcal{J}}^{(2)} = \mathcal{J}^{(2)}) \to 1$, where $\hat{\mathcal{J}}^{(2)} = \{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ is a so-called active set. From Lemma 4, we see that for any $j \in \mathcal{J}^{(2)}$, $\mathrm{P}(j \in \hat{\mathcal{J}}^{(2)}) \to 1$ and thus $\mathrm{P}(\hat{\mathcal{J}}^{(2)} \supset \mathcal{J}^{(2)}) \to 1$. Therefore, to establish the variable selection consistency, it suffices to show that

$$\mathrm{P}(j \in \hat{\mathcal{J}}^{(2)}) \to 0 \quad \text{for any} \quad j \notin \mathcal{J}^{(2)}. \tag{5.10}$$

Variable selection consistency of the Bridge estimator

To establish (5.10), let us consider the event $j \in \hat{\mathcal{J}}^{(2)}$. As mentioned in Section 5.3.2, $\hat{\beta}_{\lambda,j}$ satisfies the likelihood equation with probability converging to 1, that is,

$$-\sum_{i=1}^{n} \frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} + qn\lambda_n \mathrm{sgn}(\hat{\beta}_{\lambda,j})|\hat{\beta}_{\lambda,j}|^{q-1} = 0$$

for any $j \in \hat{\mathcal{J}}^{(2)}$. By using the Taylor's theorem, we see that

$$-\sum_{i=1}^{n} \frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_\lambda} = -n^{1/2} s_{n,j} + n\sum_{k=1}^{p} \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}(\hat{\beta}_{\lambda,k} - \beta_k^*),$$

and then we have

$$s_{n,j} - \sum_{k=1}^{p} \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = q\lambda n^{(q_0-q)/2}\mathrm{sgn}(\hat{\beta}_{\lambda,j})|n^{1/2}\hat{\beta}_{\lambda,j}|^{q-1}, \tag{5.11}$$

where $\boldsymbol{\beta}^\dagger$ is a segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\hat{\boldsymbol{\beta}}^*$. From (R2), $\boldsymbol{s}_n = \mathrm{O_p}(1)$ and $n^{1/2}$-consistency of $\hat{\boldsymbol{\beta}}_\lambda$, the left-hand side of (5.11) reduces to $\mathrm{O_p}(1)$. Moreover, the right-hand side of (5.11) goes to infinity since $q < q_0$ and $\hat{\boldsymbol{\beta}}_\lambda$ is $n^{1/2}$-consistent. Therefore, we obtain that $\mathrm{P}(j \in \hat{\mathcal{J}}^{(2)})$ is bounded above by

$$\mathrm{P}\left(s_{n,j} - \sum_{k=1}^{p} \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = q\lambda n^{(q_0-q)/2}\mathrm{sgn}(\hat{\beta}_{\lambda,j})|n^{1/2}\hat{\beta}_{\lambda,j}|^{q-1}\right)$$

and this converges to 0 for any $j \in \mathcal{J}^{(2)}$. As a result, we have the variable selection consistency of the Bridge estimator.

Theorem 6 (variable selection consistency of the Bridge estimator)**.** Under the conditions (C1) and (C2), the Bridge estimator has the variable selection consistency.

Variable selection consistency of the ℓ_1-type regularized estimator

Let us consider the event $j \in \hat{\mathcal{J}}^{(2)}$ as in the above. Then we have

$$s_{n,j} - \sum_{k=1}^{p} \boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)_{jk}\{n^{1/2}(\hat{\beta}_{\lambda,k} - \beta_k^*)\} = n^{1/2} p'_{\lambda_n}(\hat{\beta}_{\lambda,j}) \tag{5.12}$$

for any $j \in \hat{\mathcal{J}}^{(2)}$, and the left-hand side on the equation is $\mathrm{O_p}(1)$ by the same argument. On the other hand, from (5.2) and $n^{1/2}\lambda_n \to \infty$, the right-hand side on the equation goes to infinity.

Theorem 7 (variable selection consistency of the ℓ_1-type regularized estimator)**.** Under the conditions (C1), (C2) and (Q1)–(Q5), the ℓ_1-type regularized estimator has the variable selection consistency when $1 < q_0 < 2$.

5.4 Information criterion

Now, we consider an AIC-type information criterion by the same argument as in Chapter 4. To do this, we need to consider the bias of the Kullback-Leibler divergence defined by

$$\mathrm{E}\left[\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}\left[\sum_{i=1}^{n} \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)\right]\right], \tag{5.13}$$

where $\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)$ is a log-likelihood based on the copy of $\boldsymbol{y}_i$, i.e., $\tilde{\boldsymbol{y}}_i$, and $\tilde{\mathrm{E}}$ denotes the expectation based only on the copy $(\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_n)$. We derive an AIC by evaluating $\mathrm{E}[z^{\mathrm{limit}}]$, where z^{limit} is the limit to which $\sum_{i=1}^n g_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{\mathrm{E}}[\sum_{i=1}^n \tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda)]$ converges in distribution, and we call $\mathrm{E}[z^{\mathrm{limit}}]$ an asymptotic bias. In the case of the Bridge estimation, we can see the same result as in Theorem 3, that is, the asymptotic bias reduces to $|\mathcal{J}^{(2)}|$, so we consider the asymptotic bias only for the ℓ_1-type regularized estimator.

By Theorems 4 and 5, we know, both in the case of the Bridge estimator and the ℓ_1-type regularized estimator, that $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}$ with probability converging to 1, and $n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) = \boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} + \mathrm{o_p}(1)$ for $q_0 \neq 1$. To derive the asymptotic bias of the Kullback-Leibler divergence, note that (5.13) can be rewritten by the expectation of

$$\sum_{i=1}^{n}\{g_i(\hat{\boldsymbol{\beta}}_\lambda) - g_i(\boldsymbol{\beta}^*)\} - \sum_{i=1}^{n}\{\tilde{g}_i(\hat{\boldsymbol{\beta}}_\lambda) - \tilde{g}_i(\boldsymbol{\beta}^*)\}. \tag{5.14}$$

Using the Taylor's theorem, the first term in (5.14) can be expressed as

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{s}_n - n(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\boldsymbol{\beta}^\dagger)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2,$$

where $\boldsymbol{\beta}^\dagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. From Theorem 5 and (R2), we see that it can be rewritten by

$$\boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} - \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1),$$

and this converges in distribution to

$$\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} - \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2.$$

On the other hand, the second term in (5.14) can be expressed as

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\tilde{\boldsymbol{s}}_n - n(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\boldsymbol{\beta}^\ddagger)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2,$$

where $\boldsymbol{\beta}^\ddagger$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. By the same way as in the above, it can be rewritten by

$$\boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}_n^{(2)} - \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1),$$

and this converges in distribution to

$$\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)} - \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}/2,$$

where $\tilde{\boldsymbol{s}}_n^{(2)}$ and $\tilde{\boldsymbol{s}}^{(2)}$ is a copy of $\boldsymbol{s}_n^{(2)}$ and $\boldsymbol{s}^{(2)}$, respectively. Therefore, we have

$$z^{\mathrm{limit}} = \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} - \boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}.$$

Because $\boldsymbol{s}^{(2)}$ and $\tilde{\boldsymbol{s}}^{(2)}$ are independently distributed according to $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$, the asymptotic bias reduces to

$$\mathrm{E}[z^{\mathrm{limit}}] = \mathrm{E}[\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}] = |\mathcal{J}^{(2)}|.$$

As for the case of $q_0 = 1$, we see, from Theorem 5 and (R2), that the first term and second term in (5.14) can be rewritten as

$$\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{s}_n^{(1|2)} + \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} - \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n^{(1)}/2 - \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1),$$

and

$$\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\tilde{\boldsymbol{s}}_n^{(1|2)} + \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}_n^{(2)} - \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n^{(1)}/2 - \boldsymbol{s}_n^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}/2 + \mathrm{o_p}(1),$$

respectively, where $\boldsymbol{s}_n^{(1|2)} = \boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}$, $\tilde{\boldsymbol{s}}_n^{(1|2)}$ is a copy of $\boldsymbol{s}_n^{(1|2)}$, and $\boldsymbol{u}_n^{(1)}$ is defined in (5.7). Note that, from convexity lemma in Hjort and Pollard (1993), we have that $\hat{\boldsymbol{u}}_n^{(1)}$ converges in distribution to

$$\hat{\boldsymbol{u}}^{(1)} = \underset{\boldsymbol{u}^{(1)}}{\operatorname{argmin}}\{\boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1\}, \tag{5.15}$$

and that the asymptotic bias can be reduced to

$$\mathrm{E}[z^{\mathrm{limit}}] = \mathrm{E}[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}] + \mathrm{E}[\boldsymbol{s}^{(2)\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}] = \mathrm{E}[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}] + |\mathcal{J}^{(2)}|$$

by using the independency of $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$, where $\boldsymbol{s}^{(1|2)}$ is the limit to which $\boldsymbol{s}_n^{(1|2)}$ converges in distribution. Now we have the following theorem.

Theorem 8. The asymptotic bias of the Kullback-Leibler divergence reduces to

$$\mathrm{E}[z^{\mathrm{limit}}] = |\mathcal{J}^{(2)}|$$

under the conditions (C1) and (C2) for the Bridge estimator, and the conditions (C1), (C2), and (Q1)–(Q5) for the ℓ_1-type regularized estimator with $q_0 \neq 1$. Moreover, under the conditions (C1), (C2), and (Q1)–(Q5), we have

$$\mathrm{E}[z^{\mathrm{limit}}] = |\mathcal{J}^{(2)}| + K$$

for the ℓ_1-type regularized estimator with $q_0 = 1$, where $K = \mathrm{E}[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}]$ and $\hat{\boldsymbol{u}}^{(1)}$ is defined in (5.15).

Because the asymptotic bias derived in Theorem 8 depends on an unknown value $\boldsymbol{\beta}^*$ as described in Chapter 4. Replacing $\mathcal{J}^{(2)}$ by the active set $\hat{\mathcal{J}}^{(2)} = \{j;\ \hat{\beta}_{\lambda,j} \neq 0\}$ and K by its empirical mean $\hat{K}$ obtained by generating samples from $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}_n(\hat{\boldsymbol{\beta}}_\lambda))$, we propose the following index as an AIC for the non-concave penalized maximum likelihood method:

$$\mathrm{AIC}_\lambda^{\mathrm{oracle}} = -2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}^{(2)}|, \tag{5.16}$$

for the Bridge estimator or the ℓ_1-type regularized estimator with $q_0 \neq 1$, and

$$\mathrm{AIC}_\lambda = -2\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}}_\lambda) + 2|\hat{\mathcal{J}}^{(2)}| + \hat{K}, \tag{5.17}$$

for the ℓ_1-type regularized estimator with $q_0 = 1$. Note that (5.16) is a consistent estimator of the Kullback-Leibler divergence because of the variable selection consistency, although (5.17) is not consistent.

5.5 Simulation study

In this Section, we conduct some simulation studies to check the performance of the tuning parameter selection based on the AIC in (5.16) and (5.17). Concretely speaking, we consider same setting as in Chapter 4, and the performances for the AIC in (5.16) and (5.17) are compared. As regularization methods, we consider the SCAD and the MCP. This is because that we want to check the performance and the computation time in the case of the existence of $\hat{K}$ or not. The simulation settings were the same as in Chapter 4 and fifty simulations were conducted.

We assessed the performance by evaluating the average and standard deviation of the KL as a main indices, where KL is defined in (4.26). We evaluated the expectation using an empirical mean of 500 samples As mentioned in Chapter 4, we interpret that the criterion giving the small value of the KL is good. In addition, we assessed the standard deviation of the difference between KL evaluated by each criteria:

$$\Delta = \mathrm{KL}_{q_0=1} - \mathrm{KL}_{q_0\neq 1},$$

where $\mathrm{KL}_{q_0\neq 1}$ and $\mathrm{KL}_{q_0=1}$ are the value of KL using $\hat{\lambda}$ which is the argmin of (5.16) and (5.17), respectively. The standard deviation of Δ represents a stability of the difference between $\mathrm{KL}_{q_0\neq 1}$ and $\mathrm{KL}_{q_0=1}$, so we interpret that the difference between $\mathrm{KL}_{q_0\neq 1}$ and $\mathrm{KL}_{q_0=1}$ is significant when the value of the standard deviation of Δ is small. We also checked the averages of the FP and the FN as a secondary index, where FP and FN

are defined in (4.27). Moreover, we assessed an average of a computation time (sec) to compute the AIC in (5.16) and (5.17) for fixed λ in each criteria. The AIC we use is that corresponding to the case $q_0 = 1.5$ in (5.16), and that corresponding to the case $q_0 = 1$ in (5.17). The value of K in (5.17) were evaluated by 100 samples drawn from $\mathrm{N}(\mathbf{0}, \boldsymbol{J})$.

Tables 5.1, 5.2, 5.3, and 5.4 show the results for the linear setting with SCAD, the logistic setting with SCAD, the linear case with MCP, and the logistic setting with MCP, respectively, and in each table, the averages and standard deviations of the KL, the averages of the FP and FN, the averages of the computation time (Time), and the standard deviation of Δ are provided for the linear and the logistic regression settings. We can immediately see, in each Tables, that almost the values of the KL of $\mathrm{AIC}_{\mathrm{oracle}}$ are smaller than those of AIC. This is because that, as mentioned in the above, the AIC in (5.16) is a consistent estimator of the Kullback-Leibler divergence while the AIC in (5.17) is not consistent. We also see that the average of the standard deviation of Δ become small as n increases, and thus we can say that the difference between the values of the KL given by each methods is significant. Moreover, the values of the FP (FN) of $\mathrm{AIC}_{\mathrm{oracle}}$ is smaller (larger) than that of AIC, and so the tuning parameter selected by the AIC in (5.17) tend to small compared to the AIC in (5.16). On the other hand, the computational time of the AIC requires about a 6 to 8 times compared to $\mathrm{AIC}_{\mathrm{oracle}}$ because $\mathrm{AIC}_{\mathrm{oracle}}$ can be expressed by a simple form. Thus we see that the computation time of $\mathrm{AIC}_{\mathrm{oracle}}$ is robust to scale changes. On the whole, we can conclude that the AIC in (5.16) is better than the AIC in (5.17).

		Case 1					Case 2				
(p,k,n)		KL (sd)	FP	FN	Time	Δ	KL (sd)	FP	FN	Time	Δ
(8,2,50)	AIC	0.605 (0.059)	2.20	0.90	0.631	0.028	0.617 (0.046)	2.38	0.58	0.686	0.023
	$\text{AIC}_{\text{oracle}}$	0.597 (0.059)	0.98	1.32	0.087		0.607 (0.044)	0.96	0.94	0.096	
(8,2,100)	AIC	0.516 (0.030)	2.68	0.52	0.616	0.013	0.528 (0.025)	2.80	0.22	0.520	0.013
	$\text{AIC}_{\text{oracle}}$	0.515 (0.032)	0.86	1.18	0.088		0.523 (0.027)	0.92	0.52	0.075	
(8,2,150)	AIC	0.498 (0.018)	2.78	0.42	0.560	0.007	0.507 (0.014)	2.60	0.10	0.524	0.006
	$\text{AIC}_{\text{oracle}}$	0.492 (0.017)	0.90	1.14	0.081		0.508 (0.014)	0.88	0.52	0.076	
(8,1,100)	AIC	0.530 (0.023)	3.50	0.36	0.641	0.016	0.533 (0.028)	4.64	0.06	0.552	0.015
	$\text{AIC}_{\text{oracle}}$	0.517 (0.022)	0.94	0.78	0.086		0.520 (0.026)	1.22	0.48	0.074	
(8,3,100)	AIC	0.542 (0.023)	1.30	0.82	0.603	0.014	0.571 (0.027)	1.42	0.40	0.482	0.009
	$\text{AIC}_{\text{oracle}}$	0.541 (0.027)	0.66	1.52	0.092		0.573 (0.026)	0.74	0.94	0.076	
(12,3,100)	AIC	0.562 (0.034)	4.02	0.60	0.691	0.018	0.574 (0.036)	4.26	0.28	0.631	0.013
	$\text{AIC}_{\text{oracle}}$	0.556 (0.035)	1.54	1.38	0.091		0.566 (0.036)	1.70	0.78	0.086	
(16,4,100)	AIC	0.620 (0.044)	6.08	0.70	0.812	0.023	0.595 (0.036)	6.56	0.30	0.668	0.018
	$\text{AIC}_{\text{oracle}}$	0.596 (0.041)	2.64	1.64	0.104		0.579 (0.033)	2.68	1.10	0.088	

Table 5.1: Comparison between the AIC in (5.16) and (5.17) for the SCAD penalty in the linear regression setting. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is (0.1,0.5) or (0.2,1), respectively.

		Case 1					Case 2				
(p,k,n)		KL (sd)	FP	FN	Time	Δ	KL (sd)	FP	FN	Time	Δ
(8,2,50)	AIC	0.498 (0.149)	1.12	1.02	0.664	0.059	0.438 (0.111)	1.04	0.76	0.649	0.047
	$\text{AIC}_{\text{oracle}}$	0.482 (0.148)	0.74	1.20	0.109		0.431 (0.107)	0.82	0.74	0.110	
(8,2,100)	AIC	0.402 (0.042)	1.28	0.74	0.668	0.018	0.297 (0.033)	0.94	0.22	0.606	0.012
	$\text{AIC}_{\text{oracle}}$	0.396 (0.041)	0.76	0.86	0.128		0.293 (0.032)	0.58	0.18	0.118	
(8,2,150)	AIC	0.389 (0.029)	1.32	0.50	0.689	0.012	0.315 (0.022)	1.20	0.12	0.589	0.011
	$\text{AIC}_{\text{oracle}}$	0.385 (0.024)	0.60	0.60	0.158		0.312 (0.023)	0.58	0.18	0.137	
(8,1,100)	AIC	0.532 (0.039)	1.64	0.38	0.702	0.031	0.464 (0.036)	1.62	0.12	0.604	0.024
	$\text{AIC}_{\text{oracle}}$	0.525 (0.031)	0.90	0.36	0.129		0.455 (0.035)	0.94	0.02	0.112	
(8,3,100)	AIC	0.349 (0.031)	0.64	1.26	0.641	0.013	0.277 (0.036)	0.76	0.62	0.544	0.012
	$\text{AIC}_{\text{oracle}}$	0.348 (0.033)	0.34	1.46	0.129		0.279 (0.036)	0.34	0.90	0.113	
(12,3,100)	AIC	0.417 (0.052)	2.00	1.18	0.757	0.042	0.293 (0.044)	1.48	0.48	0.640	0.019
	$\text{AIC}_{\text{oracle}}$	0.410 (0.050)	1.12	1.36	0.144		0.288 (0.044)	0.82	0.56	0.123	
(16,4,100)	AIC	0.391 (0.051)	2.52	1.68	0.847	0.039	0.295 (0.064)	1.74	1.30	0.716	0.042
	$\text{AIC}_{\text{oracle}}$	0.384 (0.056)	1.44	1.98	0.160		0.280 (0.050)	1.10	1.26	0.138	

Table 5.2: Comparison between the AIC in (5.16) and (5.17) for the SCAD penalty in the logistic regression setting. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is (0.5,1.5) or (1,2), respectively.

		Case 1					Case 2				
(p,k,n)		KL (sd)	FP	FN	Time	Δ	KL (sd)	FP	FN	Time	Δ
(8,2,50)	AIC	0.614 (0.055)	2.18	0.82	0.626	0.031	0.581 (0.072)	1.32	0.84	0.677	0.028
	$\text{AIC}_{\text{oracle}}$	0.606 (0.058)	1.12	1.52	0.087		0.575 (0.070)	0.66	1.22	0.094	
(8,2,100)	AIC	0.543 (0.034)	2.56	0.50	0.616	0.014	0.579 (0.020)	2.40	0.28	0.559	0.010
	$\text{AIC}_{\text{oracle}}$	0.531 (0.029)	0.74	1.24	0.087		0.576 (0.019)	0.88	0.78	0.079	
(8,2,150)	AIC	0.534 (0.015)	2.14	0.56	0.570	0.007	0.503 (0.019)	2.64	0.14	0.522	0.009
	$\text{AIC}_{\text{oracle}}$	0.534 (0.014)	0.78	1.16	0.082		0.502 (0.020)	0.70	0.58	0.076	
(8,1,100)	AIC	0.531 (0.026)	3.66	0.32	0.637	0.016	0.543 (0.023)	3.50	0.18	0.562	0.012
	$\text{AIC}_{\text{oracle}}$	0.516 (0.024)	1.16	0.68	0.086		0.537 (0.021)	1.24	0.50	0.074	
(8,3,100)	AIC	0.518 (0.027)	1.20	1.08	0.601	0.011	0.539 (0.023)	1.06	0.50	0.486	0.011
	$\text{AIC}_{\text{oracle}}$	0.515 (0.026)	0.40	1.92	0.090		0.539 (0.019)	0.44	1.02	0.075	
(12,3,100)	AIC	0.518 (0.032)	3.90	0.86	0.701	0.023	0.543 (0.030)	3.68	0.34	0.629	0.018
	$\text{AIC}_{\text{oracle}}$	0.501 (0.031)	1.26	1.78	0.091		0.534 (0.028)	1.02	1.12	0.085	
(16,4,100)	AIC	0.574 (0.043)	5.32	0.88	0.828	0.027	0.594 (0.050)	5.30	0.70	0.691	0.025
	$\text{AIC}_{\text{oracle}}$	0.553 (0.043)	1.56	2.24	0.104		0.581 (0.043)	1.36	1.46	0.089	

Table 5.3: Comparison between the AIC in (5.16) and (5.17) for the MCP penalty in the linear regression setting. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is (0.1,0.5) or (0.2,1), respectively.

		Case 1					Case 2				
(p,k,n)		KL (sd)	FP	FN	Time	Δ	KL (sd)	FP	FN	Time	Δ
(8,2,50)	AIC	0.509 (0.091)	0.88	1.48	0.695	0.066	0.401 (0.083)	0.74	0.78	0.703	0.032
	$\text{AIC}_{\text{oracle}}$	0.499 (0.113)	0.74	1.38	0.117		0.407 (0.088)	0.72	0.66	0.122	
(8,2,100)	AIC	0.390 (0.030)	0.88	0.88	0.644	0.020	0.317 (0.031)	0.62	0.30	0.545	0.013
	$\text{AIC}_{\text{oracle}}$	0.387 (0.025)	0.68	0.88	0.128		0.313 (0.029)	0.58	0.18	0.111	
(8,2,150)	AIC	0.388 (0.025)	1.42	0.60	0.662	0.012	0.301 (0.023)	0.86	0.16	0.568	0.011
	$\text{AIC}_{\text{oracle}}$	0.384 (0.024)	0.76	0.66	0.157		0.300 (0.023)	0.64	0.14	0.137	
(8,1,100)	AIC	0.511 (0.030)	1.58	0.40	0.676	0.024	0.425 (0.026)	0.94	0.08	0.576	0.018
	$\text{AIC}_{\text{oracle}}$	0.509 (0.023)	1.18	0.32	0.128		0.421 (0.026)	0.76	0.02	0.109	
(8,3,100)	AIC	0.372 (0.035)	0.42	1.44	0.604	0.021	0.252 (0.036)	0.24	0.90	0.509	0.011
	$\text{AIC}_{\text{oracle}}$	0.369 (0.028)	0.24	1.60	0.128		0.250 (0.032)	0.18	0.88	0.108	
(12,3,100)	AIC	0.397 (0.058)	1.28	1.54	0.724	0.030	0.312 (0.051)	1.12	0.58	0.603	0.021
	$\text{AIC}_{\text{oracle}}$	0.395 (0.058)	1.22	1.42	0.142		0.310 (0.054)	1.06	0.52	0.121	
(16,4,100)	AIC	0.379 (0.075)	1.88	2.14	0.807	0.036	0.254 (0.060)	1.02	1.60	0.678	0.041
	$\text{AIC}_{\text{oracle}}$	0.378 (0.080)	1.66	2.14	0.160		0.248 (0.054)	0.98	1.44	0.135	

Table 5.4: Comparison between the AIC in (5.16) and (5.17) for the MCP penalty in the logistic regression setting. For Case 1 or 2, the true parameter vector (β_1^*, β_2^*) is (0.5,1.5) or (1,2), respectively.

5.6 Summary in this chapter

In this chapter, we have derived the oracle property of the estimator defined in (5.1), and developed the AIC-type information criterion by the same way as in Chapter 4. The resulting AIC has had a almost the same asymptotic bias for the Bridge penalty or ℓ_1-type penalty with $q_0 = 1$, although the AIC has became more simply for the ℓ_1-type penalty with $q_0 \neq 1$. This phenomenon has occurred from the fact that whether the estimator has the sparsity. We have also discussed the consistency of the AIC in (5.16), and found that the AIC for non-concave regularized method is a consistent estimator of the Kullback-Leibler divergence when the variable selection consistency is hold. It is interesting that the asymptotic behavior of the Bridge estimator and the ℓ_1-type regularized estimator are different by the value of q_0. That is, we require $q < q_0 < 1$ for the Bridge estimator and $1 < q_0 < 2$ for the ℓ_1-type regularized estimator to assure the oracle property. This is because the Bridge penalty has a property that the derivative at the origin diverges and that it is not bounded uniformly, although the ℓ_1-type penalty behaves like the Lasso around the origin and is uniformly bounded. However we have only considered the Bridge penalty for ℓ_q-type regularization method in this chapter, we can generalize the result for a general ℓ_q penalty to assume about the behavior of the first derivative of such penalty.

In the simulation study, it has been shown that the AIC in (5.16) is superior to the AIC in (5.17) both of the prediction accuracy and the computation time for the SCAD and MCP. Because of the simple expression of the AIC in (5.16), it will be significant in handling large-scale data same as in the Bridge regularization method.

It is an important future work to show that a similar asymptotic property for a high-dimensional data as mentioned in Chapter 4.

Chapter 6

Appendix

6.1 Some useful matrix algebra

Here, we describe some useful matrix algebra that we have frequently used in Chapter 3. All the proofs can be found in Harville (1997).

Lemma 6 (Theorem 8.5.11 in Harville (1997))**.** Let $A \in \mathbb{R}^{p\times p},\ B \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{q\times q}$, and assume that A is non-singular. Then

$$\begin{pmatrix} A & B \\ B^{\mathrm{T}} & C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BD^{-1}B^{\mathrm{T}}A^{-1} & -A^{-1}BD^{-1} \\ -D^{-1}B^{\mathrm{T}}A^{-1} & D^{-1} \end{pmatrix},$$

if and only if $D = C - B^{\mathrm{T}}A^{-1}B$ is non-singular.

Lemma 7 (Theorem 18.2.8 in Harville (1997))**.** Let $A \in \mathbb{R}^{p\times p},\ B \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{q\times q}$, and assume that A and C are non-singular. Then

$$(A + BCB^{\mathrm{T}})^{-1} = A^{-1} - A^{-1}B(C^{-1} + B^{\mathrm{T}}A^{-1}B)^{-1}B^{\mathrm{T}}A^{-1},$$

if and only if $C^{-1} + B^{\mathrm{T}}A^{-1}B$ is non-singular.

Lemma 2 is also known as Woodbury's formula, and in the special case when $A = \alpha I_p,\ B = \boldsymbol{b} \in \mathbb{R}^p$, and $C = \beta \in \mathbb{R}$ such that $\alpha \neq 0$ and $\alpha + \beta\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b} \neq 0$, we have

$$(\alpha I_p + \beta\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}})^{-1} = \frac{1}{\alpha}\left(I_p - \frac{\beta}{\alpha + \beta\boldsymbol{b}^{\mathrm{T}}\boldsymbol{b}}\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}\right) \tag{6.1}$$

Lemma 8 (Theorem 16.2.2 in Harville (1997))**.** For matrices $A \in \mathbb{R}^{p\times q},\ B \in \mathbb{R}^{q\times q}$, and $C \in \mathbb{R}^{p\times p}$, we have

$$\mathrm{tr}(A^{\mathrm{T}}CAB^{\mathrm{T}}) = \mathrm{vec}(A)^{\mathrm{T}}(B \otimes C)\mathrm{vec}(A)$$

From Lemma 8, we can immediately see that

$$\mathrm{tr}(D^{\mathrm{T}}A^{\mathrm{T}}CAB^{\mathrm{T}}) = \mathrm{vec}(A)^{\mathrm{T}}(DB \otimes C)\mathrm{vec}(A) \tag{6.2}$$

for matrices $A,\ B,\ C$ defined in Lemma 8, and $D \in \mathbb{R}^{q\times q}$

6.2 Proofs

6.2.1 Proofs in Chapter 2

Derivation of (2.5)

From Bayes' theorem and $\boldsymbol{z}|X \sim \mathrm{N}(\boldsymbol{0}, K)$, we have

$$\begin{aligned}\log f(\boldsymbol{z}|X, \boldsymbol{y}) &= \log f(\boldsymbol{y}|\boldsymbol{z}, X) + \log f(\boldsymbol{z}|X) + \text{const.} \\ &= \boldsymbol{z}^{\mathrm{T}}\boldsymbol{y} - \sum_{i=1}^{n}\log(1+e^{z_i}) - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}K^{-1}\boldsymbol{z} + \text{const.},\end{aligned}$$

where const. is a constant term which does not depend on $\boldsymbol{z}$. The first and the second derivatives of the log-likelihood function is given by $\boldsymbol{y} - \boldsymbol{p} - K^{-1}\boldsymbol{z}$ and $-P - K^{-1}$, respectively. Because $\boldsymbol{p}$ and P are depending on $\boldsymbol{z}$, the update equation of $\boldsymbol{z}$ is given by

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - (-P^{(t)} - K^{-1})^{-1}(\boldsymbol{y} - \boldsymbol{p}^{(t)} - K^{-1}\boldsymbol{z}^{(t)}),$$

and we obtain (2.5) by a simple calculation.

Derivation of (2.9) and (2.10)

To evaluate (2.9) and (2.10), we use the inequality as follows:

$$\log 2 + \frac{1}{2}z \le \log(1+\exp(z)) \le \log 2 + z1_{\{z>0\}} \quad {}^{\forall}z \in \mathbb{R}. \tag{6.3}$$

By using the left side inequality in (6.3) and $\mathrm{E}[Z|\boldsymbol{y}] = \nu$, we can easily obtain (2.10). On the other hand, by using the right side inequality in (6.3), we have a upper bound

$$\mathrm{E}[Z(Z-\nu)\log(1+e^Z)|\boldsymbol{y}] \leq \mathrm{E}[Z(Z-\nu)|\boldsymbol{y}]\log 2 + \mathrm{E}[Z^2(Z-\nu)1_{\{Z>0\}}|\boldsymbol{y}].$$

Since $\mathrm{V}[Z|\boldsymbol{y}] = \tau^2$, the first term in the right-had side is rewritten as $\tau^2 \log 2$. Note that the second term in the right-hand side can be decomposed as

$$\mathrm{E}[(Z-\nu)^3 1_{\{Z>0\}}|\boldsymbol{y}] + 2\nu\mathrm{E}[(Z-\nu)^2 1_{\{Z>0\}}|\boldsymbol{y}] + \nu^2\mathrm{E}[(Z-\nu)1_{\{Z>0\}}|\boldsymbol{y}].$$

Because $Z|\boldsymbol{y}$ is distributed according to a Gaussian distribution with mean ν and variance τ^2, the first term and the third term can be evaluated analytically as

$$\mathrm{E}[(Z-\nu)^3 1_{\{Z>0\}}|\boldsymbol{y}] = \frac{\tau(\nu^2+2\tau^2)}{\sqrt{2\pi}}\exp\left(-\frac{\nu^2}{2\tau^2}\right)$$

and

$$\nu^2\mathrm{E}[(Z-\nu)1_{\{Z>0\}}|\boldsymbol{y}] = \frac{\tau\nu^2}{\sqrt{2\pi}}\exp\left(-\frac{\nu^2}{2\tau^2}\right),$$

respectively. On the other hand, since

$$0 \leq \mathrm{E}[(Z-\nu)^2 1_{\{Z>0\}}|\boldsymbol{y}] \leq \mathrm{E}[(Z-\nu)^2|\boldsymbol{y}] = \tau^2,$$

the second term can be bounded from above by $2\tau^2 \max\{\nu, 0\}$. Combining all the above, we obtain (2.9).

6.2.2 Proofs in Chapter 3

Derivation of (3.2)

Noting that $A'(AA')^{-1}A = I_K$ by the definition of A, it follows that

$$Y_i - X_i B^\dagger A = E_i - \sum_{j=1}^{n} X_i (X'X)^{-1} X_j' E_j. \tag{6.4}$$

In addition, for an $(M \times M)$-dimensional matrix M, we can see that

$$\mathrm{E}[E_i' M E_i] = \mathrm{tr}(M\Xi)\Sigma. \tag{6.5}$$

In fact, the (j,k)-th element of $E_i'ME_i$ is given by $\boldsymbol{\varepsilon}_{ij}'M\boldsymbol{\varepsilon}_{ik}$ for $j,k=1,2,\ldots,K$, and thus we have

$$\mathrm{E}[\boldsymbol{\varepsilon}_{ij}'M\boldsymbol{\varepsilon}_{ik}] = \mathrm{tr}(M\mathrm{E}[\boldsymbol{\varepsilon}_{ik}\boldsymbol{\varepsilon}_{ij}']) = \mathrm{tr}(M\Xi)\sigma_{kj}.$$

This and the symmetry of Σ imply (6.5). From (6.4), (6.5), and the independence of E_i, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\mathrm{E}[(Y_i - X_iB^\dagger A)'(Y_i - X_iB^\dagger A)] \\
=&\sum_{i=1}^{n}\mathrm{E}[E_i'E_i] - 2\sum_{i,j}\mathrm{E}[E_j'X_j(X'X)^{-1}X_i'E_i] + \sum_{i,j,k}\mathrm{E}[E_j'X_j(X'X)^{-1}X_i'X_i(X'X)^{-1}X_k'E_k] \\
=&n\mathrm{tr}(\Xi)\Sigma - \sum_{j=1}^{n}\mathrm{tr}\{X_j(X'X)^{-1}X_j'\Xi\}\Sigma = \{n\mathrm{tr}(\Xi) - \tilde{J}\}\Sigma,
\end{aligned}$$

and this complete the proof.

Derivation of (3.13)

By the same argument in Section 3.3, we see that

$$(X'X)^{-1} = \begin{pmatrix} J^{-2}\mathrm{tr}(\Delta^{-1}) & J^{-1}\mathbf{1}_{J-1}'Q \\ J^{-1}Q'\mathbf{1}_{J-1} & R \end{pmatrix},$$

where $Q = \Delta_{J-1}^{-1} - J^{-1}\mathrm{tr}(\Delta^{-1})I_{J-1}$ and $R = (C'C)^{-1}C'\Delta^{-1}C(C'C)^{-1}$. Then we have

$$\begin{aligned}
&X_j(X'X)^{-1}X_j' \\
=&\frac{\mathrm{tr}(\Delta^{-1})}{J^2}\mathbf{1}_M\mathbf{1}_M' + \frac{1}{J}\tilde{X}_jCQ'\mathbf{1}_{J-1}\mathbf{1}_M' + \frac{1}{J}\mathbf{1}_M\mathbf{1}_{J-1}'QC'\tilde{X}_j' + \tilde{X}_jCRC\tilde{X}_j'.
\end{aligned}$$

Note that $CQ'\mathbf{1}_{J-1} = (\Delta^{-1} - J^{-1}\mathrm{tr}(\Delta^{-1})I_J)\mathbf{1}_J$ and $CRC = (I_J - J^{-1}\mathbf{1}_J\mathbf{1}_J')\Delta^{-1}(I_J - J^{-1}\mathbf{1}_J\mathbf{1}_J')$ by a simple calculation, and that $\tilde{X}_j\mathbf{1}_J = \mathbf{1}_M$. Hence, $X_j(X'X)^{-1}X_j'$ can be reduced to $\tilde{X}_j\Delta^{-1}\tilde{X}_j'$ and we see that this is a diagonal matrix. Finally, since $\Xi = (1-\rho)I_M + \rho\mathbf{1}_M\mathbf{1}_M'$ and

$$\sum_{j=1}^{n}\mathrm{tr}\{X_j(X'X)^{-1}X_j'\mathbf{1}_M\mathbf{1}_M'\} = \sum_{j=1}^{n}\mathrm{tr}\{\tilde{X}_j\Delta^{-1}\tilde{X}_j'\mathbf{1}_M\mathbf{1}_M'\} = \sum_{j=1}^{n}\mathbf{1}_M'\tilde{X}_j\Delta^{-1}\tilde{X}_j'\mathbf{1}_M$$

$$= \sum_{j=1}^{n} \mathrm{tr}\{\tilde{X}_j \Delta^{-1} \tilde{X}_j'\} = J,$$

we obtain

$$\begin{aligned}\sum_{j=1}^{n} \mathrm{tr}\{X_j (X'X)^{-1} X_j' \Xi\} =& (1-\rho) \sum_{j=1}^{n} \mathrm{tr}\{X_j (X'X)^{-1} X_j'\} + \rho \sum_{j=1}^{n} \mathrm{tr}\{X_j (X'X)^{-1} X_j' \mathbf{1}_M \mathbf{1}_M'\} \\ =& (1-\rho) J + \rho J = J.\end{aligned}$$

This and $\mathrm{tr}(\Xi) = M$ imply (3.13).

6.2.3 Proofs in Chapter 4

Proof of (R3)

For any given p-dimensional vector $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\| = 1$, let $\boldsymbol{y}_i^* = a''(\boldsymbol{X}_i\boldsymbol{\beta}^*)^{-1/2}\{\boldsymbol{y}_i - a'(\boldsymbol{X}_i\boldsymbol{\beta}^*)\}$ and $\omega_{ni} = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} g_i'(\boldsymbol{\beta}^*) = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \boldsymbol{X}_i^{\mathrm{T}}\{\boldsymbol{y}_i - a'(\boldsymbol{X}_i\boldsymbol{\beta})\}$. By using the Cauchy-Schwarz inequality, it follows that

$$\omega_{ni}^2 \le \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta}^*) \boldsymbol{X}_i \{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \boldsymbol{\alpha} \times \|\boldsymbol{y}_i^*\|^2.$$

Letting $\gamma_{ni} = \boldsymbol{\alpha}^{\mathrm{T}}\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \boldsymbol{X}_i^{\mathrm{T}} a''(\boldsymbol{X}_i\boldsymbol{\beta}^*) \boldsymbol{X}_i \{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \boldsymbol{\alpha}$, we see that $\max_i \gamma_{ni} \to 0$ because the minimum eigenvalue of $n\boldsymbol{J}_n(\boldsymbol{\beta}^*)$ goes to infinity from (R2). Note that $\sum_{i=1}^n \gamma_{ni} = 1$ and, because of the existing of the moments of $\boldsymbol{y}_i$ for our model, $\mathrm{E}[\|\boldsymbol{y}_i^*\|^4] < C$ for some constant $C > 0$. Then for any $\varepsilon > 0$, we have

$$\sum_{i=1}^{n} \mathrm{E}[\omega_{ni}^2 1_{\{|\omega_{ni}| > \varepsilon\}}] \le \sum_{i=1}^{n} \mathrm{E}\left[\gamma_{ni} \|\boldsymbol{y}_i^*\|^2 1_{\{\|\boldsymbol{y}_i^*\|^2 > \varepsilon^2/\gamma_{ni}\}}\right] \le \sum_{i=1}^{n} \frac{\gamma_{ni}^2}{\varepsilon^2} \mathrm{E}[\|\boldsymbol{y}_i^*\|^4] \le \frac{C}{\varepsilon^2} \max_i \gamma_{ni},$$

and the right-hand side of the inequality converges to 0 as $n \to \infty$. Thus the Lindeberg central limit theorem and the Cramér-Wold device imply that

$$\{n\boldsymbol{J}_n(\boldsymbol{\beta}^*)\}^{-1/2} \sum_{i=1}^{n} g'(\boldsymbol{\beta}^*) \xrightarrow{\mathrm{d}} N(\mathbf{0}, I_p),$$

From this and (R2), we obtain (R3).

Proof of Lemma 2

Let us define a random function as follows:

$$\mu_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} + n^{-1/2}\sum_{j=1}^{p} p_\lambda(\beta_j).$$

From (R1), the first term in the right-hand side converges in probability to $h(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$. In addition, from the convexity of $\mu_n(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, it holds

$$\sup_{\boldsymbol{\beta}\in K}\left| n^{-1}\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} - h(\boldsymbol{\beta})\right| \xrightarrow{\mathrm{p}} 0$$

for any compact set K (Andersen and Gill 1982; Pollard 1991). Then we have

$$\sup_{\boldsymbol{\beta}\in K}|\mu_n(\boldsymbol{\beta}) - h(\boldsymbol{\beta})| \xrightarrow{\mathrm{p}} 0. \tag{6.6}$$

Note that in the following inequality,

$$\mu_n(\boldsymbol{\beta}) \geq n^{-1}\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} \equiv \mu_n^{(0)}(\boldsymbol{\beta}),$$

the argmin of the right-hand side is the maximum likelihood estimator and is $\mathrm{O_p}(1)$. Also note that for some $M > 0$,

$$\mathrm{P}(\|\hat{\boldsymbol{\beta}}_\lambda\| > M) \leq \mathrm{P}\left(\inf_{\|\boldsymbol{\beta}\|>M}\mu_n(\boldsymbol{\beta}) \leq \mu_n(\mathbf{0})\right) \leq \mathrm{P}\left(\inf_{\|\boldsymbol{\beta}\|>M}\mu_n^{(0)}(\boldsymbol{\beta}) \leq \mu_n^{(0)}(\mathbf{0})\right)$$

because $p_\lambda(0) = 0$ from (P2). Therefore we have

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname*{agmin}_{\boldsymbol{\beta}\in\mathcal{B}}\mu_n(\boldsymbol{\beta}) = \mathrm{O_p}(1). \tag{6.7}$$

From (6.6) and (6.7), we obtain

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{B}}\mu_n(\boldsymbol{\beta}) \xrightarrow{\mathrm{p}} \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{B}} h(\boldsymbol{\beta}) = \boldsymbol{\beta}^*.$$

Proof of (4.6)

Let $\boldsymbol{u} = \tilde{\boldsymbol{u}}_n + l\boldsymbol{w}$, where $\boldsymbol{w}$ is a unit vector, and let $l \in (\delta, \xi)$. The strong convexity of $\eta_n(\boldsymbol{u})$ implies

$$(1-\delta/l)\eta_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\eta_n(\boldsymbol{u}) > \eta_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}),$$

and we thus have

$$\begin{aligned}(\delta/l)\{\nu_n(\boldsymbol{u}) - \nu_n(\tilde{\boldsymbol{u}}_n)\} >& \nu_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \nu_n(\tilde{\boldsymbol{u}}_n) \\ &+ (1-\delta/l)\phi_n(\tilde{\boldsymbol{u}}_n) + (\delta/l)\phi_n(\boldsymbol{u}) - \phi_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) \\ &+ (1-\delta/l)\psi_n(\tilde{\boldsymbol{u}}_n^\dagger) + (\delta/l)\psi_n(\boldsymbol{u}^\dagger) - \psi_n(\tilde{\boldsymbol{u}}_n^\dagger + \delta\boldsymbol{w}^\dagger).\end{aligned}$$

Since it follows that

$$\begin{aligned}&\nu_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \nu_n(\tilde{\boldsymbol{u}}_n) \\ &= \{\nu_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w})\} + \{\tilde{\nu}_n(\tilde{\boldsymbol{u}}_n + \delta\boldsymbol{w}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n)\} + \{\tilde{\nu}_n(\tilde{\boldsymbol{u}}_n) - \nu_n(\tilde{\boldsymbol{u}}_n)\} \\ &\geq \Upsilon_n(\delta) - 2\Delta_n(\delta),\end{aligned}$$

we obtain from (4.4) and (4.5) that, for any ε (> 0),

$$(\delta/l)\{\nu_n(\boldsymbol{u}) - \nu_n(\tilde{\boldsymbol{u}}_n)\} > \Upsilon_n(\delta) - 2\Delta_n(\delta) - \varepsilon$$

for sufficiently large n and sufficiently small γ. If $2\Delta_n(\delta)+\varepsilon < \Upsilon_n(\delta)$, then $\nu_n(\boldsymbol{u}) \geq \nu_n(\tilde{\boldsymbol{u}}_n)$ for any $\boldsymbol{u}$ such that $|\boldsymbol{u}^\dagger| \leq \gamma$ and $\delta \leq |\boldsymbol{u} - \tilde{\boldsymbol{u}}_n| \leq \xi$. This means $\boldsymbol{u}_n$ must satisfy $|\boldsymbol{u}_n^\dagger| > \gamma$ or $|\boldsymbol{u}_n - \tilde{\boldsymbol{u}}_n| \notin [\delta, \xi]$ in order for $\boldsymbol{u}_n$ to be the argmin of $\nu_n(\boldsymbol{u})$. Hence, we obtain (4.6).

Proof of (4.15)

Let us define a random function as follows:

$$\omega_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\{g_i(\boldsymbol{\beta}^*) - g_i(\boldsymbol{\beta})\} + n^{-1/2}\sum_{j=1}^{p}\{p_\lambda(\beta_j) - p_\lambda(\beta_j^*)\}.$$

Since $p_\lambda(0) = 0$ from (P2), we have

$$\begin{aligned}\omega_n(\hat{\boldsymbol{\beta}}_\lambda) =& -n^{-1/2}\boldsymbol{s}_n^{\mathrm{T}}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2 \\ &+ n^{-1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j}) + n^{-1/2}\sum_{j\in\mathcal{J}^{(2)}} p'_\lambda(\beta_j^*)(\hat{\beta}_{\lambda,j} - \beta_j^*)\{1 + \mathrm{o_p}(1)\},\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is a vector on the segment from $\hat{\boldsymbol{\beta}}_\lambda$ to $\boldsymbol{\beta}^*$. Then, we have

$$0 \geq \omega_n(\hat{\boldsymbol{\beta}}_\lambda) - \omega_n(\boldsymbol{\beta}^*) \geq \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|) + (\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)/2$$

because $\boldsymbol{s}_n = \mathrm{O_p}(1)$. From (C2), $\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})$ is positive definite for sufficiently large n, and therefore it follows that

$$\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^* = \mathrm{O_p}(n^{-1/2}). \tag{6.8}$$

Let us express $\omega_n(\boldsymbol{\beta})$ by $\omega_n(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$. Because $0 \geq \omega_n(\hat{\boldsymbol{\beta}}_\lambda^{(1)}, \hat{\boldsymbol{\beta}}_\lambda^{(2)}) - \omega_n(\mathbf{0}, \hat{\boldsymbol{\beta}}_\lambda^{(2)})$, we see that

$$-n^{-1/2}\boldsymbol{s}_n^{(1)\mathrm{T}}\hat{\boldsymbol{\beta}}_\lambda^{(1)} + \hat{\boldsymbol{\beta}}_\lambda^{(1)\mathrm{T}}\boldsymbol{J}_n^{(11)}(\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}_\lambda^{(1)}/2 + \hat{\boldsymbol{\beta}}_\lambda^{(1)\mathrm{T}}\boldsymbol{J}_n^{(11)}(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)}) + n^{-1/2}\sum_{j\in\mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j})$$

is non-positive. Here we use that $\sum_{j\in\mathcal{J}^{(1)}} p_\lambda(\hat{\beta}_{\lambda,j})$ reduces to $\lambda\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q\{1 + \mathrm{o_p}(1)\}$ from (P2) and (6.8) and that $\boldsymbol{J}_n(\tilde{\boldsymbol{\beta}})$ is positive definite for sufficiently large n. Then we have

$$\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|^2 + n^{-1/2}\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q\{1 + \mathrm{o_p}(1)\} \leq \mathrm{O_p}(n^{-1/2}\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|)$$

and thus $\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|_q^q \leq \mathrm{O_p}(\|\hat{\boldsymbol{\beta}}_\lambda^{(1)}\|)$. Hence it holds

$$\mathrm{P}(\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathbf{0}) \to 1 \tag{6.9}$$

because $0 < q < 1$ and $\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \mathrm{o_p}(1)$. This implies the former in (4.15). Since $\tilde{\boldsymbol{u}}_n^{(2)}$ is trivially $\mathrm{O_p}(1)$, we obtain the latter of (4.15) from (6.8) and (6.9).

Proof of (4.17) and (4.18)

Let $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ be the one with $q = 1$ in (4.7), and let $\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{s}_n + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{u}/2$ in place of (4.8). Then, we can obtain $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \mathrm{o_p}(1)$ by taking a Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\mathbf{0}, \mathbf{0})$. In addition, let $\phi_n(\boldsymbol{u})$ and $\phi(\boldsymbol{u})$ be $\phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ and $\phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$ with $q = 1$ in (4.9), (4.10) and (4.11), let $\boldsymbol{u}^\dagger$ be empty vector and $\psi_n(\boldsymbol{u}^\dagger) = \psi(\boldsymbol{u}^\dagger) = 0$, and define $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi_n(\boldsymbol{u}) + \psi_n(\boldsymbol{u}^\dagger)$ and $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi(\boldsymbol{u}) + \psi(\boldsymbol{u}^\dagger)$ again. Here, note that

$$(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}) = \operatorname*{argmin}_{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})} \nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})).$$

Next, because

$$\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}\{-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})\}\|^2_{\boldsymbol{J}^{(22)}}/2 + \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\|\boldsymbol{u}^{(1)}\|_1 - \|\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}\|^2_{\boldsymbol{J}^{(22)-1}}/2,$$

we see that

$$(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) = \underset{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\operatorname{argmin}} \tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\hat{\boldsymbol{u}}_n^{(1)}, -\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)} + \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)}))$$

by using $\hat{\boldsymbol{u}}_n^{(1)}$ in (4.16), where we denote $\boldsymbol{x}^{\mathrm{T}}A\boldsymbol{x}$ by $\|\boldsymbol{x}\|^2_A$ for an appropriate size of matrix A and vector $\boldsymbol{x}$. Now we apply Lemma 3 and evaluate the right-hand side in (4.2). First, it follows that $\Delta_n(\delta)$ converges in probability to 0 by the same way as in (4.13). Next, because of the definition of $\tilde{\boldsymbol{u}}_n^{(1)}$, it must hold

$$\boldsymbol{J}^{(1|2)}\tilde{\boldsymbol{u}}_n^{(1)} - \boldsymbol{\tau}_\lambda(\boldsymbol{s}_n) + \lambda\boldsymbol{\gamma} = \boldsymbol{0},$$

where $\boldsymbol{\gamma}$ is a $|\mathcal{J}^{(1)}|$-dimensional vector such that $\gamma_j = 1$ when $\hat{u}_{n,j}^{(1)} > 0$, $\gamma_j = -1$ when $\hat{u}_{n,j}^{(1)} < 0$, and $\gamma_j \in [-1, 1]$ when $\hat{u}_{n,j}^{(1)} = 0$. Thus, noting that $\tilde{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{\gamma} = \|\tilde{\boldsymbol{u}}_n^{(1)}\|_1$, we can write $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ as

$$\|\boldsymbol{u}^{(1)} - \tilde{\boldsymbol{u}}_n^{(1)}\|^2_{\boldsymbol{J}^{(1|2)}}/2 + \lambda \sum_{j \in \mathcal{J}^{(1)}} (|u_j| - \gamma_j u_j) + \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}\{-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})\}\|^2_{\boldsymbol{J}^{(22)}}/2 \tag{6.10}$$

by a simple calculation. Let $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ be unit vectors such that $\boldsymbol{u}^{(1)} = \tilde{\boldsymbol{u}}_n^{(1)} + \zeta\boldsymbol{w}_1$ and $\boldsymbol{u}^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + (\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2$, where $0 \le \zeta \le \delta$. Then, letting $\rho^{(22)}$ and $\rho^{(1|2)}$ (> 0) be half the smallest eigenvalues of $\boldsymbol{J}^{(22)}$ and $\boldsymbol{J}^{(1|2)}$, respectively, it follows that

$$\Upsilon_n(\delta) \ge \min_{0 \le \zeta \le \delta} \left\{\rho^{(1|2)}\zeta^2 + \rho^{(22)}\|(\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2 + \zeta\boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\boldsymbol{w}_1\|^2\right\} > 0$$

because the second term in (6.10) is non-negative. Hence the first term on the right-hand side in (4.2) converges to 0. In addition, because $(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)})$ is $\mathrm{O_p}(1)$ from (6.8) and $(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ is also $\mathrm{O_p}(1)$ of course, the second term on the right-hand side in (4.12) can be reduced to arbitrarily small by considering a sufficiently large ξ. Thus we have $\|\boldsymbol{u} - \tilde{\boldsymbol{u}}_n\| = \mathrm{o_p}(1)$, and as a consequence, we obtain (4.17) and (4.18).

Proof of (4.24)

Because $n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)} = \hat{\boldsymbol{u}}_n^{(1)} + \mathrm{o_p}(1)$ from Theorem 2, the terms including $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$ do not reduce to $\mathrm{o_p}(1)$ in this case. Therefore (4.22) is expressed as

$$\begin{aligned}&\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}(\boldsymbol{s}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)}) + (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}_n^{(2)} \\ &- \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n/2 - (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})/2 + \mathrm{o_p}(1),\end{aligned}$$

and this converges in distribution to

$$\begin{aligned}&\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)} + (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} \\ &- \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}/2 - (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})/2.\end{aligned}$$

By the same way, (4.23) is expressed as

$$\begin{aligned}&\hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}(\tilde{\boldsymbol{s}}_n^{(1)} - \boldsymbol{J}^{(12)}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}_n^{(2)}) + (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}_n^{(2)} \\ &- \hat{\boldsymbol{u}}_n^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}_n/2 - (\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{s}_n^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})/2 + \mathrm{o_p}(1),\end{aligned}$$

and this converges in distribution to

$$\begin{aligned}&\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\tilde{\boldsymbol{s}}^{(1|2)} + (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)} \\ &- \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\hat{\boldsymbol{u}}/2 - (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)}(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})/2,\end{aligned}$$

where $\tilde{\boldsymbol{s}}_n^{(1)}$, $\tilde{\boldsymbol{s}}^{(2)}$, $\tilde{\boldsymbol{s}}^{(1|2)}$ and $\tilde{\boldsymbol{s}}^{(2)}$ are copies of $\boldsymbol{s}_n^{(1)}$, $\boldsymbol{s}^{(2)}$, $\boldsymbol{s}^{(1|2)}$ and $\boldsymbol{s}^{(2)}$, respectively. Thus we see that

$$z^{\mathrm{limit}} = \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)} + (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)} - \hat{\boldsymbol{u}}^{(1)\mathrm{T}}\tilde{\boldsymbol{s}}^{(1|2)} - (\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\tilde{\boldsymbol{s}}^{(2)}.$$

Since $\tilde{\boldsymbol{s}}$ and $\boldsymbol{s}$ are independently distributed according to $\mathrm{N}(\boldsymbol{0}, \boldsymbol{J}^{(22)})$, the asymptotic bias reduces to

$$\mathrm{E}[z^{\mathrm{limit}}] = \mathrm{E}[\hat{\boldsymbol{u}}^{(1)\mathrm{T}}\boldsymbol{s}^{(1|2)}] + \mathrm{E}[(\boldsymbol{s}^{(2)} - \boldsymbol{p}_\lambda^{\prime(2)})^{\mathrm{T}}\boldsymbol{J}^{(22)-1}\boldsymbol{s}^{(2)}].$$

As a result, we obtain (4.24).

6.2.4 Proofs in Chapter 5

Proof of (5.8) and (5.9)

Let us define

$$\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \sum_{i=1}^{n} \{g_i(\boldsymbol{\beta}^{*(1)}, \boldsymbol{\beta}^{*(2)}) - g_i(n^{-1/2}\boldsymbol{u}^{(1)}, n^{-1/2}\boldsymbol{u}^{(2)} + \boldsymbol{\beta}^{*(2)})\}$$

and

$$\tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = -\boldsymbol{u}^{\mathrm{T}}\boldsymbol{s}_n + \boldsymbol{u}^{\mathrm{T}}\boldsymbol{J}\boldsymbol{u}/2.$$

in the same way as in Section 6.2.3. Then we obtain $\eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \mathrm{o_p}(1)$ again by using the Taylor expansion around $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\boldsymbol{0}, \boldsymbol{0})$. In addition, let

$$\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = n \sum_{j \in \mathcal{J}^{(1)}} p_{\lambda_n}(n^{-1/2}u_j) + n \sum_{i \in \mathcal{J}^{(2)}} \{p_{\lambda_n}(n^{-1/2}u_j + \beta_j^*) - p_{\lambda_n}(\beta_j^*)\}$$

and $\phi(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \lambda\|\boldsymbol{u}^{(1)}\|_1$. By noting that $n^{1/2}\lambda_n = \lambda$ since $q_0 = 1$, we see that

$$\phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \phi(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \mathrm{o}(1)$$

from (Q2), (Q4), (Q5) and the Taylor's theorem, and the convergence is uniform over $(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ in a compact set. Define $\nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \eta_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$ and $\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = \tilde{\eta}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) + \phi(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$. First, note that

$$(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)}) = \underset{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\operatorname{argmin}} \nu_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (n^{1/2}\hat{\boldsymbol{\beta}}_\lambda^{(1)}, n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^{(2)} - \boldsymbol{\beta}^{*(2)})).$$

Next, because

$$\begin{aligned}\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) =& \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + \boldsymbol{s}_n^{(2)})\|_{\boldsymbol{J}^{(22)}}^2/2 \\ &+ \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{J}^{(1|2)}\boldsymbol{u}^{(1)}/2 - \boldsymbol{u}^{(1)\mathrm{T}}\boldsymbol{s}_n^{(1|2)} + \lambda\|\boldsymbol{u}^{(1)}\|_1 - \|\boldsymbol{s}_n^{(2)}\|_{\boldsymbol{J}^{(22)-1}}^2/2,\end{aligned}$$

we see that

$$(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) = \underset{(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})}{\operatorname{argmin}} \tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) = (\hat{\boldsymbol{u}}_n^{(1)}, \boldsymbol{J}^{(22)-1}(\boldsymbol{s}_n^{(2)} - \boldsymbol{J}^{(21)}\hat{\boldsymbol{u}}_n^{(1)}))$$

by using $\hat{\boldsymbol{u}}_n^{(1)}$ in (5.7). Now we apply Lemma 3 and evaluate the right-hand side in (4.2). First, it follows that $\Delta_n(\delta)$ converges in probability to 0 by the same way as in (4.13). Next, we can write

$$\begin{aligned}\tilde{\nu}_n(\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)}) - \tilde{\nu}_n(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)}) =& \|\boldsymbol{u}^{(1)} - \tilde{\boldsymbol{u}}_n^{(1)}\|^2_{\boldsymbol{J}^{(1|2)}}/2 + \lambda \sum_{j \in \mathcal{J}^{(1)}} (|u_j| - \gamma_j u_j) \\ &+ \|\boldsymbol{u}^{(2)} - \boldsymbol{J}^{(22)-1}(-\boldsymbol{J}^{(21)}\boldsymbol{u}^{(1)} + \boldsymbol{s}_n^{(2)})\|^2_{\boldsymbol{J}^{(22)}}/2 \qquad (6.11)\end{aligned}$$

by the same way as in the proof of Theorem 2, where $\boldsymbol{\gamma}$ is a $|\mathcal{J}^{(1)}|$-dimensional vector such that $\gamma_j = \mathrm{sgn}(\hat{u}_{n,j}^{(1)})$ when $\hat{u}_{n,j}^{(1)} \neq 0$ and $\gamma_j \in [-1, 1]$ when $\hat{u}_{n,j}^{(1)} = 0$. Let $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ be unit vectors such that $\boldsymbol{u}^{(1)} = \tilde{\boldsymbol{u}}_n^{(1)} + \zeta \boldsymbol{w}_1$ and $\boldsymbol{u}^{(2)} = \tilde{\boldsymbol{u}}_n^{(2)} + (\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2$, where $0 \leq \zeta \leq \delta$. Then, letting $\rho^{(22)}$ and $\rho^{(1|2)}$ (> 0) be half the smallest eigenvalues of $\boldsymbol{J}^{(22)}$ and $\boldsymbol{J}^{(1|2)}$, respectively, it follows that

$$\Upsilon_n(\delta) \geq \min_{0 \leq \zeta \leq \delta} \left\{ \rho^{(1|2)}\zeta^2 + \rho^{(22)} \|(\delta^2 - \zeta^2)^{1/2}\boldsymbol{w}_2 + \zeta \boldsymbol{J}^{(22)-1}\boldsymbol{J}^{(21)}\boldsymbol{w}_1\|^2 \right\} > 0$$

because the second term in (6.11) is non-negative. Hence the first term on the right-hand side in (4.2) converges to 0. In addition, because $(\boldsymbol{u}_n^{(1)}, \boldsymbol{u}_n^{(2)})$ and $(\tilde{\boldsymbol{u}}_n^{(1)}, \tilde{\boldsymbol{u}}_n^{(2)})$ are $\mathrm{O_p}(1)$ from $n^{1/2}$-consistency of $\hat{\boldsymbol{\beta}}_\lambda$, the second term on the right-hand side in (4.12) can be reduced to arbitrarily small by considering a sufficiently large ξ. Thus we have $\|\boldsymbol{u} - \tilde{\boldsymbol{u}}_n\| = \mathrm{o_p}(1)$, and as a consequence, we obtain (5.8) and (5.9).

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *Proceeding of the 2nd International Symposium on Information Theory*, eds. Petrov, B. N. and Csaki, F, Akademiai Kiado, 267–281.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10**, 1100–1120.

Bach, F. R., Lanckriet, G. R., and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm, In *Proceedings of the 21 international conference on Machine learning*, ACM, 6.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences*, **2**, 183–202.

Benzécri, J.-P. (1992). *Correspondence analysis handbook*, New York: Marcel Dekker.

Bishop, C. M. (2006). *Pattern recognition and machine learning*, New York: springer.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*, New York: Cambridge university press.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization, *Foundations of Computational mathematics*, **9**, 717–772.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, New York: Cambridge university press.

Dicker, L., Huang, B., and Lin, X. (2012). Variable selection and estimation with the seamless-L_0 penalty, *Statistica Sinica*, **23**, 929–962.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics*, **32**, 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **32**, 928–961.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood, *Journal of the Royal Statistical Society: Series B*, **75**, 531–552.

Fienberg, S. E. and Hinkley, D. V. (1980). *R. A. Fisher: An Appreciation*, Lecture Notes in Statistics, New York-Berlin: Springer-Verlag.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109–135.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, New York: Springer.

Hida, T. and Si, S. (2008). *Lectures on white noise functionals*, Singapore: World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.

Hjort, N. L. and Pollard, D. (1993). Asymptotics for minimisers of convex processes, *arXiv preprint arXiv:1107.3806*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender systems: an introduction*, New York: Cambridge University Press.

Katayama, K., Hagiwara, S., Tsutsui, H., Ochi, H., and Sato, T. (2010). Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis, In *Proceedings of the International Conference on Computer-Aided Design*, IEEE Press, 703–708.

Kiefer, J. (1961). Optimum designs in regression problems, II, *The Annals of Mathematical Statistics*, **32**, 298–325.

——— (1959). Optimum experimental designs, *Journal of the Royal Statistical Society. Series B*, **21**, 272–319.

Kiefer, J. and Wolfowitz, J. (1959). Optimum designs in regression problems, *The Annals of Mathematical Statistics*, **30**, 271–294.

Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. (2011). ℓ_p-norm multiple kernel learning, *The Journal of Machine Learning Research*, **12**, 953–997.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators, *The Annals of Statistics*, **28**, 1356–1378.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *The Annals of Statistics*, **39**, 2302–2329.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*, Springer Series in Statistics, New York: Springer, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The annals of mathematical statistics*, **22**, 79–86.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima, *Journal of Machine Learning Research*, **16**, 559–616.

Mallows, C. L. (1995). More comments on C_p, *Technometrics*, **37**, 362–372.

——— (1973). Some comments on C_p, *Technometrics*, **15**, 661–675.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties, *Journal of the American Statistical Association*, **106**, 1125–1138.

McCullagh, P. and Nelder, J. A. (1983). *Generalized linear models*, Monographs on Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.

Nagai, I. (2011). Modified C_p Criterion for Optimizing Ridge and Smooth Parameters in the MGR Estimator for the Nonparametric GMANOVA Model, *Open Journal of Statistics*, **1**, 1–14.

Ninomiya, Y. and Kawano, S. (2014). AIC for the LASSO in generalized linear models, In *ISM Research Memorandum*, **1187**.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *Advances in large margin classifiers*, **10**, 61–74.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, **7**, 186–199.

Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems, In *Biometrika*, ed. 51, **3/4**, 313–326.

Radchenko, P. (2005). Reweighting the lasso, In *2005 Proceedings of the American Statistical Association* [CD-ROM], Available at http://www-rcf.usc.edu/ radchenk/.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, Cambridge, MA: MIT Press.

Rockafellar, R. T. (1970). *Convex Analysis*, Princeton Mathematical Series, New York: Princeton university press.

——— (1976). Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Mathematics of Operations Research*, **1**, 97–116.

Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*, Cambridge, MA: MIT press.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*, New York: Cambridge university press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution, *The Annals of Statistics*, **9**, 1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society: Series B*, **36**, 111–147.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's, *Communications in Statistics-Theory and Methods*, **7**, 13–26.

Sun, S. and Li, X. (2014). Fast statistical analysis of rare circuit failure events via subset simulation in high-dimensional variation space, In *Computer-Aided Design (ICCAD), 2014 IEEE/ACM International Conference on*, IEEE, 324–331.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

Tong, S. and Koller, D. (2002). Support vector machine active learning with applications to text classification, *The Journal of Machine Learning Research*, **2**, 45–66.

Umezu, Y., Matsuoka, H., Ikeda, H., and Ninomiya, Y. (2015). Defect rate evaluation via simple active learning, *Pacific Journal of Mathematics for Industry*, **7**, 1–8.

Umezu, Y. and Ninomiya, Y. (2015). AIC for Non-concave Penalized Method, *arXiv preprint arXiv:1509.01688*.

——— (2013). Optimal experimental design based on Gaussian process classification (in Japanese), In *Proceedings of the Japanese Joint Statistical Meeting*, 327.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, **94**, 553–568.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society: Series B*, **71**, 671–683.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model, *Biometrika*, **94**, 19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion, *Journal of the American Statistical Association*, **105**, 312–323.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso, *The Annals of Statistics*, **35**, 2173–2192.