中央大学博士論文

# Bayesian Sparse Regression Modeling

Ibuki HOSHINA
保科　架風

博士（理学）

中央大学大学院
理 工 学 研 究 科
数　学　専　攻

平成 28 年度
2017 年 3 月

# Acknowledgments

To complete this doctor thesis, I have received a lot of support. I am deeply grateful to my supervisor, Prof. Sadanori Konishi, for providing innumerable advice and tutelage. Although I have been not an allegiant student, he had kept taking care of me. Further, he gave me a lot of chances to grow me. Thanks to him, I could join a lot of academic meetings. Such opportunities gave me much stimulus for my researches. Without his support, this doctor thesis must not have been possible.

I would like to show my greatest appreciation to Prof. Fumitake Sakaori, for providing invaluable advice and constant encouragement. He is my best adviser and gave me an academic freedom. He permitted me to challenge at a lot of situations, and I could decide with his support when I was troubled over a choice. The opportunity he gave me enabled me to obtain another world.

It gives me great pleasure to acknowledge to the encouragement, suggestions, and wisdom of my co-supervisors; Prof. Yoshikazu Kobayashi, Prof. Toshinari Kamakura, and Prof. Yoshinori Kawasaki of the Institute of Statistical Mathematics.

I would like to express my gratitude to Prof. Akimichi Takemura of the Shiga University, Prof. Yasuto Yoshizoe of the Aoyama Gakuin University, and the other professors of the Japan Statistical Society Certificate and the Japanese Inter-university Network for Statistical Education, for their kind assistance. They cared for my doctor degree and supported my livelihood.

I would like to offer my special thanks to Prof. Yasunori Fujikoshi of the Hiroshima University, Prof. Kanta Naito of the Shimane University, and the other professors in the statistical society, for giving me valuable advice. Prof. Fujikoshi, he is my first professor, taught me what is the research. Prof. Naito had kept

giving some encouragement to me.

I also would like to express many thanks to professors, staffs, senior associates, and the other members in the Chuo University for their support.

I would like to thank to Prof. Takuma Yoshida of the Kagoshima University, Prof. Masayo Hirose of the Institute of Statistical Mathematics, Dr. Daeju Kim of SoftBank Corp, Prof. Heewon Park of the Yamaguchi University, and the other peers, for their encouragement. A great time with them helped me to take my own way. Thanks to them I could research with philosophy.

I am grateful to Prof. Shuichi Kawano of the University of Electro-Communications, Prof. Kei Hirose of the Kyushu University, Prof. Tetsuto Himeno of the Shiga University, and the other senior associates of the Konishi laboratory, for their great deal of encouragement and advice. I was helped by their advice a lot, and they also gave me a motivation to research.

I would like to thank the all members of the Konishi Laboratory and the Sakaori Laboratory, for giving a lot of finds.

I also would like to thank to the members of the IREP Co., Ltd for giving me the opportunity to bring the unacademic life. The days that I have spent in the IREP are irreplaceable. I believe that the experience of those days expanded the possibilities for my life.

Furthermore, I would like to thank my family for their constant support and enormous encouragement.

Finally, my deepest appreciation goes to the late Prof. Fumiyuki Momose for his help. I believe that no my doctor's course without his help, from bottom of my heart.

# Contents

# Chapter 1

# Introduction

The advancements of the computer and sensor technology have enabled us to get and save the high-dimensional, complex, and huge data in various fields of natural and social sciences such as biotechnology, bioinformatics, system engineering, marketing, and information technology. The statistical modeling plays an important role for extracting useful information and knowledge from the data. The linear regression modeling is used to model a linear relationship between a response variable and several explanatory variables, and it represents the mechanisms of phenomena by linear combinations of explanatory variables. The model tells us various things; which variables have larger influence on the phenomena or which have no influence on them.

The estimation of regression parameters and variable selection are fundamentally important in the linear regression modeling. The parameter estimation corresponds to the estimation of the amount of the impact of the factors for the phenomena, and the variable selection corresponds to the selection of the factors, respectively. The parameters are usually estimated by using the ordinary least squares or maximum likelihood procedures. Variable selection follows the best subset selection based on the model selection criteria such as the AIC (Akaike, 1973) and the BIC (Schwarz, 1978). The cross-validation is also widely used as a model selection criterion. For model selection criteria, we refer to Konishi and Kitagawa (2008). For high-dimensional regression, however, these modeling procedures lead models with poor prediction accuracy. The least square procedures

often yield model estimates with large variances, especially when there is a problem of multicollinearity. The best subset selection is often unstable because of its inherent discreteness (Breiman, 1996). Further, the computational costs of the parameter estimation and the model evaluation complicate the modeling because we need to calculate the inverse matrix of high-dimensional matrix and the number of the candidate models are vast when the dimension of the data increases.

In order to overcome these issues, Tibshirani (1996) proposed the lasso (least absolute shrinkage and selection operator), which tends to shrink some regression coefficients toward exactly zero by imposing an $L_1$ norm penalty on regression coefficients. A distinctive feature of the lasso is its capability for simultaneous model estimation and variable selection. The modeling procedures via the $L_1$ norm regularization is called the "sparse regression modeling" because they can produce sparse estimates of the regression coefficients. For the last 20 years, various sparse regression procedures which are inspired by the lasso have been proposed; e.g. the bridge regression (Frank and Friedman, 1993), the SCAD (smoothly clipped absolute deviation; Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), the adaptive lasso (Zou, 2006), the group lasso (Yuan and Lin, 2006) and the MCP (minimax concave penalty; Zhang, 2010).

Although the least square or the maximum likelihood procedures give us the closed form of the estimators of regression coefficients, analytical derivation of the estimators for $L_1$ regularizations is difficult, since $L_1$ penalty is non-differentiable at the origin. For this problem, several efficient algorithms have been proposed to solve the $L_1$ regularizations. Fu (1998) proposed the shooting algorithm for the bridge regression, and the coordinate descent algorithm (Friedman *et al.* , 2010) is an improved of the shooting algorithm. Mazumder *et al.* (2011) proposed the SparseNet, which is an extension of the coordinate descent algorithm for the non-convex optimization. The development of the LARS algorithm (Efron *et al.* , 2004) touched off the growth of the area of the $L_1$ regularizations. The GPS algorithm (Friedman, 2012) is also known procedure for these problems. For non-convex regularizations such as the bridge, the SCAD, and the MCP, the local quadratic approximation (LQA; Fan and Li, 2001) and the local linear approximation (LLA;

Zou and Li, 2008) are proposed.

In sparse regression modeling, the selection of adjusted tuning parameters including in the $L_1$ norm penalty is a crucial issue since these procedures depend on the values of tuning parameters that control the bias-variance trade-off in resulting estimates. Tuning parameters also identify a set of variables included in a model. Ordinary model selection criteria, such as the AIC and the BIC, are often hard to evaluate the goodness of estimated models. Although the AIC and the BIC are the consistent estimators of the Kullback Leiblar information and logarithms of the marginal likelihoods, respectively, when the regression coefficients and the error variance are estimated by the maximum likelihood procedure, these criteria have the estimation bias. In the estimation bias, the degrees of freedom (e.g. Ye, 1998; Efron, 1986; Efron, 2004) is often used to quantify the model complexity, and it plays a key role in model selection. In the lasso, Efron (2004) showed that Mallows' $C_p$ type criteria (Mallows, 1973) are unbiased estimators of the true prediction error when degrees of freedom is given, and often provide better accuracy than the cross-validation. It is, however, difficult to derive a closed form of the degrees of freedoms of the sparse regression modelings. For this problems, estimators of the degrees of freedom of the lasso have been integrated by Zou *et al.* (2007), Kato (2009), Tibshirani and Taylor (2012) and Hirose *et al.* (2013). Especially Zou *et al.* (2007) showed that the number of non-zero estimates for regression coefficients is an unbiased estimator of the degrees of freedom of the lasso.

The regularization procedures have the relationship with the Bayes model. The Bayes model is one of the statistical modeling techniques, and its fundamental characteristic is in evaluating the posterior probability distribution. In non-Bayes modeling, the estimation of the model does through the evaluating of the likelihood or the loss functions. On the other hand, the Bayes modeling evaluates the posterior probability derived from a product of the likelihood and the prior probability. The regularization procedures are formed as the combination of the loss function and a penalty term, and we can interpret it as the Bayes model (the loss function and the penalty term correspond to the likelihood and the prior, respectively). The GBIC (Konishi *et al.* , 2004) used as a criterion for evaluating

models estimated by regularization methods, have been proposed from a Bayesian viewpoint. Tibshirani (1996) indicated that the lasso estimates can be interpreted as a MAP (maximum a posteriori) estimates when the regression coefficients have independent and identical Laplace prior and the likelihood is taken to be normal linear regression model. The Bayesian lasso (Park and Casella 2008, Hans 2009) is a fully Bayesian analysis, and they suggested Gibbs sampling for the lasso with Laplace prior in the hierarchical model. The Bayesian lasso provides the Bayesian credible intervals of the lasso, and it guides the variable selection.

Compared to non-Bayesian modeling, the Bayesian lasso also has two advantages:

1. estimating error variance.
2. choosing the value of tuning parameter.

In the lasso, the estimate of error variance is not directly obtained, and efficient procedures were studied (see e.g. Reid *et al.*, 2014). On the other hand, the Bayesian lasso determines it as mode, median, or mean of posterior. Tuning parameters which can be viewed as the Bayesian hyper parameters, are estimated by hierarchical or empirical Bayesian method,

The Bayesian lasso has two drawbacks: it is difficult to calculate the posterior mode of regression coefficients, and the resulting regression coefficients are not sparse. Although the posterior mode of the Bayesian lasso coefficients is equivalent to the lasso estimates, it is difficult to calculate the posterior mode because the posterior function is not differentiable at zero. Kernel density estimation may be applicable for this problem. It is however difficult to calculate a stable posterior mode in high-dimensional density estimation. Furthermore, Park and Casella (2008) indicate that the Bayesian lasso (point) estimates for regression coefficients do not take zero value exactly.

To overcome these drawbacks, we propose three new methodologies:

A. The sparse algorithm (Hoshina, 2012).
B. aPIC: New model selection criterion that evaluates a Bayesian predictive distribution (Kawano *et al.* , 2015).

C. The MAP Bayesian lasso (Hoshina, 2015).

## A. Sparse algorithm

The lack of the sparsity of the Bayesian lasso estimates is cased by the model estimation, using MCMC or the Gibbs sampling. Since the estimates are calculated by the random sample from the posterior, it is hard to take zero values as the estimates. To overcome this issue, Hoshina (2012) proposed the sparse algorithm that gives exactly zero values for some of the estimated coefficients according to the posterior probability.

## B. aPIC

Park and Casella (2008) proposed the method to select the value of the tuning parameter taking an empirical Bayes approach. Hans (2010) proposed a variable selection procedure that can model uncertainty based on the marginal likelihood. We propose a new model selection criterion for evaluating a Bayesian predictive distribution of the Bayesian lasso, which is used to choose appropriate values of hyper-paramters included in a prior.

## C. MAP Bayesian lasso

It is hard to derive the MAP estimates of the Bayesian lasso because of the non-differentiability of the posterior function. For this problem, we propose a new methodology that approximates posterior function by Monte Carlo integration; estimating the posterior mode by Newton's method, and modifying the resulting estimates of regression coefficients to be sparse along a posterior probability.

The remainder of this thesis is organized as follows:

- Chapter 2 introduces $L_1$ regularization procedures including the ridge, the lasso, the elastic net, the adaptive lasso, the group lasso, the bridge regression, the SCAD, and the MCP for the linear regression modeling. Especially,

the oracle property that is the asymptotic property of the sparse modeling is provided. We describe the estimation algorithms for the $L_1$ and the non-convex regularizations; the LARS, the coordinate descent algorithm, the LQA, and the LLA. The degrees of freedom of the $L_1$ regularizations are introduced and the algorithm which calculates the degrees of freedom of the LARS are described. A number of the $L_1$ regularization procedures are compared in terms of the sparsity.

- In Chapter 3, we presents a review of the Bayes-type $L_1$ regularizations, the Bayesian lasso and its extensions. Some important properties on the Bayesian lassos are described, and the unimodality of several Bayes-type $L_1$ regularizations is shown.
- Chapter 4 introduces new procedures for the sparse regression modeling via the Bayesian lasso: an algorithm to correct the resulting regression coefficients as sparse, a model selection criterion for the selection of appropriate values of hyper-paramters included in a prior distribution of the Bayesian lasso, and a new sparse modeling procedure which based on the MAP estimation of the Bayesian lasso.
- In Chapter 5, we introduce the Bayesian information criteria; the BIC and the GBIC, and a new model selection criterion for the elastic net is introduced. This procedure evaluates the approximated marginal likelihood of the elastic net or the Bayesian elastic net.
- Chapter 6 presents numerical studies to investigate the proposed procedures through Monte Carlo simulations and the analyses of artificial and real data sets.
- Chapter 7 gives some concluding remarks.

# Chapter 2

# Lasso and $L_1$ regularizations

## 2.1 Background

We consider the linear regression model

$$\boldsymbol{y} = \beta_0 \mathbf{1}_n + X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an $n$-dimensional response vector, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ design matrix, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the $p$-dimensional observations for predictor variables, the elements of $\boldsymbol{x}_i$ are given as $x_{i1}, \ldots, x_{ip}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional regression coefficient vector, $\mathbf{1}_n$ is an $n$-dimensional vector whose all components are one, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector. It is assumed that the elements of $\boldsymbol{\varepsilon}$ are independent and identically distributed according to a normal distribution with mean zero and unknown variance $\sigma^2$. Without loss of generality, we assume that the predictors are standardized:

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad j = 1, \ldots, p. \tag{2.2}$$

The linear regression model is usually fitted by the ordinary least squares procedure (OLS) or the maximum likelihood estimator (MLE). The OLS estimates of

$\beta_0$ and $\boldsymbol{\beta}$ are obtained by minimizing the sum of squared error

$$\begin{aligned} R(\beta_0, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \\ &= (\boldsymbol{y} - \beta_0 \mathbf{1}_n - X\boldsymbol{\beta})^T (\boldsymbol{y} - \beta_0 \mathbf{1}_n - X\boldsymbol{\beta}). \end{aligned} \tag{2.3}$$

Differentiating with respect to $\beta_0$ and $\boldsymbol{\beta}$ we obtain

$$\begin{aligned} &\frac{\partial}{\partial \beta_0} R(\beta_0, \boldsymbol{\beta}) = -2(\boldsymbol{y}^T \mathbf{1}_n - \beta_0), \quad \frac{\partial^2}{\partial \beta_0^2} R(\beta_0, \boldsymbol{\beta}) = 2, \\ &\frac{\partial}{\partial \boldsymbol{\beta}} R(\beta_0, \boldsymbol{\beta}) = -2\boldsymbol{X}^T (\boldsymbol{y} - X\boldsymbol{\beta}), \quad \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} R(\beta_0, \boldsymbol{\beta}) = 2X^T X. \end{aligned} \tag{2.4}$$

Assuming that $X$ has a full column rank ($X^T X$ is positive definite), we have the normal equation

$$-2(\boldsymbol{y}^T \mathbf{1}_n - \beta_0) = 0, \quad -2\boldsymbol{X}^T (\boldsymbol{y} - X\boldsymbol{\beta}) = \mathbf{0}. \tag{2.5}$$

Thus the OLS estimates of $\beta_0$ and $\boldsymbol{\beta}$ are given by

$$\hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}. \tag{2.6}$$

Since the error vector $\boldsymbol{\varepsilon}$ has an $n$-dimensional normal distribution $N_n(\mathbf{0}, \sigma^2 I_n)$, we also have the likelihood function for the response vector $\boldsymbol{y}$ in the form

$$\begin{aligned} p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right\} \\ &= N_n(\boldsymbol{y}|\beta_0 \mathbf{1}_n + X\boldsymbol{\beta}, \sigma^2 I_n), \end{aligned} \tag{2.7}$$

where $N_q(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma)$ is a probability density function of a $q$-dimensional normal distribution with variable $\boldsymbol{z}$, the mean vector $\boldsymbol{\mu}$ and the variance covariance matrix $\Sigma$, and $I_n$ is an $n \times n$ identity matrix.

This leads to the log-likelihood function

$$\log p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - \beta_0\mathbf{1}_n - X\boldsymbol{\beta}\|^2. \tag{2.8}$$

Thus, the MLEs for $\beta_0$ and $\boldsymbol{\beta}$ in model (2.1) are defined by

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{\beta_0, \boldsymbol{\beta}}{\operatorname{argmax}} \left[ -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \beta_0\mathbf{1}_0 - X\boldsymbol{\beta}\|^2 \right]. \tag{2.9}$$

The maximizer of (2.9) is equivalent to the minimizer (2.6), and the OLS and the MLE for $\beta_0$ and $\boldsymbol{\beta}$ have the same values as in the Gaussian models.

## 2.2 Estimation accuracy and ridge regression

In the OLS or the MLE procedures, the mean vector and the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ are respectively given by

$$\mathrm{E}\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}, \quad \mathrm{Cov}\left[\hat{\boldsymbol{\beta}}\right] = \sigma^2(X^TX)^{-1}. \tag{2.10}$$

This means that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ and the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ depends on $X^TX$. When some column elements of $X$ are highly correlated, the determinant of $X^TX$ decreases ($X^TX$ is close to singular) and the diagonal elements of $(X^TX)^{-1}$ become extremely large (this is the multicollinearity problem). This problem is also encountered in the high-dimensional case. Even when the true variance covariance matrix is an identity matrix (i.e., correlations between any predictors are sufficiently small), the determinant of the variance covariance matrix $|S|$ ($S = X^TX/n$) can be small (even if the sample size is sufficient). Fig. 2.1 shows that $S$ or $X^TX$ approaches to singularity as the dimension increases.

The OLS procedures often yield poor prediction because of the large variance of the estimator. The regularization techniques overcome this problem. The ridge regression introduced by Hoerl and Kennard (1970) is known as one of the regularization procedures. The ridge estimates are defined by minimizing $R(\beta_0, \boldsymbol{\beta})$

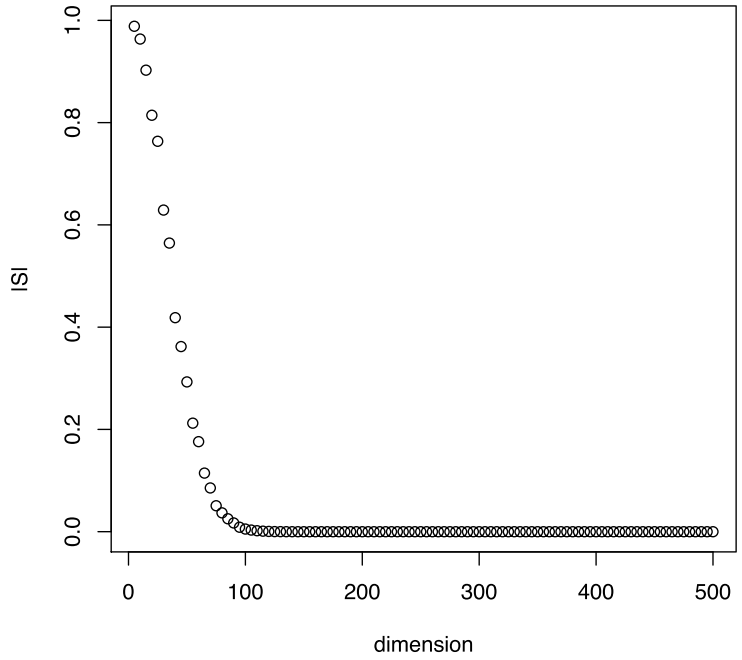



Fig. 2.1 The relationship between dimensionality and determinant of the variance covariance matrix: The variance covariance matrix $S$ is computed by 1000 random samples from $N_p(\mathbf{0}_p, I_p)$.

with a bound on the $L_2$ norm of the coefficients:

$$\begin{aligned}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := &\underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} R(\beta_0, \boldsymbol{\beta}), \\ &\text{subject to} \quad \sum_{j=1}^{p} \|\beta_j\|^2 \leq t,\end{aligned} \tag{2.11}$$

or equivalently,

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta})}{\operatorname{argmin}} R(\beta_0, \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \|\beta_j\|^2, \tag{2.12}$$

where a tuning parameter $t$ and a regularization parameter $\lambda$ ($t, \lambda \geq 0$) control the degrees of shrinkage. The ridge shrinks coefficients $\boldsymbol{\beta}$ toward $\mathbf{0}$ as $t$ decreases or $\lambda$ increases, although $\hat{\beta}_0$ is $\bar{y}$ for any $t$ and $\lambda$ (without loss of generality, we can assume that $\bar{y} = 0$ and hence we omit $\beta_0$ from the model for convenience).

The ridge estimator is given by

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T \boldsymbol{y}, \tag{2.13}$$

and there is a one-to-one correspondence between $t$ and $\lambda$:

$$\lambda = \frac{1}{t}(X\hat{\boldsymbol{\beta}}^{\text{ridge}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{ridge}}). \tag{2.14}$$

The mean vector and the variance covariance matrix of the ridge estimator are respectively given by

$$\begin{aligned} \mathrm{E}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}}\right] &= (X^T X + \lambda I_p)^{-1} X^T X \boldsymbol{\beta} \\ \mathrm{Cov}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}}\right] &= \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1}. \end{aligned} \tag{2.15}$$

(2.15) indicates that although the ridge estimator is not an unbiased estimator of $\boldsymbol{\beta}$, the ridge estimator has a smaller variance than OLS does, that is, the ridge is more stable than the OLS. If $X^T X$ is singular, $X^T X + \lambda I_p$ remains nonsingular taking an appropriate value of $\lambda > 0$. The cause of instability of the OLS is the singular or near singular matrix $X^T X$, and it often appears when there are a set of highly correlated predictors or high dimensionality. Therefore, the ridge performs a more stable estimation and achieves a better prediction accuracy than the OLS does in such cases.

## 2.3 $L_1$ regularization

### 2.3.1 Lasso

Frank and Friedman (1993) extended the ridge for the $L_q$ regularization called the bridge regression. The bridge estimator is given by

$$\hat{\boldsymbol{\beta}}^{\text{bridge}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2, \quad \text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t, \tag{2.16}$$

or,

$$\hat{\boldsymbol{\beta}}^{\text{bridge}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q, \tag{2.17}$$

where $q > 0$. The bridge regression includes the ridge with $q = 2$ as a special case. If we take $q = 0$, we have the following optimization of least square loss and $L_0$ penalty term:

$$\hat{\boldsymbol{\beta}}^{\text{L}_0} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} I(|\beta_j| > 0), \tag{2.18}$$

where $I(\cdot)$ is the indicator function. $L_0$ bridge produces a parsimonious model (i.e., some coefficients are estimated to be exactly zero) because it penalizes the number of non-zero coefficients in the model and performs as a variable selection procedure. However, it is hard to solve this $L_0$ optimization problem because of its non-convexity. It is known that the $L_0$ optimization is equivalent to the "subset selection" based on some information criteria. If we set $\lambda$ in (2.18) to be proportional to some constant or $\log n$, the $L_0$ optimization can be reckoned as the subset selection based on the AIC or BIC, respectively, and known as traditional model selection procedures.

Although variable selection enables us to improve prediction performance and helps us to interpret the fitted model, an increase in the number of predictors hinder the application of subset selection. Further, high dimensionality adversely affects the prediction accuracy of the resulting model. In such cases, the subset selection is computationally expensive because it needs to choose the most moderate combination of predictors (the number of the candidate model is $2^p$). In addition, the subset selection often becomes extremely variable because of its inherent discreteness. Since predictors are either retained or dropped from the model, the prediction accuracy of the resulting model becomes poor (Breiman 1996).

For these problems, the efficiency of the lasso (Least Absolute Shrinkage and Selection Operator; Tibshirani 1996) is well known. The lasso minimizes the least square loss subject to the sum of the absolute values of the coefficients ($L_1$ norm

of the coefficient vector) being less than a constant $t$,

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} := \underset{\boldsymbol{\beta}}{\text{argmin}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2, \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \tag{2.19}$$

Further, the lasso estimates have the Lagrangian form with $L_1$ penalty

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} := \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.20}$$

and they correspond to the case of the bridge with $q = 1$.

The lasso continuously shrinks the coefficients toward zero as $\lambda$ increases. In the case of $\lambda = 0$ and $n > p$, $\hat{\boldsymbol{\beta}}$ is equivalent to the OLS or the MLE. Further, $\hat{\boldsymbol{\beta}}$ becomes sparse, that is, some coefficients are shrunk to exactly zero when the scale of $\lambda$ is sufficiently large, because of the nature of the $L_1$ penalty.

In the case of $X^T X = I_p$ (i.e., $X$ is orthonormal),

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{lasso}} &= \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ -2\boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ -2\hat{\boldsymbol{\beta}}^T \boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{\beta} + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{j=1}^{p} \left[ -2\hat{\beta}_j \beta_j + \beta_j^2 + \lambda |\beta_j| \right] \\
&= \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ \sum_{\beta_j \geq 0} \left\{ \beta_j^2 - (2\hat{\beta}_j - \lambda)\beta_j \right\} + \sum_{\beta_j < 0} \left\{ \beta_j^2 - (2\hat{\beta}_j + \lambda)\beta_j \right\} \right],
\end{aligned} \tag{2.21}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ is the OLS estimate of $\boldsymbol{\beta}$. If $\beta_j \geq 0$ and $2\hat{\beta}_j - \lambda \leq 0$, $\text{argmin}_{\beta_j} \beta_j^2 - (2\hat{\beta}_j - \lambda)\beta_j = 0$. If $\beta_j \leq 0$ and $2\hat{\beta}_j + \lambda \geq 0$, then $\text{argmin}_{\beta_j} \beta_j^2 +$

$(2\hat{\beta}_j - \lambda)\beta_j = 0$. Further,

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\beta_j^2 - (2\hat{\beta}_j - \lambda)\beta_j\Big|_{\beta_j=\hat{\beta}_j^{\text{lasso}}} &= 2\hat{\beta}_j^{\text{lasso}} - (2\hat{\beta}_j - \lambda) = 0, \\
&\qquad \text{if } \hat{\beta}_j^{\text{lasso}} \geq 0 \text{ and } 2\hat{\beta}_j - \lambda > 0, \\
\frac{\partial}{\partial \beta_j}\beta_j^2 - (2\hat{\beta}_j + \lambda)\beta_j\Big|_{\beta_j=\hat{\beta}_j^{\text{lasso}}} &= 2\hat{\beta}_j^{\text{lasso}} - (2\hat{\beta}_j + \lambda) = 0, \\
&\qquad \text{if } \hat{\beta}_j^{\text{lasso}} \leq 0 \text{ and } 2\hat{\beta}_j + \lambda < 0.
\end{aligned} \tag{2.22}$$

Thus, we can see that the lasso estimates in the orthonormal case are given by

$$\hat{\beta}_j^{\text{lasso}} = \text{sign}(\hat{\beta}_j) \cdot \left(|\hat{\beta}_j| - \frac{\lambda}{2}\right)_+, \tag{2.23}$$

where

$$(x)_+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}. \tag{2.24}$$

That is, if $\lambda$ is sufficiently large, then some coefficients of the lasso are shrunk to exactly zero. In more general cases, since the $L_1$ penalty is not differentiable at $\beta_j = 0$, the lasso estimates are not analytically derived. Several efficient algorithms have been proposed to compute the lasso estimates, and these are discussed in Section 2.4.

### Comparison between the lasso and the ridge

In the orthonormal case, the ridge estimates are given by

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}, \tag{2.25}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimate vector. The ridge and lasso estimates given by (2.23) are compared in Fig. 2.2. This shows the following:

- The ridge yields a proportional shrinkage and the lasso translates each coefficient by a constant truncating at zero.

- The lasso estimates often can take zero values and the ridge takes zero only when the OLS is zero (i.e., "non-sparsity" of the ridge).

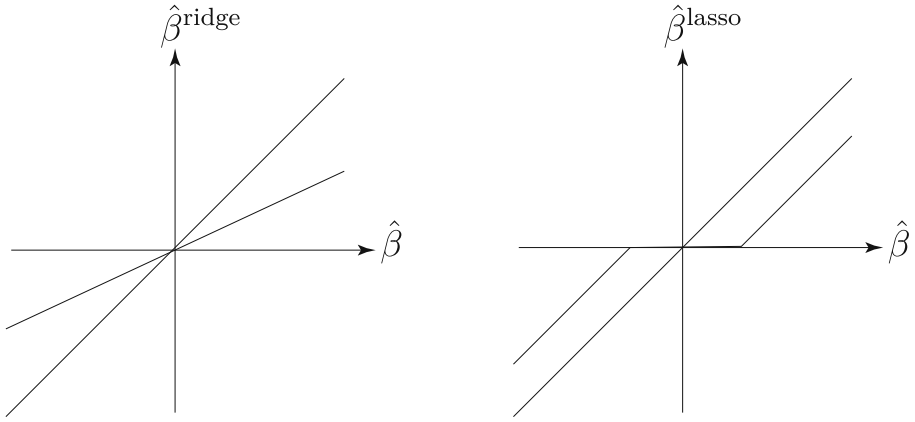



Fig. 2.2 The relationship between the ridge (left) and the lasso (right) estimates in orthonormal case: Real lines indicate the estimates and dashed lines represent $\hat{\beta}^{\text{ridge}} = \hat{\beta}$ or $\hat{\beta}^{\text{lasso}} = \hat{\beta}$, where $\hat{\beta}$ is the OLS estimates.

In the non-orthonormal case, the elementary differential geometry helps us to show the non-sparsity of the ridge. The minimizing problem of the least square loss function $\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$ can be interpreted as minimizing the quadratic function

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \tag{2.26}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ denotes the OLS estimates, which implies the ellipsoid contour for fixed loss $\ell$. Thus, the ridge estimates and the lasso estimates can be interpreted as the points where the ellipsoids hit the sphere $\sum_j^p \|\beta_j\|^2 = t$ and the cube $\sum_j^p |\beta_j| = t$, respectively, where $t$ is some fixed value (Fig. 2.3).

At the ridge or the lasso estimates in Fig. 2.3, the ellipsoids and the ridge sphere or the lasso cube have the same tangent plane. Generally, the tangent plane of a curved surface $\mathcal{S} := \{\boldsymbol{x} = (x_1, ..., x_p) \in \mathbb{R}^p;\ G(\boldsymbol{x}) = 0\}$ has the equation

$$\sum_{j=1}^{p} \frac{\partial}{\partial x_j} G(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{x}^*} (x_j - x_j^*) = 0 \tag{2.27}$$

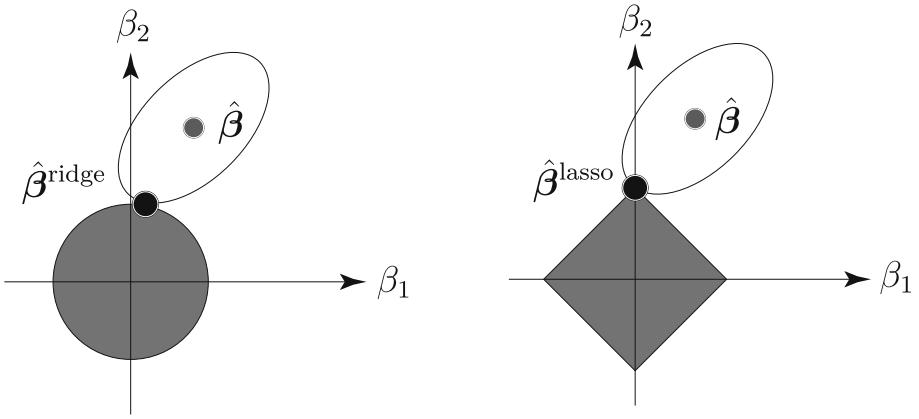



Fig. 2.3 Graphical representation of the ridge (left) and the lasso (right) estimates: The dark grey sphere and cube mean the areas $\sum_j^p \|\beta_j\|^2 \leq t$ and $\sum_j^p |\beta_j| \leq t$, respectively. Both estimates are the points where the loss ellipsoids hit the penalty sphere or cube.

at the point $\boldsymbol{x}^* = (x_1^*, ..., x_p^*) \in \mathcal{S}$. In matrix form, we have

$$\left(\frac{\partial G(\boldsymbol{x}^*)}{\partial \boldsymbol{x}}\right)^T (\boldsymbol{x} - \boldsymbol{x}^*) = \mathbf{0}. \tag{2.28}$$

Let $F(\boldsymbol{\beta})$ be

$$F(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) - \ell = 0, \tag{2.29}$$

then, we have the tangent plane of the loss ellipsoid at $\boldsymbol{\beta}^*$ as the following:

$$\begin{aligned}\left(\frac{\partial F(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\right)^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*) &= \left\{X^T X(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})\right\}^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &= (\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})^T X^T X (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &= 0.\end{aligned} \tag{2.30}$$

The tangent planes for the ridge sphere at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ are given by

$$\left(\left.\frac{\partial \boldsymbol{\beta}^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\right)^T (\boldsymbol{\beta} - \boldsymbol{\beta}^*) = 2\boldsymbol{\beta}^{*T}(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \tag{2.31}$$

In $p = 2$, let $\boldsymbol{\beta}^* = (0, \sqrt{t})^T$ be the ridge estimates and $S = X^T X$ have the components

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}. \tag{2.32}$$

Then, we have

$$\begin{aligned} \begin{pmatrix} -\hat{\beta}_1 \\ \sqrt{t} - \hat{\beta}_2 \end{pmatrix}^T \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 - \sqrt{t} \end{pmatrix} = 0, \quad {}^\forall \beta_1, \beta_2 \in \mathbb{R}, \\ 2 \begin{pmatrix} 0 \\ \sqrt{t} \end{pmatrix}^T \begin{pmatrix} \beta_1 \\ \beta_2 - \sqrt{t} \end{pmatrix} = 0, \quad {}^\forall \beta_1, \beta_2 \in \mathbb{R}, \end{aligned} \tag{2.33}$$

and

$$\begin{aligned} &\begin{pmatrix} -\hat{\beta}_1 \\ \sqrt{t} - \hat{\beta}_2 \end{pmatrix}^T \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} \\ &\qquad = \begin{pmatrix} -\hat{\beta}_1 \\ \sqrt{t} - \hat{\beta}_2 \end{pmatrix}^T \begin{pmatrix} c_{11} \\ c_{21} \end{pmatrix} \beta_1, \quad {}^\forall \beta_1 \in \mathbb{R}. \end{aligned} \tag{2.34}$$

From the above, the OLS $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$ satisfies

$$-(c_{11}\hat{\beta}_1 + c_{21}\hat{\beta}_2 - c_{21}\sqrt{t})\beta_1 = 0, \quad {}^\forall \beta_1 \in \mathbb{R}. \tag{2.35}$$

This means that the ridge estimates have a zero component when the OLS exists on the line such that

$$c_{11}\hat{\beta}_1 + c_{21}\hat{\beta}_2 - c_{21}\sqrt{t} = 0, \tag{2.36}$$

however, this is an event with probability zero (Fig. 2.4).

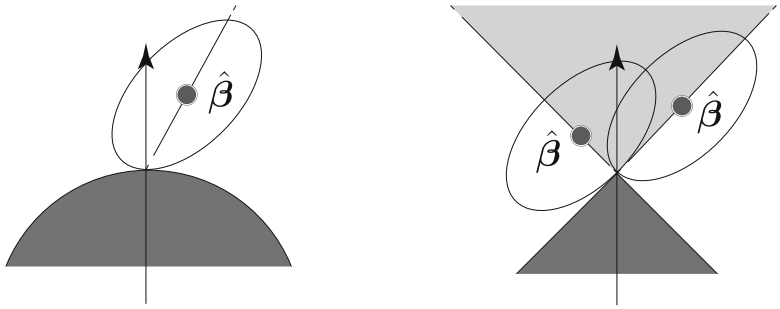



Fig. 2.4 The conditions in which the ridge and lasso estimates become sparse: The ridge estimates (left) have zero component only when the OLSs on the dashed line but this event has zero probability. The lasso estimates (right) are sparse when the OLS exists on the light grey area. This is why the ridge has no sparsity but the lasso does.

On the other hand, the tangent planes of the lasso cube at $\boldsymbol{\beta}^* = (0, t)^T$ are not specified uniquely, which is given by

$$s\beta_1 + \beta_2 - t = 0, \quad s \in [-1, 1]. \tag{2.37}$$

From (2.30) and (2.37), we have

$$\begin{pmatrix} -\hat{\beta}_1 \\ t - \hat{\beta}_2 \end{pmatrix}^T \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 - t \end{pmatrix} = 0, \quad {}^\forall\beta_1, \beta_2 \in \mathbb{R}. \tag{2.38}$$

Further we have the equation

$$-\{(c_{11} - sc_{12})\hat{\beta}_1 + (c_{21} - sc_{22})\hat{\beta}_2 - t(c_{21} - sc_{22})\}\beta_1 = 0, \quad {}^\forall\beta_1 \in \mathbb{R}, \tag{2.39}$$

for fixed $s \in [-1, 1]$. That is, the lasso estimates have a zero component if the OLS exists in the area that satisfies (2.39) (Fig. 2.4).

### 2.3.2 Elastic net

The lasso enables us to do both continuous shrinkage and automatic variable selection simultaneously. However, some limitations of the lasso have been pointed out. Zou and Hastie (2005) mentioned the following limitations of the lasso :

- In the case of $p > n$, the lasso only takes in at most $n$ predictors into the model because the nature of the convex optimization. (However, some algorithms that solve the lasso optimization take in more than $n$ predictors into the model because they approximate the optimization problems.)
- If there are groups of predictors among which the pairwise correlations are very strong, the lasso takes in only one predictor from the groups.
- In the case of $n > p$, as Tibshirani (1996) suggested, the prediction performance of the lasso often is dominated by the ridge if there are high correlations between predictors.

These limitations indicate that the lasso leads to poor prediction and model selection accuracy in high-dimensional or highly correlated situations. To overcome these drawbacks, Zou and Hastie (2005) proposed an $L_1 + L_2$ type regularization procedure, called the "elastic net". Similar to the lasso, the elastic net does sparse estimation, but in contrast to the lasso, it takes in all members from the group of the highly correlated predictors into the model.

The elastic net for the linear regression model are given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{EN}} := (1+\lambda_2) \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right], \tag{2.40}$$

where $\lambda_1, \lambda_2 (> 0)$ are the tuning parameters that control the strength of the $L_1$ or $L_2$ penalties. The elastic net includes the lasso with $\lambda_2 = 0$ as a special case.

Further, the elastic net has the conditional optimization form:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\mathrm{EN}} := (1+\lambda_2) \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \right], \\ \text{subject to } \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \le t, \end{aligned} \tag{2.41}$$

where $t (> 0)$ and $\alpha \in [0, 1]$ are the tuning parameters. Thus, it is shown that the elastic net estimates can be interpreted as the point where the squared loss ellipsoid hits the shape $\alpha \sum_j^p |\beta_j| + (1-\alpha) \sum_j^p \beta_j^2 = t$, where $t$ and $\alpha$ are some fixed values (Fig. 2.5).

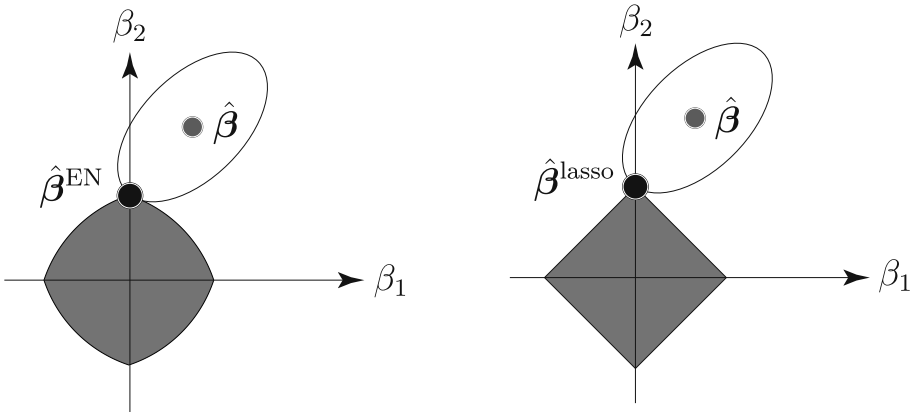



Fig. 2.5 Graphical representation of the elastic net (left) and lasso (right) estimates: The dark grey shape on the left-hand side means $\alpha \sum_j^p |\beta_j| + (1-\alpha) \sum_j^p \|\beta_j\|^2 \leq t$ ($\alpha = 0.5$). The elastic net estimate is the point where the loss ellipsoid hits the penalty shape.

Note that the elastic net, which has two penalty terms in the objective function of the optimization problem, incurs the "double shrinkage" without scaling by $(1+\lambda_2)$. It is observed in an empirical evidence of Zou and Hastie (2005) that the elastic net without scaling does not perform well compared with the ridge and the lasso.

### 2.3.3 Adaptive lasso

In the field of regression modeling, several studies (Fan and Li, 2001; Fan and Peng, 2004; Zou, 2006) have claimed that a good regression procedure should have the oracle property, where the oracle property is defined by the following:

A. Consistency in variable selection: $\Pr(\mathcal{A} = \mathcal{A}^*) \to 1$ $(n \to \infty)$.
B. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^*_{\mathcal{A}^*}) \xrightarrow{d} N_q(\mathbf{0}_q, \Sigma^*)$,

where the active set $\mathcal{A}$ is the set of predictors that are included in the estimated model based on $n$ observations, $\mathcal{A}^*$ is the true active set, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ is the estimated regression coefficient vector according to $\mathcal{A}$, $\boldsymbol{\beta}^*_{\mathcal{A}^*}$ is part of the true regression coefficient vector with nonzero component, $\xrightarrow{d}$ means the convergence in distribution, and $\Sigma^*$ is the variance covariance matrix knowing $\mathcal{A}^*$.

Zou (2006) showed that the lasso must satisfy some nontrivial condition to have consistency in variable selection and proposed the adaptive lasso that has the oracle property . The adaptive lasso is given by

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right], \tag{2.42}$$

where $\lambda_n$ is a tuning parameter dependant on sample size $n$, $\hat{w}_j = 1/|\tilde{\beta}_j|^\gamma$ $(j = 1, \ldots, p)$, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ is the root-$n$ consistent estimator of true regression coefficients $\boldsymbol{\beta}^*$ (e.g., the OLS or the MLE), and $\gamma > 0$ is a tuning parameter.

### 2.3.4 Group lasso

The linear regression modeling is used to model a linear relationship between a response variable and predictors, and it is widely used for the purpose of identifying the true structure that generates the response variable. We usually interpret the resulting models as meaning that the predictors, included in the model, are the explanatory factors of the response variable. These explanatory factors sometimes consist of a group of predictors. In the ANOVA (analysis of variance) model, a group of dummy variables compounds the predictor (e.g., we analyze the sexual influence by using a dummy variable, such that we set the male variable as one and female variable as zero for some observation from the male). The additive model in the nonlinear regression model also consists of a group of basis functions.

In these cases, the variable selection amounts to the selection of the important factor. The lasso, however, does not perform as the factor selection because it evaluates only each predictor in the penalty term. In order to overcome this difficulty, Yuan and Lin (2006) extended the group lasso, which is a lasso for factor selection.

We consider the regression model with $J$ factors:

$$\begin{aligned} \boldsymbol{y} &= \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon} \\ &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.43}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$ is the response vector, $X = (X_1, ..., X_J)$ is the $n \times p$ design matrix, $X_j = (\boldsymbol{x}_{j1}, ..., \boldsymbol{x}_{jp_j})$ is the $n \times p_j$ matrix corresponding to the $j$th factor, $\boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{1p_j}$ are the predictors of $j$th factor, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_J^T)^T$ is the coefficient vector, $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jp_j})^T$ is the coefficient vector of $j$th factor, and $\boldsymbol{\varepsilon} = (\varepsilon_1, .., \varepsilon_n)^T$ is the vector of independent and identically distributed error with mean 0 and variance $\sigma^2$. Similar to the lasso, we assume that the response and the predictor are centered without loss of generality. (Note that we do not assume that the predictors are standardized).

For a $q$-dimensional vector $\boldsymbol{x}$ ($q \geq 1$), we denote

$$\|\boldsymbol{x}\|_K = \sqrt{\boldsymbol{x}^T K \boldsymbol{x}}, \tag{2.44}$$

where $K$ is a $q \times q$ positive definite matrix. Let us have positive definite matrices $K_1, \ldots, K_J$, then the group lasso is given by

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \frac{1}{2} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\|_{K_j} \right], \tag{2.45}$$

where $\lambda (> 0)$ is a tuning parameter that controls the strength of regularization.

### 2.3.5 Bridge regression

As mentioned in Section 2.3.1, Frank and Friedman (1993) proposed the bridge regression as a generalization of the ridge estimates, and it is given by

$$\hat{\boldsymbol{\beta}}^{\text{bridge}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right], \tag{2.46}$$

where $\lambda, q(>0)$ are the tuning parameters (the bridge regression includes the ridge and lasso with $q = 2$ or $q = 1$).

Further, the bridge regression also has the conditional optimization form:

$$\hat{\boldsymbol{\beta}}^{\text{bridge}} := \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \right], \quad \text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t, \tag{2.47}$$

where $t(>0)$ is a tuning parameter. Thus, it is shown that the bridge estimates can be interpreted as the point where the squared loss ellipsoid hits the shape $\sum_j^p |\beta_j|^q = t$, where $t$ and $q$ are some fixed values (Fig. 2.6). When $0 < q \leq 1$, the

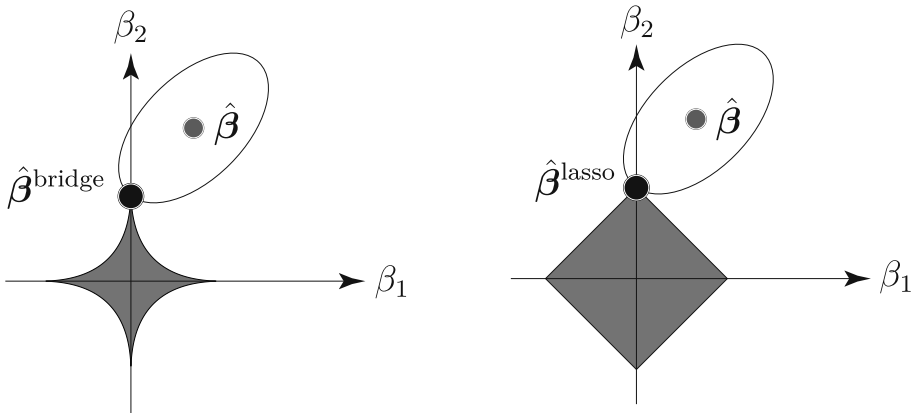



Fig. 2.6 Graphical representation of the bridge regression (left) and lasso (right) estimates: The dark grey shape on the left-hand side means $\sum_j^p |\beta_j|^q \leq t$ ($q = 0.5$). The bridge regression estimates is the point where the loss ellipsoid hits the penalty shape.

bridge regression enables us to obtain the sparse solution, and it also does stable estimation when $q > 1$.

Fig. 2.7 compares the penalty functions of the bridge regression, the adaptive lasso, and the lasso. The left-hand side panel is the case of $\hat{w} = 1/|\tilde{\beta}| = 1/2$ (the OLS is large) and the right-hand side one is the case of $\hat{w} = 1/|\tilde{\beta}| = 1/0.5$ (the OLS is small), respectively. It is shown that the adaptive lasso penalty becomes flat when the OLS is large and it sharpens when the OLS is small. This is a part of the reason of the oracle property of the adaptive lasso. On the other hand, the bridge penalty is sharp around the origin and becomes flat when $\beta$ is large in Fig. 2.7. From this, it is also known that the $q < 1$ bridge has the oracle property.

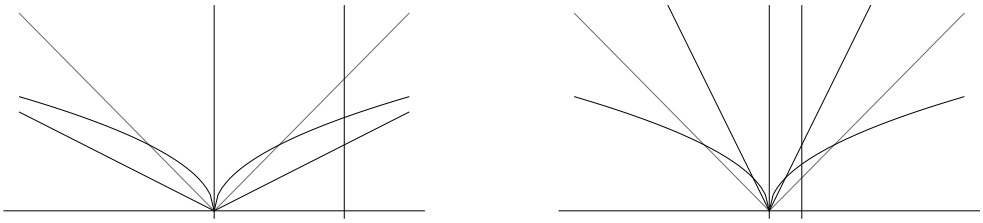


Fig. 2.7 Penalty functions of the lasso (grey dashed line), bridge regression (black real line, $q = 0.5$), and adaptive lasso (black dotted line, left: $\hat{w} = 0.5$, right: $\hat{w} = 2$).

Further, Huang *et al.* (2008) showed the oracle property of the bridge regression in high-dimensional models.

### 2.3.6 SCAD and MCP

The oracle property has another definition. We consider a convex loss function $l_n(\boldsymbol{\beta})$ such as $l_n(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$ or $\log N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n)$ and a penalty function $P_\lambda(\boldsymbol{\beta})$.

As in Fan *et al.* (2014), the oracle estimator is defined as

$$\hat{\boldsymbol{\beta}}^{\text{oracle}} := \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}^*} \\ \mathbf{0} \end{pmatrix} = \underset{\boldsymbol{\beta}; \boldsymbol{\beta}_{\mathcal{B}^*} = \mathbf{0}}{\text{argmin}}\, l_n(\boldsymbol{\beta}), \tag{2.48}$$

where $\mathcal{A}^*$ is a true active set and $\mathcal{B}^*$ is a complementary set of $\mathcal{A}^*$. Here, we assume that

$$\nabla_j l_n(\hat{\boldsymbol{\beta}}^{\text{oracle}}) = 0, \quad {}^\forall j \in \mathcal{A}^*, \tag{2.49}$$

where $\nabla_j$ denotes the sub-gradient with respect to the $j$-th element of $\boldsymbol{\beta}$.

When an estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\, [l_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})] \tag{2.50}$$

has the same asymptotic distribution as the oracle estimator, it is said to have the oracle property. Moreover, $\hat{\boldsymbol{\beta}}$ is said to have the strong oracle property if $\hat{\boldsymbol{\beta}}$ converges in probability on the oracle estimator.

Fan *et al.* (2014) showed that $\hat{\boldsymbol{\beta}}$ has the strong oracle property if the folded concave penalty function $P_\lambda(\boldsymbol{\beta})$ defined on $\beta_j \in (-\infty, \infty)$ satisfies the following four conditions:

(i). $P_\lambda(\beta_j)$ is increasing and concave in $\beta_j \in [0, \infty)$ with $P_\lambda(0) = 0$.

(ii). $P_\lambda(\beta_j)$ is differentiable in $\beta_j \in (0, \infty)$ with $P'_\lambda(0+) \geq a_1\lambda$, where $a_1$ is some fixed positive constant.

(iii). $P'_\lambda(\beta_j) \geq a_1\lambda$ for $\beta_j \in (0, a_2\lambda)$, where $a_2$ is some fixed positive constant.

(iv). $P'_\lambda(\beta_j) = 0$ for $\beta_j \in [a\lambda, \infty)$ with the pre-specified positive constant $a > a_2$.

Although the bridge penalty $P_\lambda = |\beta_j|^q$ $(0 < q < 1)$ satisfies the conditions (i), (ii) and (iii), it does not satisfy condition (iv). Condition (iv) means that the penalty function becomes completely flat for $\beta_j > c$ ($c$ is some constant). If $P_\lambda(\beta_j)$ that is flat at around $\tilde{\beta}_j$ ($\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$) is the minimizer of $l_n(\boldsymbol{\beta})$), the bias $|\tilde{\beta}_j - \hat{\beta}_j|$ decreases. However $|\beta_j|^q$ is not bounded, and thus, the bridge regression has the oracle property but does not have the strong oracle property.

There are two famous procedures that have the strong oracle property: the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010).

The penalty function of the smoothly clipped absolute deviation penalty (SCAD) is continuous differentiable, and its derivatives are defined by

$$\frac{\partial}{\partial \beta_j} P_\lambda(\beta_j) = \lambda \left\{ I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda) \right\}, \quad a > 2. \tag{2.51}$$

Thus, the SCAD estimates are given by

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \lambda|\beta_j| \cdot I(|\beta_j| < a\lambda) \right.$$
$$- \frac{(|\beta_j| - \lambda)^2}{2(\alpha - 1)} \cdot I(\lambda \leq |\beta_j| < a\lambda) \tag{2.52}$$
$$\left. + \frac{(\alpha + 1)\lambda^2}{2} \cdot I(|\beta_j| \geq a\lambda) \right].$$

The value of tuning parameter $a$ is often taken to be 3.7 in terms of the Bayes risks.

From (2.52), it can be seen that the SCAD is a bridge of $L_1$ and the OLS. When $|\beta_j| < \lambda$, the SCAD penalty is the $L_1$ norm of $\beta_j$. Further, for $|\beta_j| \geq a\lambda$, the SCAD does not penalize $\beta_j$. The tuning parameter $a$ controls a length of the transition interval from $L_1$ to the OLS.

When the minimax concave penalty (MCP) resembles the SCAD, the penalty function is also continuous differentiable. The MCP estimates are given by

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \lambda \int_0^{|\beta_j|} \left(1 - \frac{t}{a\lambda}\right) dt \right], \quad a > 1. \tag{2.53}$$

The integral in (2.53) is

$$\int_0^{|\beta_j|} \left(1 - \frac{t}{a\lambda}\right) dt = \frac{|\beta_j|(2a\lambda - |\beta_j|)}{2a} \cdot I(|\beta_j| < a\lambda) + \frac{a\lambda^2}{2} \cdot I(|\beta_j| \geq a\lambda). \tag{2.54}$$

From (2.54), it can be seen that the MCP is also a bridge of $L_1$ and the OLS, When $|\beta_j| \to 0$, the MCP penalty is the $L_1$ norm of $\beta_j$. Further for $|\beta_j| \geq a\lambda$, the MCP does not penalize $\beta_j$.

In the conditions of the strong oracle property, $a_1 = a_2 = 1$ for the SCAD, and $a_1 = 1 - 1/a$, $a_2 = 1$ for the MCP.

## 2.4 Algorithms for $L_1$ regularizations

In this section, we describe the typical algorithms to derive the lasso solution, LARS (Efron *et al.*, 2004) and the coordinate descent algorithm (Friedman *et al.*, 2010). Furthermore, we describe the algorithms for the non-convex regularizations such as the bridge, the SCAD and the MCP.

### 2.4.1 LARS

The least angle regression (LAR; Efron *et al.* 2004) builds a model continuously and enables us to obtain sparse models, that is, some coefficients shrink to exactly zero. LAR is very similar to the lasso, and we can obtain the lasso solution by slightly correcting LAR (this is called the "LARS").

Initially, the active set $\mathcal{A} = \emptyset$, the inactive set $\mathcal{B} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_p\}$, and coefficients according to $\mathcal{B}$ are all set to zero. First, LAR identifies the predictor most correlated with $\boldsymbol{y}$ in $\mathcal{B}$, which we denote as $\boldsymbol{x}_j$, and shifts it to $\mathcal{A}$. Although the best subset selection based on the least square procedure fits $\boldsymbol{x}_j$ completely, LAR gradually moves the coefficient of $\boldsymbol{x}_j$ $(= \beta_j)$ continuously from zero towards its least square value, causing its correlation with the current residual $\boldsymbol{r} = \boldsymbol{y} - \beta_j \boldsymbol{x}_j$ to decrease in terms of absolute value (this correlation equals zero when $\beta_j$ reaches its least square value). As soon as another predictor $\boldsymbol{x}_k$ has correlation with $\boldsymbol{r}$ as much, the process is paused and $\boldsymbol{x}_k$ is shifted to $\mathcal{A}$.

Next, LAR moves the coefficients of $\mathcal{A}$ together in a way that keeps their correlation with $\boldsymbol{r}$ $(= \boldsymbol{y} - \beta_j \boldsymbol{x}_j - \beta_k \boldsymbol{x}_k)$ tied and decreasing (they have their respective joint least square values in this direction). Then, when some other predictor $\boldsymbol{x}_\ell$ in $\mathcal{B}$ has correlation with the current residual as much, $\boldsymbol{x}_\ell$ is shifted to $\mathcal{A}$.

This process is continued until all the variables in the model are used, and it ends at the least squares estimates of $\min(n-1, p)$ predictors. If $p > n-1$, the residual becomes zero when the size of $\mathcal{A}$ is $n-1$ and the coefficients of $\mathcal{A}$ reach their joint least square values (i.e., the coefficients of $\mathcal{B}$ remains zero at the end of process).

The variables corresponding to $\mathcal{A}$ are tied in their absolute correlation with the current residuals, and we can express this as

$$\boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta}) = c_j\gamma, \tag{2.55}$$

where $c_j = \text{sign}[\boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta})]$, $j \in \mathcal{A}$, and $\gamma$ is some positive constant value. On the other hand, the lasso estimate $\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)$ for a given value of $\lambda$ is the minimizer of

$$R(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{2.56}$$

Suppose that $\mathcal{D} = \{j \ ; \ \hat{\beta}_j^{\text{lasso}}(\lambda) \neq 0\}$, $R(\boldsymbol{\beta})$ is differentiable for the variables corresponding to $\mathcal{D}$, and we have the following relationship

$$\boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \frac{\lambda}{2}\cdot\text{sign}(\hat{\beta}_j^{\text{lasso}}), \quad {}^\forall j \in \mathcal{D}. \tag{2.57}$$

From (2.55) and (2.57), the LAR estimates and the lasso estimates are identical only if $\text{sign}[\boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta})] = \text{sign}(\hat{\beta}_j^{\text{lasso}})$ and $\mathcal{A} = \mathcal{D}$. Hence the LAR and the lasso have similar estimates. However, when some coefficient of LAR passes through zero, $\mathcal{A}$ is not equivalent to $\mathcal{D}$. Therefore, we can calculate the lasso estimate by LARS with simple modification (Algorithm 1).

**Algorithm 1** Least angle regression with lasso modification

1. Start with the residual $\boldsymbol{r} = \boldsymbol{y}$, $\boldsymbol{\beta} = \mathbf{0}_p$, the active set $\mathcal{A} = \emptyset$, and the inactive set $\mathcal{B} = \{1, ..., p\}$.
2. Find the predictor most correlated with $\boldsymbol{r}$ from $\mathcal{B}$, and add it to $\mathcal{A}$.
3. Move $\boldsymbol{\beta}_\mathcal{A}$ toward $\hat{\boldsymbol{\beta}}^\mathcal{A} = (X_\mathcal{A}^T X_\mathcal{A})^{-1} X_\mathcal{A}^T \boldsymbol{y}$, gradually.
4. Stop above movement if
   a. some component of $\boldsymbol{\beta}_\mathcal{A}$ hits zero, and drop its variable from $\mathcal{A}$.
   b. some member of $\mathcal{B}$ has much correlation with the current residual.
5. Repeat steps 2, 3, 4 until all $p$ predictors have been entered.

Fig 2.8 shows all possible LAR solutions $\boldsymbol{\beta}$ for the diabetes data of Efron *et al.* (2004), as $t = \sum_{j=1} |\beta_j|$ increases from zero ($\boldsymbol{\beta} = \mathbf{0}$) to 3460, where $\boldsymbol{\beta}$ equals the least square value, which we call the "solution path". It is desired that a single estimate is chosen from the solution path, that is the model selection process of LAR.

### 2.4.2 Coordinate descent algorithm

The coordinate descent algorithm of Friedman *et al.* (2010) is proposed for solving $L_1 + L_2$ type regularization. In the recent years, most researchers used this algorithm to derive the lasso solution because of its extremely high speed.

We consider the elastic net problem,

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \right]. \tag{2.58}$$

Here, we try to partially optimize (2.58) with respect to $\beta_j$. Suppose that we have

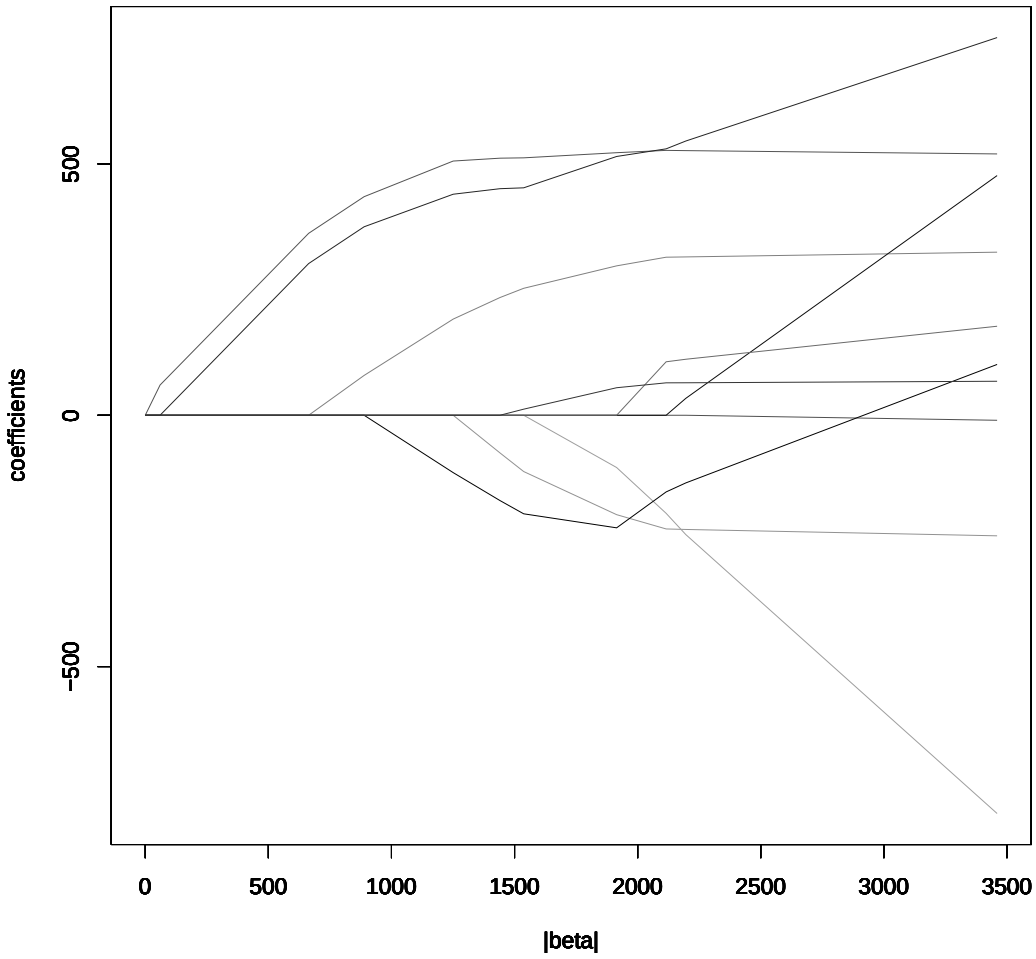



Fig. 2.8 Solution path of diabetes data

estimates $\tilde{\beta}_\ell$ $(\ell \neq j)$; then, the gradient at $\beta_j = \tilde{\beta}_j$ $(\tilde{\beta}_j \neq 0)$ is

$$\begin{aligned} &\frac{\partial}{\partial \beta_j} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 \right\} \Bigg|_{\beta_j = \tilde{\beta}_j} \\ &\qquad = -\frac{1}{n} \sum_{i=1}^{n} x_{ij} (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}) + \lambda_1 \operatorname{sign}(\tilde{\beta}_j) + \lambda_2 \tilde{\beta}_j. \end{aligned} \tag{2.59}$$

From the Karush-Kuhn-Tucker (KKT) conditions about this optimization, it is

shown that the estimates $\tilde{\beta}_j$ $(\neq 0)$ always maintain

$$\tilde{\beta}_j = \frac{1}{\lambda_2}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - \lambda_1 \operatorname{sign}(\tilde{\beta}_j)\right). \tag{2.60}$$

Thus, we have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\right| > \lambda_1, \quad \text{if and only} \quad \tilde{\beta}_j \neq 0, \\
&\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})\right| \leq \lambda_1, \quad \text{if and only} \quad \tilde{\beta}_j = 0.
\end{aligned} \tag{2.61}$$

Hence, the coordinate descent algorithm for an elastic net is given by an iterative algorithm that updates from $\tilde{\beta}_j^{(t)}$ to $\tilde{\beta}_j^{(t+1)}$ by

$$\tilde{\beta}_j^{(t+1)} \leftarrow \frac{1}{\lambda_2} S\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_{i(j)}^{(t)}),\ \lambda_1\right), \tag{2.62}$$

where

$$\begin{aligned}
\tilde{y}_{i(j)}^{(t)} &= \sum_{\ell \neq j} x_{i\ell}\tilde{\beta}_\ell^{(t)}, \\
S(x,\lambda) = \operatorname{sign}(x)(|x| - \lambda)_+ &= \begin{cases} x - \lambda & \text{if } x > 0 \text{ and } \lambda < |x|, \\ x + \lambda & \text{if } x < 0 \text{ and } \lambda < |x|, \\ 0 & \text{if } \lambda \geq |x|. \end{cases}
\end{aligned} \tag{2.63}$$

Since the coordinate descent algorithm does not need to calculate any inverse matrix, it is able to obtain the elastic net solution within a short time. Mazumder *et al.* (2011) said that "*Coordinate-wise optimization algorithms appear to be the fastest for computing the regularization paths for a variety of loss functions, and scale well*".

### 2.4.3 Local approximation procedures

The estimation algorithms for the lasso enable us to derive the solutions of other $L_1$ regularizations such as the elastic net and the adaptive lasso. We can transform

the elastic net optimization into the lasso problem (this is shown in Lemma 1 of Zou and Hastie (2005)):

$$\begin{aligned}&\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 \right] \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \boldsymbol{y}_*^T \boldsymbol{y}_* - 2\boldsymbol{y}_*^T X_* \boldsymbol{\beta} + \boldsymbol{\beta}^T X_*^T X_* \boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right],\end{aligned} \tag{2.64}$$

where

$$\boldsymbol{y}_* = \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0}_p \end{pmatrix}, \quad X_* = \begin{pmatrix} X \\ \sqrt{\lambda}_2 I_p \end{pmatrix}. \tag{2.65}$$

Thus we can obtain the elastic net solution by the following $L_1$ optimization:

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y}_* - X_*\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right]. \tag{2.66}$$

We can also obtain the solution of the adaptive lasso. The adaptive lasso problem can be transformed as follows:

$$\begin{aligned}&\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right] \\ &= \underset{\boldsymbol{\beta}=W\boldsymbol{\gamma}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - XW^{-1}\boldsymbol{\gamma}\|^2 + \lambda \sum_{j=1}^{p} |\gamma_j| \right],\end{aligned} \tag{2.67}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ ($\gamma_j = w_j\beta_j$) and $W = \operatorname{diag}(w_1, \ldots, w_p)$. Thus, the adaptive lasso is estimated by the following:

$$\hat{\boldsymbol{\beta}} = W^{-1} \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X_{**}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.68}$$

where $X_{**} = XW^{-1}$.

The solution of the $L_1$+$L_2$ or the weighted $L_1$ regularizations can be transformed

as the lasso problem. However, it is difficult to obtain the solutions of the non-convex regularizations such as the bridge, the SCAD, and the MCP. Fan and Li (2001) proposed the local quadratic approximation (LQA) procedure for this drawback.

We consider the regularized least square problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \right], \tag{2.69}$$

where $P_\lambda(|\beta_j|)(j = 1, \ldots, p)$ are some penalty functions such as $P_\lambda(|\beta_j|) = \lambda|\beta_j|^q$. In this optimization problem, it is difficult to solve the optimal value since the non-differentiability at the origin and the non-convexity of $P_\lambda(|\beta_j|)$ with respect to $\beta_j$. Hence, Fan and Li (2001) use a locally quadratic approximation to $P_\lambda(|\beta_j|)$:

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{P'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}), \quad \beta_j \approx \beta_j^{(0)}, \tag{2.70}$$

where

$$P'_\lambda(|\beta_j|) = \frac{\partial}{\partial \beta_j} P_\lambda(|\beta_j|). \tag{2.71}$$

The LQA enables us to solve the optimization problem of (2.69) with an iterative update,

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \frac{1}{2}\sum_{j=1}^{p} \frac{P'_\lambda(|\beta_j^{(k)}|)}{|\beta_j^{(k)}|}\beta_j^2 \right], k = 1, \ldots, . \tag{2.72}$$

The initial values of the LQA often use the OLS or the MLE.

However, the LQA does not derive a sparse solution for any regularization problems. Fan and Li (2001) suggested that if an absolute value of some component of the estimated regression coefficient vector in (2.72) is smaller than a pre-specified value $\varepsilon_0$, then we need to set it to zero and delete the corresponding predictor from iterations. Zou and Li (2008) listed two drawbacks of this procedure. First, if the LQA deletes a predictor at any step from a model, the predictor never returns to

the model. Second, the size of $\varepsilon_0$ affects the degrees of sparsity and the speed of convergence.

To overcome these drawbacks, Zou and Li (2008) proposed the local linear approximation (LLA) procedure. They approximated $P_\lambda(|\beta_j|)$ by the linear function,

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(0)}|) + P'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|), \quad \beta_j \approx \beta_j^{(0)}, \tag{2.73}$$

where

$$P'_\lambda(|\beta_j|) = \frac{\partial}{\partial|\beta_j|} P_\lambda(|\beta_j|). \tag{2.74}$$

Note that the LLA uses the derivation of $P_\lambda(|\beta_j|)$ by $|\beta_j|$, although the LQA differentiates it by $\beta_j$.

Thus, we have a local linear approximated non-convex regularization as follows:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P'_\lambda(|\beta_j^{(k)}|)|\beta_j| \right], k = 1, \ldots. \tag{2.75}$$

We can easily obtain the solution of this optimization problem because it is an adaptive lasso-type regularization in the case of $w_j = P'_\lambda(|\beta_j^{(k)}|)$.

Further, Zou and Li (2008) proposed the following procedure, which is called the "one-step local linear approximation":

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} P'_\lambda(|\beta_j^{(0)}|)|\beta_j| \right], \tag{2.76}$$

where $\beta_j^{(0)}$ is the OLS or the MLE. Although the penalty functions prefer that the resulting estimator is continuous, the bridge regression is not. On the other hand, the one-step LLA bridge is continuous. Furthermore, Zou and Li (2008) showed that the one-step LLA procedures have the oracle property, and Kanba and Naito (2011) proposed the model selection method for the one-step LLA procedures using results of Zou and Li (2008).

## 2.5 Degrees of freedom of the $L_1$ regularizations

In regression modeling, Mallows' $C_p$ type criteria (Mallows 1973) estimates the prediction error. The "degrees of freedom", which is often used to quantify the model complexity of modeling procedure, plays an important role in $C_p$. With the degrees of freedom, $C_p$ is an unbiased estimator of true prediction error, and Efron (2004) showed that in some setting, it offers substantially better accuracy than the cross-validation does. However, it is difficult to derive the closed form of the degrees of freedom of most continuous modeling, including LAR. The unbiased estimators of the degrees of freedom were used by several previous works.

For this problem, we show that the degrees of freedom of LAR are derived by the property of LAR. In this section, first, the definition of the degrees of freedom is described. Then, a new procedure that calculate the degrees of freedom of LAR is introduced. Note that this work is unpublished because we need to validate the efficiency of procedure.

### 2.5.1 Degrees of freedom

Let the expectation and the variance covariance matrix of response vector $\boldsymbol{y}$ be

$$\mathrm{E}\left[\boldsymbol{y}\right] = \boldsymbol{\mu}, \quad \mathrm{Var}\left[\boldsymbol{y}\right] = \sigma^2 I_n, \tag{2.77}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)^T$ is a true mean vector, $\sigma^2$ is a true variance, and $I_n$ is an $n$-dimensional identity matrix. We define a modeling procedure $\mathcal{M}$ as

$$\mathcal{M} : \boldsymbol{y} \to \hat{\boldsymbol{\mu}}, \tag{2.78}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_n)^T$, and we often use the notation $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\boldsymbol{y})$ to emphasize the dependence of $\hat{\boldsymbol{\mu}}$ on $\boldsymbol{y}$. Then, degrees of freedom of $\mathcal{M}$ are defined as (Ye 1998,

Efron 1986, 2004)

$$\mathrm{DF} = \sum_{i=1}^{n} \frac{\mathrm{cov}(\hat{\mu}_i, y_i)}{\sigma^2}, \tag{2.79}$$

where $\mathrm{cov}(\hat{\mu}_i, y_i)$ refers to the sampling covariance between $\hat{\mu}_i$ and $y_i$.

For example, in the simple case that $\mathcal{M}$ is the identity map, i.e. $\hat{\boldsymbol{\mu}}(\boldsymbol{y}) = \boldsymbol{y}$, the degrees of freedom is $n$. When $\hat{\boldsymbol{\mu}}$ is given in the linear form of $\hat{\boldsymbol{\mu}} = H\boldsymbol{y}$, where $H$ is a matrix that does not depend on $\boldsymbol{y}$, degrees of freedom is $\mathrm{tr}H$. The matrix $H$ is called a *hat matrix* or *smoother matrix*, which is widely used to select the optimal values of several tuning parameters, such as the ridge parameter and the smoothing parameters.

Degrees of freedom plays a key role in Mallows' $C_p$ criterion, which is an unbiased estimator of the true prediction error. Define the expected error as

$$\mathrm{Err} := \mathrm{E}\left[\mathrm{E}_*\left\{(\hat{\boldsymbol{\mu}} - \boldsymbol{y}^*)^T(\hat{\boldsymbol{\mu}} - \boldsymbol{y}^*)\right\}\right], \tag{2.80}$$

where the expectation "$\mathrm{E}_*$" is taken over $\boldsymbol{y}^* \sim (\boldsymbol{\mu}, \sigma^2 I)$ independent of $\boldsymbol{y}$. Err can be expressed as

$$\mathrm{Err} = \mathrm{E}\left[\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2\sigma^2 \mathrm{df}\right]. \tag{2.81}$$

This shows that $C_p$ criterion, defined by

$$C_p = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 + 2\sigma^2 \mathrm{df}, \tag{2.82}$$

is an unbiased estimator of Err with degrees of freedom.

### 2.5.2 DFLAR algorithm

As mentioned in Section 2.4.1, LAR moves the coefficients of active set $\mathcal{A}$ towards its least square solution, and when some predictor in the inactive set $\mathcal{B}$ has as much correlation with the current residual, it pauses the movement and shifts this predictor to $\mathcal{A}$. That is, LAR changes the direction of coefficients, movement at

the point where $\mathcal{A}$ has a new member (we call this point as the "turn point") and moves straightly between turn points. We show that this property enables us to obtain an estimate of the degrees of freedom of the LAR.

At first, we introduce the following new theorem.

Theorem 2.5.1 $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are some estimates of the regression coefficient vector. When an estimate $\boldsymbol{\beta}^*$ is defined by

$$\boldsymbol{\beta}^* := m\boldsymbol{\beta}^{(1)} + (1-m)\boldsymbol{\beta}^{(2)}, \tag{2.83}$$

where $m$ is a positive constant in $[0,1]$, the degrees of freedom of $\boldsymbol{\beta}$ is

$$m \cdot \mathrm{df}(\boldsymbol{\beta}^{(1)}) + (1-m) \cdot \mathrm{df}(\boldsymbol{\beta}^{(2)}), \tag{2.84}$$

where $\mathrm{df}(\boldsymbol{\beta})$ denotes the degrees of freedom of $\boldsymbol{\beta}$.

For example, if $H_1$ and $H_2$ are hat matrices and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are coefficient vectors according to $H_1$ and $H_2$, then degrees of freedom of these average coefficients $\boldsymbol{\beta}_* = (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)/2$ is $\mathrm{df}(\boldsymbol{\beta}_1)/2 + \mathrm{df}(\boldsymbol{\beta}_2)/2$, because

$$\hat{\boldsymbol{\mu}}_* = X\boldsymbol{\beta}_* = \frac{1}{2}X(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) = \frac{1}{2}(H_1 + H_2)\boldsymbol{y}. \tag{2.85}$$

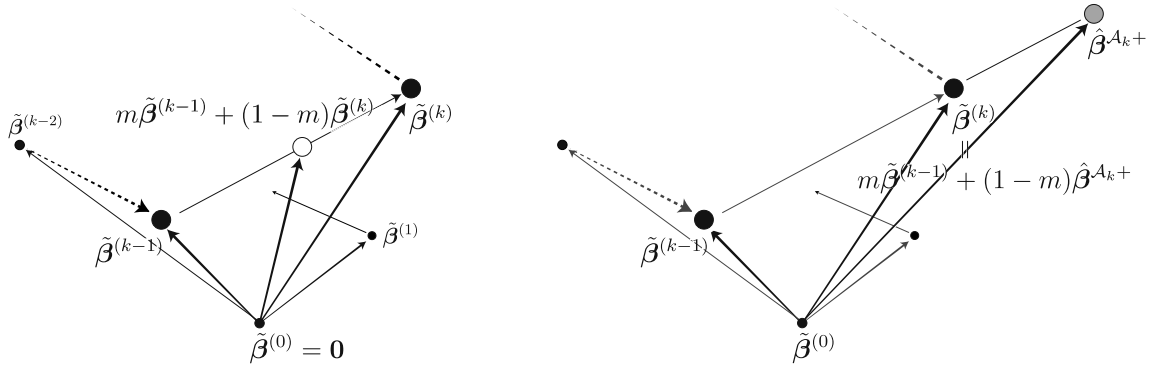



Fig. 2.9 The property of LAR

This theorem plays a key role in deriving the degrees of freedom of LAR. When $\boldsymbol{\beta}$ is any possible LAR estimate, it is the inner point of two turn points as shown in the left-hand side panel of Fig. 2.9. We can represent this as follows:

$$^{\forall}\boldsymbol{\beta} \in \{\boldsymbol{\beta}_{\mathrm{LAR}}\}, \ ^{\exists}m, k \quad \text{s.t.} \quad \boldsymbol{\beta} = m\tilde{\boldsymbol{\beta}}^{(k-1)} + (1-m)\tilde{\boldsymbol{\beta}}^{(k)}, \tag{2.86}$$

where $\boldsymbol{\beta}_{\mathrm{LAR}}$ represents any LAR estimate, $m \in [0,1]$, $k = 0, 1, ..., \min(n-1, p)-1$, and $\tilde{\boldsymbol{\beta}}^{(k)}$ is the $k$th turn point of LAR ($\tilde{\boldsymbol{\beta}}^{(0)}$ is $\mathbf{0}$). Thus, it follows from Theorem 2.5.1 that if we can obtain the degrees of freedom of all turn points of LAR, we can also obtain degrees of freedom of all possible estimates of LAR.

As given in the right-hand side panel of Fig. 2.9, the least square value of $\mathcal{A}$ is an extension of the next turn point from the previous turn point. Then, every turn point is also the inner point of the previous turn point and least square estimate, which we can represent as follows:

$$\begin{aligned} ^{\forall}\boldsymbol{\beta} \in \left\{\tilde{\boldsymbol{\beta}}^{(j)} \mid j = 1, 2, ..., \min(p, n-1)\right\}, \\ ^{\exists}m, \quad \text{s.t.} \quad \boldsymbol{\beta} = m\tilde{\boldsymbol{\beta}}^{(k-1)} + (1-m)\hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}, \end{aligned} \tag{2.87}$$

where $m$ is some constant in $[0,1]$, $\tilde{\boldsymbol{\beta}}^{(k-1)}$ is the previous turn point, $\hat{\boldsymbol{\beta}}^{\mathcal{A}}$ is the least square estimate on predictors in $\mathcal{A}$, $\mathcal{A}_k$ is the current active set, $\hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}$ is a $p$ dimensional vector whose elements corresponding to the predictors in $\mathcal{A}_k$ are their joint least square values, and other elements are zero. From the above two results, we can obtain the following new theorem.

Theorem 2.5.2 For a single LAR estimate $\boldsymbol{\beta}$, there is $k$ such that

$$\boldsymbol{\beta} = m\tilde{\boldsymbol{\beta}}^{(k-1)} + (1-m)\hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}, \tag{2.88}$$

where $m$ is some constant in $[0,1]$.

If we obtain all turn points of LAR, we can calculate $m$ for a given estimate $\boldsymbol{\beta}$

by

$$m = \frac{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}\|}{\|\tilde{\boldsymbol{\beta}}^{(k-1)} - \hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}\|}. \tag{2.89}$$

However, the calculation of $m$ can be computationally expensive in high-dimensional data because computing the least square solution needs the calculation of the inverse matrix. We calculate $m$ by using the relationship

$$\begin{aligned} \boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta}) &= \boldsymbol{x}_j^T \left\{ \boldsymbol{y} - X(m\tilde{\boldsymbol{\beta}}^{(k-1)} + (1-m)\hat{\boldsymbol{\beta}}^{\mathcal{A}_k+} \right\} \\ &= m \left\{ \boldsymbol{x}_j^T(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}^{(k-1)}) \right\} + (1-m) \left\{ \boldsymbol{x}_j^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\mathcal{A}_k+}) \right\} \\ &= m \left\{ \boldsymbol{x}_j^T(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}^{(k-1)}) \right\} \end{aligned} \tag{2.90}$$

where $\boldsymbol{x}_j$ is some predictor in $\mathcal{A}_k$,

$$m = \frac{\boldsymbol{x}_j^T(\boldsymbol{y} - X\boldsymbol{\beta})}{\boldsymbol{x}_j^T(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}^{(k-1)})}. \tag{2.91}$$

Thus, we obtain the following algorithm that computes the solution path of LAR and its degrees of freedom.

LAR is closely related to $L_1$ regularizations. From this relationship, we can obtain the degrees of freedom of various $L_1$ type regularizations, such as lasso, adaptive lasso, group lasso, and elastic net, by using the proposed algorithm. This is an unpublished result. We need more validation for this procedure to publish.

## 2.6 Strength of the sparsity of the $L_1$ regularizations

Several $L_1$ regularizations have been proposed, and they have different characteristics. For example, the elastic net works well in high-dimensional modeling, the bridge has the oracle property, and the SCAD and the MCP have the strong oracle property.

Here, we consider the strength of the sparsity. Some experience has shown that the strength of sparsity of the elastic net is weaker than that of the lasso, and the bridge has strong sparsity. However, as we do not have the definition of the

**Algorithm 2** Least angle regression with degrees of freedom

1. Start with the residual $\boldsymbol{r} = \boldsymbol{y}$, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, DF $= 0$ and set the active set $\mathcal{A} = \emptyset$ and the inactive set $\mathcal{B} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p\}$.
2. Find the predictor most correlated with $\boldsymbol{r}$ from $\mathcal{B}$, add it to $\mathcal{A}$, and let $k = |\mathcal{A}|$.
3. Move the coefficients according to member of $\mathcal{A}$ toward its least squares coefficient $\hat{\boldsymbol{\beta}}^{\mathcal{A}}$.
4. Calculate DF $= m \cdot \mathrm{df}(\tilde{\boldsymbol{\beta}}^{(k-1)}) + (1-m) \cdot k$, where $m = (\boldsymbol{x}_j^T \boldsymbol{r})/(\boldsymbol{x}_j^T(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}^{(k-1)}))$, $\boldsymbol{x}_j \in \mathcal{A}$, and $\mathrm{df}(\tilde{\boldsymbol{\beta}}^{(k-1)})$ denotes the degrees of freedom of the previous turn point.
5. Stop this move if some member of $\mathcal{B}$ has large correlation with the current residual.
6. Repeat steps 2, 3, 4, 5 until all $p$ predictors have been entered.

strength of the sparsity, it is difficult to evaluate it in a quantitative way. Therefore we establish the definition of the strength of the sparsity.

Most regularization procedures shrink the OLS or the MLE towards zero. Further, if the least squares are closed to zero, they produce zero values for the regularized estimates. Henceforth, we define the strength of the sparsity (SS) of a penalty function $P_\lambda(\beta)$ as follows:

$$\mathrm{SS} := \underset{\beta^*}{\mathrm{argmax}} \left[ \underset{\beta}{\mathrm{argmin}} \left\{ \|\beta^* - \beta\|^2 + P_1(\beta) \right\} = 0 \right]. \tag{2.92}$$

The strength of sparsity of the $L_1$ regularizations (the lasso, the elastic net ($\alpha = 0.3, 0.5, 0.7$), the adaptive lasso ($w = 1/|\beta^*|^\gamma$, $\gamma = 0.5, 1.0, 2.0$), the bridge ($q = 0.3, 0.5, 0.7$), the SCAD ($a = 3.7$) and the MCP ($a = 0.5, 1.0, 2.0$)) are given in Table 2.1. It shows that the lasso, the SCAD and the MCP have the same values of the strength of the sparsity, and the adaptive lasso and the bridge have larger

Table 2.1 The strength of the sparsity of the $L_1$ regularizations

| Procedure | SS | |
|---|---|---|
| lasso | 0.500 | |
| elastic net | 0.150 | ($\alpha$=0.3) |
| | 0.250 | ($\alpha$=0.5) |
| | 0.350 | ($\alpha$=0.7) |
| adaptive lasso | 0.630 | ($\gamma$=0.5) |
| | 0.707 | ($\gamma$=1.0) |
| | 0.794 | ($\gamma$=2.0) |
| bridge | 0.984 | ($q$=0.3) |
| | 0.945 | ($q$=0.5) |
| | 0.858 | ($q$=0.7) |
| SCAD | 0.500 | ($a$=3.7) |
| MCP | 0.500 | ($a$=0.5) |
| | 0.500 | ($a$=1.0) |
| | 0.500 | ($a$=2.0) |

values of it. The strength of the sparsity of the elastic net is smaller than that of the lasso.

# Chapter 3

# Bayes model for $L_1$ regularizations

Several regularization procedures can be interpreted as the MAP (maximum a posteriori) estimation under some Bayes model. For example, the ridge in linear regression model

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right] \tag{3.1}$$

is equivalent to the MAP estimator of the model

$$\begin{aligned} &\text{Likelihood}: \ N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n), \\ &\text{Prior on } \boldsymbol{\beta}: \ N_p\left(\boldsymbol{\beta}|\mathbf{0}, \frac{\sigma^2}{\lambda} I_p\right), \end{aligned} \tag{3.2}$$

where $N_q(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ is a probability density function of a normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\Sigma$.

In this chapter, we discuss the relationship between the $L_1$ regularizations and the Bayes models. Further, we introduce some Bayesian analysis procedures which have been extended from $L_1$ regularizations (for details, we refer to Park and Casella (2008) and Kyung *et al.* (2010)). Moreover, we show that various Bayesian procedures have the unimodality, which is a keyrole in Bayesian analysis using some computational techniques.

## 3.1 Relationship between the lasso and Bayes model

We consider a linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an $n$-dimensional response vector, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ design matrix, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the $p$-dimensional observations for predictor variables, the elements of $\boldsymbol{x}_i$ is given as $x_{i1}, \ldots, x_{ip}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector which elements have independent and identically distributed according to a normal distribution with mean zero and unknown variance $\sigma^2$. Without loss of generality, we assume that the predictors are standardized:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad j = 1, \ldots, p. \tag{3.4}$$

The lasso estimate

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right], \quad \lambda > 0, \tag{3.5}$$

can be interpreted as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ -\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 - \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ \exp(-\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2) \cdot \prod_{j=1}^{p} \exp(-\lambda|\beta_j|) \right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ \exp\left( -\frac{1}{2\sigma^2} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \right) \cdot \prod_{j=1}^{p} \exp\left( -\frac{\lambda}{2\sigma^2} |\beta_j| \right) \right] \\
&= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ N_n(\boldsymbol{y} | X\boldsymbol{\beta}, \sigma^2 I_n) \cdot \prod_{j=1}^{p} \frac{\lambda}{4\sigma^2} \exp\left( -\frac{\lambda}{2\sigma^2} |\beta_j| \right) \right].
\end{aligned} \tag{3.6}$$

Thus, the lasso estimate can be interpreted as the MAP estimates under independent Laplace priors for $\boldsymbol{\beta}$ (e.g. Tibshirani, 1996; Park and Casella, 2008).

## 3.2 Laplace distribution and Scale mixture normal distribution

However, it is difficult to obtain the posterior distribution or MCMC sample from (3.6) because of the Laplace prior. For this drawback, the result of Andrews and Mallows (1974) is applicable.

Suppose that $Z$ has a standard normal distribution and $V$ is a positive continuous random variable, and $V$ is independent of $Z$. Let $X = Z/V$, then $X$ has a probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \int_0^\infty v \cdot \exp\left(-\frac{1}{2}v^2x^2\right) f_V(v)dv, \tag{3.7}$$

where $f_V(v)$ is a probability density function of $V$.

Consider the transformation $v = \sqrt{2t}$, and we define $h(t)$ by

$$h(t) = \frac{\sqrt{t}}{\sqrt{\pi}} f_V(\sqrt{2t}) \left|\frac{dv}{dt}\right|. \tag{3.8}$$

Then, if $h(y) = f_X(\sqrt{y})$, $h(y)$ is the Laplace transformation of $h(t)$, because

$$\begin{aligned} h(y) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty v \cdot \exp\left(-\frac{1}{2}v^2y\right) f_V(v)dv \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \sqrt{2t} \cdot \exp\left(-ty\right) f_V(\sqrt{2t}) \frac{1}{\sqrt{2t}} dt \\ &= \int_0^\infty \exp(-yt) f(t) dt. \end{aligned} \tag{3.9}$$

Hence, we have the relationship between $h(y)$, $f(t)$ and $f_V(v)$ as follows:

$$\begin{aligned} h(y) &= \int_0^\infty \exp(-yt) f(t) dt, \\ f(t) &= \frac{1}{\sqrt{2\pi}} f_V(\sqrt{2t}), \\ f_V(v) &= \sqrt{2\pi} \cdot f\left(\frac{1}{2}v^2\right). \end{aligned} \tag{3.10}$$

Now, let $X$ have the Laplace distribution, i.e.

$$f_X(x) = \frac{1}{2}\exp(-|x|), \tag{3.11}$$

the inverse Laplace transformation of $h(y) = (1/2)\exp(-\sqrt{y})$ is

$$\frac{1}{2}\exp(-\sqrt{y}) = \int_0^\infty \exp(-yt)\frac{1}{4\sqrt{\pi t^3}}\exp\left(-\frac{1}{4t}\right) dt, \tag{3.12}$$

because the inverse Laplace transformation of $\exp(-a\sqrt{s})$ is given by $\{a/(2\sqrt{\pi t^3})\}\exp(-a^2/(4t))$. Thus, we have

$$\begin{aligned} f_V(v) &= \sqrt{2\pi} \cdot f\left(\frac{1}{2}v^2\right) \\ &= \frac{1}{v^3}\exp\left(-\frac{1}{2v^2}\right). \end{aligned} \tag{3.13}$$

We transform $\tau^2 = (2v^2)^{-1}$,

$$f_{\tau^2}(\tau^2) = \exp(-\tau^2). \tag{3.14}$$

From above, Andrews and Mallows (1974) showed that the Laplace distribution can be represented as the following scale mixture normals:

$$\frac{\lambda}{2}\exp(-\lambda|x|) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{x^2}{2\tau^2}\right)\frac{\lambda^2}{2}\exp\left(-\frac{\lambda}{2}\tau^2\right), \tag{3.15}$$

where $\lambda > 0$.

## 3.3 Bayesian lasso

Park and Casella (2008) proposed the Gibbs sampling for the lasso with a hierarchical Laplace prior or scale mixture normal prior based on the result of Andrews and Mallows (1974). Note that Park and Casella (2008) considered the Bayes model based on the following conditional Laplace prior:

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right). \tag{3.16}$$

This conditional prior of $\boldsymbol{\beta}$ given $\sigma^2$ guarantees a unimodal posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$, this avoids the slow convergence of the Gibbs sampler. This procedure that is called the "Bayesian lasso", is the Gibbs sampling from hierarchical representation of the following full model:

$$\begin{aligned} p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) &= N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n), \\ p(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|\mathbf{0}_p, \sigma^2 D), \\ p(\sigma^2) &= \frac{1}{\sigma^2} \text{ or } \mathrm{IG}(\sigma^2|\nu_0, \eta_0), \\ p(\tau_1^2, \ldots, \tau_p^2|\lambda) &= \prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2|\frac{\lambda^2}{2}\right), \end{aligned} \tag{3.17}$$

where $\mathbf{0}_q$ is a $q$-dimensional vector whose elements are all 0, $D = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$, $\mathrm{IG}(x|\nu, \eta)$ is a probability density function of a inverse gamma distribution with variable $x$, the shape parameter $\nu$ and the rate parameter $\eta$.

The full model (3.17) leads to the following full conditional distributions of $\boldsymbol{\beta}$, $\sigma^2$, and $1/\tau_1^2, \ldots, 1/\tau_p^2$ (when $p(\sigma^2) = 1/\sigma^2$):

$$\begin{aligned} p_{\text{full}}(\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \sigma^2 A^{-1}), \\ p_{\text{full}}(\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2) &= \mathrm{IG}(\sigma^2|\nu_1, \eta_1) \\ p_{\text{full}}(1/\tau_1^2, \ldots, 1/\tau_p^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \sigma^2, \lambda) &= \prod_{j=1}^{p} \mathrm{IGauss}(1/\tau_j^2|\mu_j', \lambda'), \end{aligned} \tag{3.18}$$

where

$$\begin{aligned} A &= X^T X + D^{-1}, \\ \nu_1 &= \frac{n+p}{2}, \quad \eta_1 = \frac{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta}}{2}, \\ \mu'_j &= \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \quad \lambda' = \lambda^2, \end{aligned} \tag{3.19}$$

and $\mathrm{IGauss}(x|\mu, \lambda)$ is a probability density function of a inverse gaussian distribution with variable $x$ ($x > 0$), the mean $\mu$, and the shape parameter $\lambda$ (if $\sigma^2$ has a inverse gamma prior, $\nu_1 = (n+p+\nu_0)/2$ and $\eta_1 = \{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})+\boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta}+ \eta_0\}/2$). Further, Park and Casella (2008) suggested how to choose the Bayesian lasso tuning parameter $\lambda$ in Bayesian analysis; empirical Bayes through marginal maximum likelihood and hierarchical Bayes through gamma priors $\mathrm{Gamma}(\lambda^2|r, \delta)$ on $\lambda^2$, where

$$\mathrm{Gamma}(\lambda^2|r, \delta) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}\exp(-\delta\lambda^2), \quad r > 0,\ \delta > 0. \tag{3.20}$$

By generating Gibbs samples according to these full conditional distributions (3.18), we can obtain some information about the posterior of $(\boldsymbol{\beta}, \sigma^2)$, even if it is difficult to derive a closed form of the posterior.

## 3.4 Other Bayes model of $L_1$ regularizations

Similar to the Bayesian lasso, various extensions for Bayesian procedures of $L_1$ regularizations have been proposed (e.g., Kyung *et al.* , 2010). Here, we introduce some Bayes-type $L_1$ regularizations, and we discuss about the unimodality of the posterior distribution of these procedures. Also, Polson *et al.* (2014) introduced an extension of the bridge regression for the Bayesian modeling.

### 3.4.1 Bayesian elastic net

The elastic net problem for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[ \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right] \tag{3.21}$$

is equivalent to the MAP problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ \exp(-\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2) \cdot \prod_{j=1}^{p} \exp\left(-\lambda_1 |\beta_j|\right) \cdot \prod_{j=1}^{p} \exp\left(-\lambda_2 \beta_j^2\right) \right]. \tag{3.22}$$

Hence, it is shown that the elastic net is also the MAP estimator when $\boldsymbol{\beta}$ has a Laplace and normal prior in a normal linear regression model. From this relationship, Kyung *et al.* (2010) suggested the Gibbs sampling from the following hierarchical model:

$$\begin{aligned} p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) &= N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n), \\ p(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|\mathbf{0}_p, \sigma^2 D), \\ p(\sigma^2) &= \frac{1}{\sigma^2} \text{ or } \mathrm{IG}(\sigma^2|\nu_0, \eta_0), \\ p(\tau_1^2, \ldots, \tau_p^2|\lambda) &= \prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2|\frac{\lambda_1^2}{2}\right), \end{aligned} \tag{3.23}$$

where $D = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) + (1/\lambda_2) I_p$.

The full model (3.23) leads to the following full conditional distributions of $\boldsymbol{\beta}$, $\sigma^2$, and $1/\tau_1^2, \ldots, 1/\tau_p^2$ (when $p(\sigma^2) = 1/\sigma^2$):

$$\begin{aligned} p_{\text{full}}(\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \sigma^2 A^{-1}), \\ p_{\text{full}}(\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2) &= \mathrm{IG}(\sigma^2|\nu_1, \eta_1) \\ p_{\text{full}}(1/\tau_1^2, \ldots, 1/\tau_p^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \sigma^2, \lambda) &= \prod_{j=1}^{p} \mathrm{IGauss}(1/\tau_j^2|\mu_j', \lambda'), \end{aligned} \tag{3.24}$$

where

$$\begin{aligned} A &= X^T X + D^{-1}, \\ \nu_1 &= \frac{n+p}{2}, \quad \eta_1 = \frac{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta}}{2}, \\ \mu'_j &= \sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \quad \lambda' = \lambda_1^2, \end{aligned} \tag{3.25}$$

and if $\sigma^2$ has a inverse gamma prior, $\nu_1 = (n+p+\nu_0)/2$ and $\eta_1 = \{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}) + \boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta} + \eta_0\}/2)$. Kyung *et al.* (2010) also suggested how to choose the Bayesian elastic net tuning parameter $\lambda$ in Bayesian analysis; empirical Bayes through marginal maximum likelihood and hierarchical Bayes through gamma priors $\text{Gamma}(\lambda_1^2|r_1, \delta_1)$ on $\lambda_1^2$ and $\text{Gamma}(\lambda_2|r_2, \delta_2)$ on $\lambda_2$.

### 3.4.2 Bayesian adaptive lasso

Further, the adaptive lasso has the MAP problem form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \left[ \exp(-\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2) \cdot \prod_{j=1}^{p} \exp\left(-\lambda_j|\beta_j|\right) \right]. \tag{3.26}$$

Hence, the Gibbs sampling of the adaptive lasso can be take from following hierarchical model:

$$\begin{aligned} p(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2) &= N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n), \\ p(\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|\mathbf{0}_p, \sigma^2 D), \\ p(\sigma^2) &= \frac{1}{\sigma^2} \text{ or } \text{IG}(\sigma^2|\nu_0, \eta_0), \\ p(\tau_1^2, \ldots, \tau_p^2|\lambda) &= \prod_{j=1}^{p} \text{Exp}\left(\tau_j^2|\frac{\lambda_j^2}{2}\right), \end{aligned} \tag{3.27}$$

where $D = \text{diag}(\tau_1^2, \ldots, \tau_p^2)$.

The full model (3.27) leads to following full conditional distributions of $\boldsymbol{\beta}$, $\sigma^2$,

and $1/\tau_1^2, \ldots, 1/\tau_p^2$ (when $p(\sigma^2) = 1/\sigma^2$):

$$\begin{aligned} p_{\text{full}}(\boldsymbol{\beta}|\boldsymbol{y}, X, \sigma^2, \tau_1^2, \ldots, \tau_p^2) &= N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \sigma^2 A^{-1}), \\ p_{\text{full}}(\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2) &= \text{IG}(\sigma^2|\nu_1, \eta_1) \\ p_{\text{full}}(1/\tau_1^2, \ldots, 1/\tau_p^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \sigma^2, \lambda) &= \prod_{j=1}^{p} \text{IGauss}(1/\tau_j^2|\mu_j', \lambda'), \end{aligned} \tag{3.28}$$

where

$$\begin{aligned} A &= X^T X + D^{-1}, \\ \nu_1 &= \frac{n+p}{2}, \quad \eta_1 = \frac{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta}}{2}, \\ \mu_j' &= \sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}, \quad \lambda' = \lambda_j^2, \end{aligned} \tag{3.29}$$

and if $\sigma^2$ has a inverse gamma prior, $\nu_1 = (n+p+\nu_0)/2$ and $\eta_1 = \{(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) + \boldsymbol{\beta}^T D^{-1}\boldsymbol{\beta} + \eta_0\}/2)$. It is considered that the we can set the gamma priors on $\lambda_j$ $\text{Gamma}(\lambda_j^2|r, \delta)$ on $\lambda_j^2$ $(j = 1, \ldots, p)$.

### 3.4.3 Bayesian group lasso

Moreover, Kyung *et al.* (2010) suggested the Bayesian extension of the group lasso,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left[ \|\boldsymbol{y} - \sum_{j=1}^{J} X\boldsymbol{\beta}_j\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\beta}_j\| \right], \tag{3.30}$$

where $J$ is the number of factor, and penalty terms based on $\|\boldsymbol{\beta}_j\|$ despite original group lasso penalty is based on $(\boldsymbol{\beta}_j^T K_j \boldsymbol{\beta}_j)^{1/2}$.

The full model and full conditional of the Bayesian group lasso is given by

$$\begin{aligned}
p(\boldsymbol{y}|X,\boldsymbol{\beta},\sigma^2) &= N_n(\boldsymbol{y}|X\boldsymbol{\beta},\sigma^2 I_n),\\
p(\boldsymbol{\beta}_j|\sigma^2,\tau_j^2) &= N_{p_j}(\boldsymbol{\beta}|\mathbf{0}_p,\sigma^2\tau_j^2),\\
p(\sigma^2) &= \frac{1}{\sigma^2} \text{ or } \mathrm{IG}(\sigma^2|\nu_0,\eta_0),\\
p(\tau_1^2,\ldots,\tau_p^2|\lambda) &= \prod_{j=1}^{J} \mathrm{Gamma}\left(\tau_j^2|\frac{\lambda_{p_j+1}^2}{2},\frac{\lambda^2}{2}\right),\\
p_{\mathrm{full}}(\boldsymbol{\beta}_j|\boldsymbol{y},X,\boldsymbol{\beta}_{(-j)},\sigma^2,\tau_1^2,\ldots,\tau_J^2) &= N_{p_j}(\boldsymbol{\beta}|A_j^{-1}X^T\tilde{\boldsymbol{y}}_{(j)},\sigma^2A_j^{-1}),\\
p_{\mathrm{full}}(\sigma^2|\boldsymbol{y},X,\boldsymbol{\beta},\tau_1^2,\ldots,\tau_p^2) &= \mathrm{IG}(\sigma^2|\nu_1,\eta_1)\\
p_{\mathrm{full}}(1/\tau_1^2,\ldots,1/\tau_p^2|\boldsymbol{y},X,\boldsymbol{\beta},\sigma^2,\lambda) &= \prod_{j=1}^{p} \mathrm{IGauss}(1/\tau_j^2|\mu_j',\lambda'),
\end{aligned} \tag{3.31}$$

where $p_j$ is the dimensionality of $\boldsymbol{\beta}_j$,

$$\begin{aligned}
\boldsymbol{\beta}_{(-j)} &= (\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_{j-1}^T,\boldsymbol{\beta}_{j+1}^T,\ldots,\boldsymbol{\beta}_J^T)^T,\\
A_j &= X_j^TX_j + \frac{1}{\tau_j^2}I_{p_j},\\
\tilde{\boldsymbol{y}}_{(j)} &= \boldsymbol{y} - \frac{1}{2}\sum_{k\neq j} X_k\boldsymbol{\beta}_k,\\
\nu_1 &= \frac{n+p}{2}, \quad \eta_1 = \frac{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta})}{2} + \sum_j^J \frac{\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j}{2\tau_j^2},\\
\mu_j' &= \sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}, \quad \lambda' = \lambda_j^2,
\end{aligned} \tag{3.32}$$

and if $\sigma^2$ has a inverse gamma prior, $\nu_1 = (n+p+\nu_0)/2$ and $\eta_1 = \{(\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}) + \sum_j^J(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j/\tau_j^2) + \eta_0\}/2)$. Kyung *et al.* (2010) suggested how to choose the Bayesian group lasso tuning parameter $\lambda$ in Bayesian analysis; empirical Bayes through marginal maximum likelihood and hierarchical Bayes through gamma priors $\mathrm{Gamma}(\lambda|r,\delta)$ on $\lambda$.

### 3.4.4 Unimodality of the posteriors

In Bayesian procedures, the unimodality of the posterior is important role to obtain the MCMC sample via the Gibbs sampler. Absence of the unimodality induces retardation of convergence of the Gibbs sampler and the point estimates becomes less meaningful.

Park and Casella (2008) proposed the Bayesian lasso that has unimodal posterior. However, Park and Casella (2008) only showed that the joint posterior of $\boldsymbol{\beta}$ and $\sigma^2$ is unimodal. Here, we show that the Bayesian lasso, the Bayesian elastic net and the Bayesian adaptive lasso have a unimodal posterior when the tuning parameters have gamma priors.

#### Unimodality of the Bayesian lasso

In the Bayesian lasso with gamma prior $\text{Gamma}(\lambda^2|r,\delta)$ on $\lambda^2$, the likelihood and priors are given by

$$\begin{aligned} &\text{Likelihood: } (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\beta}\|^2\right), \\ &\text{Priors: } \prod_{j=1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda}{\sigma^2}|\beta_j|\right)\cdot\pi(\sigma^2)\cdot\frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}\exp(-\delta\lambda^2). \end{aligned} \tag{3.33}$$

The log posterior is proportionate to

$$\begin{aligned} \log\pi(\sigma^2)-\frac{n+p}{2}\log(\sigma^2)+p\log\lambda-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{y}\|^2 \\ -\frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|+(r-1)\log\lambda^2-\delta\lambda^2. \end{aligned} \tag{3.34}$$

The unimodality of a function $F$ in some coordinates is equivalent to unimodality in transformed coordinates when the transformation is continuous with a continuous inverse at support of $F$. Using this property, we transform the coordinate as

$$\phi_j=\frac{1}{\sqrt{\sigma^2}}\beta_j\ (j=1,\ldots,p),\quad \rho=\frac{1}{\sqrt{\sigma^2}},\quad \lambda, \tag{3.35}$$

and (3.34) becomes

$$\begin{aligned}\log \pi\left(\frac{1}{\rho^2}\right) + (n+p)\log(\rho) - \frac{1}{2}\|\rho \boldsymbol{y} - X\boldsymbol{\phi}\|^2 \\ - \lambda \sum_{j=1}^{p} |\phi_j| + \left(r - 1 + \frac{p}{2}\right)\log \lambda^2 - \delta\lambda^2,\end{aligned} \tag{3.36}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$. If $\pi(\sigma^2)$ is $1/\sigma^2$ or inverse gamma density, the first term is concave. The second and sixth terms are concave in $(\boldsymbol{\phi}, \rho, \lambda)$, The fourth term is also concave in $(\boldsymbol{\phi}, \rho, \lambda)$ because it is a sum of concave function $-\lambda|\phi_j|$ , and the third term is concave quadratic in $(\boldsymbol{\phi}, \rho, \lambda)$. The fifth term is concave in $(\boldsymbol{\phi}, \rho, \lambda)$ when $r > 1 - (p/2)$ (it is always satisfied when $p \geq 2$ since $r$ takes positive value). Hence, it is showed that (3.34) is concave, and the posterior of the Bayesian lasso with prior Gamma$(\lambda^2|r, \delta)$ is unimodal when $r > 1 - (1/p)$.

### Unimodality of the Bayesian elastic net

In the Bayesian elastic net, Sepehri (2016) only showed the unimodality of the joint posterior of $(\boldsymbol{\beta}, \sigma^2)$. We show that the Bayesian elastic net has a unimodal posterior when the tuning parameters have gamma priors.

The Bayesian elastic net has the following likelihood and priors:

$$\begin{aligned}&\text{Likelihood: } (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|^2\right), \\ &\text{Priors: } \prod_{j=1}^{p}\frac{\lambda_1}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda_1}{\sigma^2}|\beta_j|\right)\cdot(2\pi)^{-p/2}(\sigma^2)^{-p/2}(\lambda_2)^{p/2}\exp\left(-\frac{\lambda_2}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\ &\qquad \cdot \pi(\sigma^2)\cdot\frac{\delta_1^{r_1}}{\Gamma(r_1)}(\lambda_1^2)^{r_1-1}\exp(-\delta_1\lambda_1^2)\cdot\frac{\delta_2^{r_2}}{\Gamma(r_2)}(\lambda_2)^{r_2-1}\exp(-\delta_2\lambda_2).\end{aligned} \tag{3.37}$$

Note that the tuning parameter $\lambda_2$ has gamma prior Gamma$(\lambda_2|r_2, \delta_2)$ (not $\lambda_2^2$).

The log posterior is proportionate to

$$\begin{aligned}\log \pi(\sigma^2) - \frac{n+2p}{2}\log(\sigma^2) + p\log\lambda_1 + \frac{p}{2}\log\lambda_2 - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\boldsymbol{y}\|^2 \\ - \frac{\lambda_1}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2 + (r_1-1)\log\lambda_1^2 - \delta_1\lambda_1^2 \\ + (r_2-1)\log\lambda_2 - \delta_2\lambda_2,\end{aligned} \tag{3.38}$$

and we transform the coordinate as

$$\phi_j = \frac{1}{\sqrt{\sigma^2}}\beta_j \ (j=1,\ldots,p), \quad \rho = \frac{1}{\sqrt{\sigma^2}}, \quad \lambda_j, \tag{3.39}$$

and (3.38) becomes

$$\begin{aligned}\log\pi\left(\frac{1}{\rho^2}\right) + (n+2p)\log(\rho) - \frac{1}{2}\|\rho\boldsymbol{y} - X\boldsymbol{\phi}\|^2 \\ - \lambda_1\sum_{j=1}^{p}|\phi_j| - \lambda_2\sum_{j=1}^{p}\phi_j^2 + \left(r_1 - 1 + \frac{p}{2}\right)\log\lambda_1^2 - \delta_1\sum_{j=1}^{p}\lambda_j^2 \\ + \left(r_2 - 1 + \frac{p}{2}\right)\log\lambda_2 - \delta_2\lambda_2,\end{aligned} \tag{3.40}$$

where $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_p)^T$. If $\pi(\sigma^2)$ is $1/\sigma^2$ or inverse gamma density, the first, second, fourth, fifth, seventh and ninth terms are concave in $(\boldsymbol{\phi},\rho,\lambda)$, and the third term is concave quadratic in $(\boldsymbol{\phi},\rho,\lambda)$. Sixth and eighth terms are concave in $(\boldsymbol{\phi},\rho,\lambda)$ when $r_1$ and $r_2$ are both greater than $1/2$.

Hence, if $r_k > 1/2$ $(k=1,2)$, (3.38) is concave, and the posterior of the Bayesian Adaptive lasso with prior Gamma$(\lambda_1^2|r_1,\delta_1)$ and Gamma$(\lambda_2|r_2,\delta_2)$ is unimodal.

### Unimodality of the Bayesian adaptive lasso

We can easily show the unimodality of the Bayesian elastic net with gamma prior Gamma$(\lambda_j^2|r,\delta)$ on $\lambda_j^2$.

The Bayesian adaptive lasso has the following likelihood and priors:

$$\begin{aligned} &\text{Likelihood: } (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\beta}\|^2\right), \\ &\text{Priors: } \prod_{j=1}^{p}\frac{\lambda_j}{2\sqrt{\sigma^2}}\exp\left(-\frac{\lambda_j}{\sigma^2}|\beta_j|\right)\cdot\pi(\sigma^2)\cdot\frac{\delta^r}{\Gamma(r)}(\lambda_j^2)^{r-1}\exp(-\delta\lambda_j^2). \end{aligned} \tag{3.41}$$

The log posterior is proportionate to

$$\begin{aligned} \log\pi(\sigma^2)-\frac{n+p}{2}\log(\sigma^2)+\sum_{j=1}^{p}\log\lambda_j-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{y}\|^2 \\ -\frac{1}{\sqrt{\sigma^2}}\sum_{j=1}^{p}\lambda_j|\beta_j|+(r-1)\sum_{j=1}^{p}\log\lambda_j^2-\delta\sum_{j=1}^{p}\lambda_j^2, \end{aligned} \tag{3.42}$$

and we transform the coordinate as

$$\phi_j=\frac{1}{\sqrt{\sigma^2}}\beta_j\ (j=1,\ldots,p),\quad \rho=\frac{1}{\sqrt{\sigma^2}},\quad \lambda_j, \tag{3.43}$$

and (3.42) becomes

$$\begin{aligned} \log\pi\left(\frac{1}{\rho^2}\right)+(n+p)\log(\rho)-\frac{1}{2}\|\rho\boldsymbol{y}-X\boldsymbol{\phi}\|^2 \\ -\sum_{j=1}^{p}\lambda_j|\phi_j|+\left(r-1+\frac{1}{2}\right)\sum_{j=1}^{p}\log\lambda_j^2-\delta\sum_{j=1}^{p}\lambda_j^2, \end{aligned} \tag{3.44}$$

where $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_p)^T$. As similar to the Bayesian lasso, if $\pi(\sigma^2)$ is $1/\sigma^2$ or inverse gamma density, the first, second, fourth and sixth terms are concave in $(\boldsymbol{\phi},\rho,\lambda)$. And the third term is concave quadratic in $(\boldsymbol{\phi},\rho,\lambda)$. Fifth term is concave in $(\boldsymbol{\phi},\rho,\lambda)$ when $r>1/2$. Hence, if $r>1/2$, (3.42) is concave, and the posterior of the Bayesian adaptive lasso with prior Gamma$(\lambda_j^2|r,\delta)$ is unimodal.

# Chapter 4

# Sparse modeling in the Bayesian lasso

## 4.1 Sparse algorithm in the Bayesian lasso

Since the Bayesian lasso enables us to treat the lasso from the Bayesian viewpoint, we can estimate the posterior distribution of the lasso. However, a crucial problem arises in the lack of the sparsity.

Although the lasso produces some coefficients exactly into zero, the Bayesian lasso does not. The cause arises from the estimation of the posterior distribution using MCMC method such as the Gibbs sampler (e.g., Bishop, 2006). In the Bayesian analysis, it is often hard to derive the posterior distribution analytically when the prior distribution is not conjugate. On the other hand, the MCMC procedure enables us to obtain the random sample from the posterior distribution even if there are no closed form of the posterior distribution. Thus, we can estimate the posterior using the MCMC.

Since it is hard to obtain the closed form of the Bayesian lasso, Park and Casella (2008) used the Gibbs sampler for the estimation of the posterior distribution. Bayesian lasso gives us the random sample from the posterior distribution of the lasso, and we can calculate the posterior mode, posterior median, and posterior mean from this sample. However, MCMC does not produce zero estimates of

coefficients, since a posterior mode estimated by MCMC is not equivalent to a mode of the true posterior distribution, exactly. Further, the posterior median and mean is not equivalent to the lasso estimate.

In order to overcome this problem, Hoshina (2012) proposed the sparse algorithm (SA). The proposed algorithm is given in Table 4.1. We focused the MAP estimation in the Bayesian lasso. If the estimated posterior mode is close to the true value enough, some components of it may be exactly zero. That is, the lack of sparsity is cased by poor estimation accuracy. We can evaluate the estimation accuracy of the MAP estimation by the posterior probability, and if there is some estimate that have larger posterior probability than current MAP estimate, we can employ it as new MAP estimate. SA is based on this idea. After MCMC process, the SA gives zero values for some components of estimated coefficient vector such that the posterior probability becomes large. An outline of SA is given in Fig. 4.1.

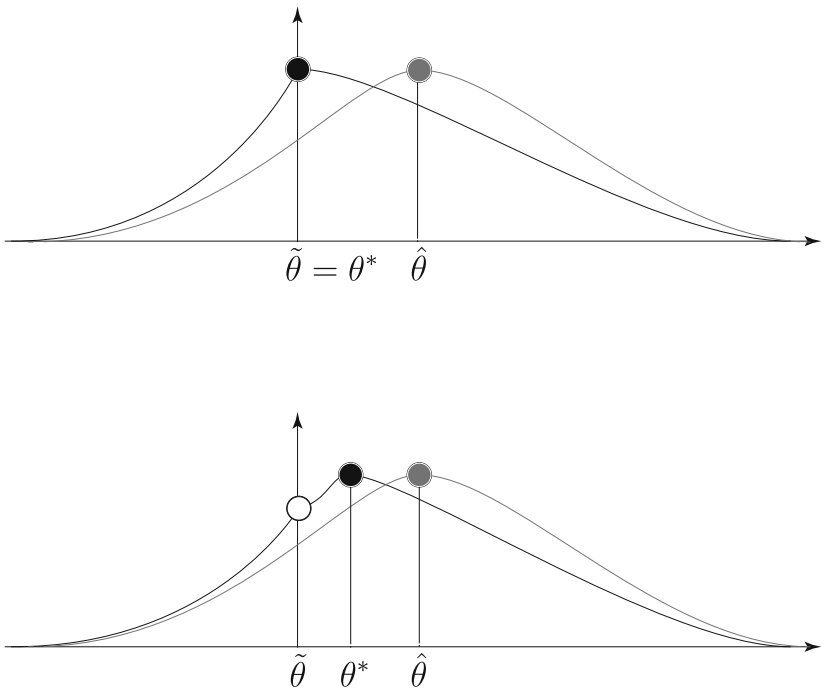



Fig. 4.1 Illustration of the sparse algorithm (Hoshina, 2012): The real line and the black circle are the true posterior density and the true posterior mode $\theta^*$. The dashed line and the grey circle are the estimated posterior density and the estimated posterior mode $\hat{\theta}$. Let $\tilde{\theta} = 0$. Then, we employ $\tilde{\theta}$ as the point estimates if $\tilde{\theta}$ has larger posterior probability than $\hat{\theta}$. On the other hand, we employ $\hat{\theta}$ as the point estimates if $\hat{\theta}$ has larger posterior probability than $\tilde{\theta}$.

Table 4.1 Sparse algorithm (Hoshina, 2012).

| **Sparse algorithm** | |
|---|---|
| 1. | Estimate the coefficient vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ |
| 2. | $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T \leftarrow \hat{\boldsymbol{\beta}}$ |
| 3. | For $j = 1, \ldots, p$, |
| | set $\tilde{\beta}_j \leftarrow 0$ |
| 3.1 | if $g(\tilde{\beta}, \hat{\boldsymbol{\xi}}, \boldsymbol{y}) \geq g(\hat{\beta}, \hat{\boldsymbol{\xi}}, \boldsymbol{y})$ then $\hat{\beta}_j \leftarrow \tilde{\beta}_j$ |
| 3.2 | else $\hat{\beta}_j \leftarrow \hat{\beta}_j$ |
| | where $g(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{y}) = \log f(\boldsymbol{y}\|\boldsymbol{\beta}, \boldsymbol{\xi}) + \log \pi(\boldsymbol{\beta}, \boldsymbol{\xi})$, $f(\boldsymbol{y}\|\boldsymbol{\beta}, \boldsymbol{\xi})$ is a likelihood, $\pi(\boldsymbol{\beta}, \boldsymbol{\xi})$ is a prior on $(\boldsymbol{\beta}, \boldsymbol{\xi})$, and $\hat{\boldsymbol{\xi}}$ is point estimates of the parameter vector $\boldsymbol{\xi} = (\sigma^2, \tau_1^2, \ldots, \tau_p^2)^T$. |

One advantage of the SA is that enables us to obtain sparse MAP estimates of the Bayesian lasso. However, this procedure only corrects for the resulting point estimates, and the numerically-computed MAP estimates are often instable. Fig. 4.2 represents the solution paths of the diabetes data (Efron *et al.* , 2004). The point estimates are the posterior mode(=MAP), median and mean of the Bayesian lasso, respectively. This figure shows the instability of the MAP estimates. To overcome this drawback, we propose another procedure, the MAP Bayesian lasso, in Section 4.3.

## 4.2 aPIC criterion for the Bayesian lasso

In the Bayesian lasso, the value of the tuning parameter $\lambda$ controls the strength of an impact of the Laplace prior on the model, and the resulting model depends on the value of $\lambda$. To choose the values of $\lambda$, Park and Casella (2008) proposed two approaches; the empirical Bayes based on maximizing the marginal likelihood and the hierarchical Bayes.

The hierarchical Bayes approach such as the MAP procedure evaluates the likelihood and the prior information, and we can avoid the overfitting because of the prior. On the other hand, the marginal likelihood evaluates the estimation accuracy of the estimated model in terms of the parameter space, and it is known that the marginal likelihood also enables us to avoid the overfitting. However, it does

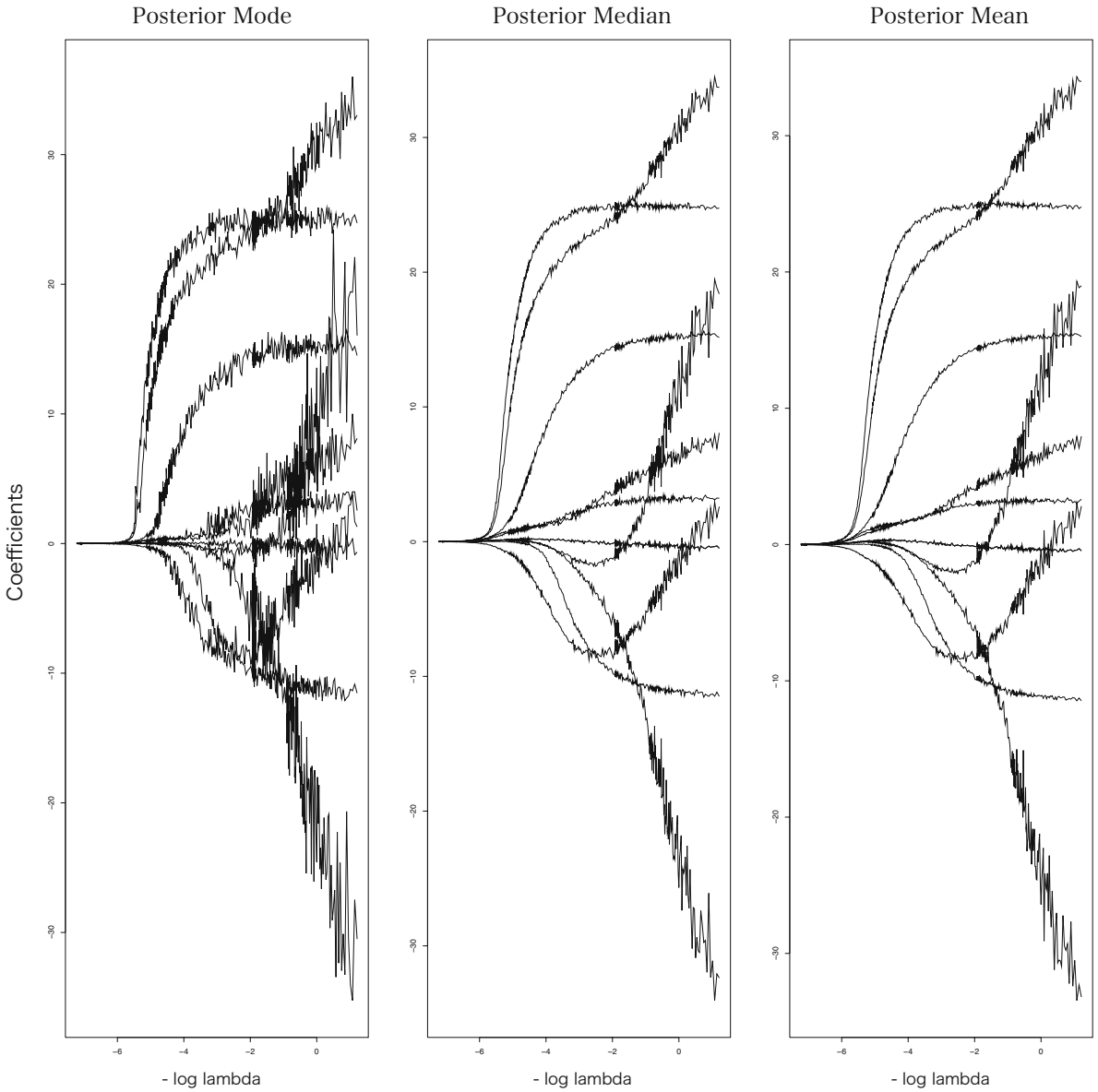



Fig. 4.2 Solution paths of the diabetes data of Efron *et al.* (2004): The posterior mode (left), median (center) and mean (right) of the Bayesian lasso for the diabetes data are represented. Each Bayesian lasso estimates were computed over a grid of $\lambda$ values, using 10000 Gibbs sample (after 1000 burn in) for each $\lambda$.

not evaluate the prediction accuracy.

In the Bayes statistics, the Bayesian predictive distribution has an information from a predictive point of view. Thus, we propose a model selection criterion for evaluating a Bayesian predictive distribution for the Bayesian lasso (Kawano *et al.* , 2015).

### 4.2.1 aPIC criterion

Kitagawa (1997) proposed the predictive information criterion (PIC) for evaluating the Bayesian predictive distribution. A Bayesian predictive distribution is, in general, given by

$$h(\boldsymbol{z}|\boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{4.1}$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is an $n$-dimensional future observation, $f(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(z_i|\boldsymbol{\theta})$ is the likelihood, $\boldsymbol{\theta}$ is a parameter vector, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior distribution on $\boldsymbol{\theta}$ defined by

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{4.2}$$

Using the Bayesian predictive distribution, Kitagawa (1997) derived the predictive information criterion (PIC)

$$\mathrm{PIC} = -2 \log h(\boldsymbol{y}|\boldsymbol{y}) + 2B_p, \tag{4.3}$$

where $B_p$ is the bias term given by

$$B_p = E_{q(\boldsymbol{y})} \left[ \log h(\boldsymbol{y}|\boldsymbol{y}) - E_{q(\boldsymbol{z})} \left[ \log h(\boldsymbol{z}|\boldsymbol{y}) \right] \right] \tag{4.4}$$

with $q(\cdot)$ being the true distribution that generates the data.

In order to derive PIC for the Bayesian lasso, we obtain the Bayesian predictive distribution in (4.1). In the Bayesian lasso, the prior distribution is formulated by

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left( -\frac{\lambda}{\sqrt{\sigma^2}} |\beta_j| \right). \tag{4.5}$$

It is, however, difficult to obtain the predictive distribution $h(\boldsymbol{z}|\boldsymbol{y})$ based on this prior in closed form, since it is difficult to analytically represent the form of the posterior distribution. This problem arises from the fact that the prior distribution

$\pi(\boldsymbol{\beta}|\sigma^2)$ is not a conjugate prior for the likelihood function. In Section 4.2.2, we approximate the prior distribution $\pi(\boldsymbol{\beta}|\sigma^2)$ by a conjugate prior distribution (a normal prior distribution) for the likelihood function.

### 4.2.2 Approximated prior distribution

Let $f(\beta)$ be the Laplace distribution

$$f(\beta) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta|}{\sqrt{\sigma^2}}\right), \tag{4.6}$$

and $g(\beta|\alpha^2)$ be the normal distribution

$$g(\beta|\alpha^2) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{\beta^2}{2\alpha^2}\right), \tag{4.7}$$

where $\alpha$ is positive.

Our aim is to find the normal distribution that is the closest to the Laplace distribution. Here, we measure the closeness between the distributions in terms of the Kullback-Leibler information (Kullback and Leibler, 1951). We determine the normal distribution $g(\beta|\hat{\alpha}^2)$, where $\hat{\alpha}^2$ is an estimator of $\alpha^2$, such that the Kullback-Leibler information between the distributions $f(\beta)$ and $g(\beta|\alpha^2)$;

$$\mathrm{KL}(f, g) = \int_{-\infty}^{\infty} f(\beta) \log \frac{f(\beta)}{g(\beta|\alpha^2)} d\beta \tag{4.8}$$

is minimized with respect to the parameter $\alpha^2$.

Theorem 4.2.1 The minimum of the Kullback-Leibler information (4.8) attains at $\hat{\alpha}^2 = 2(\sqrt{\sigma^2}/\lambda)^2$.

*Proof.* The Kullback-Leibler information between $f(\beta)$ and $g(\beta|\alpha^2)$ is calculated as

$$\mathrm{KL}(f, g) = \log\lambda - \log(2\sqrt{\sigma^2}) + \frac{1}{2}\log\left(2\pi\alpha^2\right) - 1 + \frac{1}{\alpha^2}\left(\frac{\sqrt{\sigma^2}}{\lambda}\right)^2. \tag{4.9}$$

A minimizer of (4.9) is $\hat{\alpha}^2 = 2(\sqrt{\sigma^2}/\lambda)^2$, which is obtained by solving the equation

$\partial \mathrm{KL}(f, g)/\partial \alpha^2 = 0.$

□

From this result, the Laplace distribution $f(\beta)$ can be approximated by the normal distribution $g(\beta|\hat{\alpha}^2)$, and we have

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2) &= \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left[-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right] \\ &\approx \tilde{\pi}(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{\sqrt{2\pi(2\sigma^2)}} \exp\left[-\frac{\lambda^2\beta_j^2}{2(2\sigma^2)}\right]. \end{aligned} \tag{4.10}$$

The approximated distribution $\tilde{\pi}(\boldsymbol{\beta}|\sigma^2)$ can be regarded as the closest to the Laplace distribution $\pi(\boldsymbol{\beta}|\sigma^2)$ in terms of the Kullback-Leibler information. Fig. 4.3 illustrates the case with $p = 1$ and $\lambda/\sqrt{\sigma^2} = 1$.

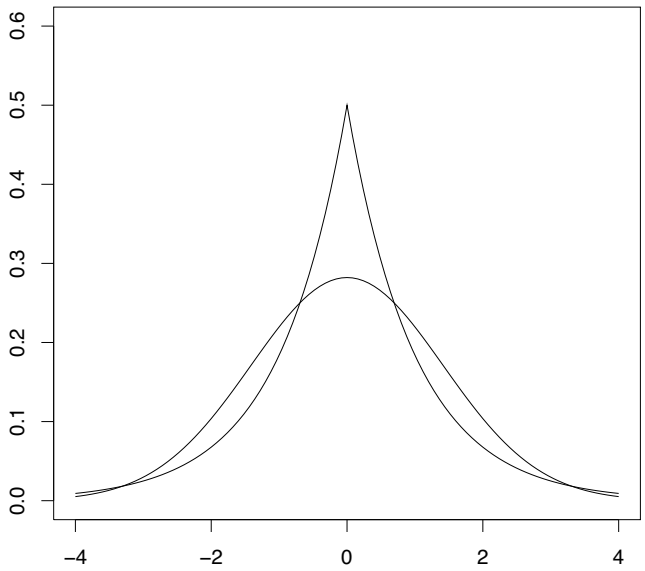



Fig. 4.3 Laplace distribution and the closest normal distribution: Dashed line is the Laplace distribution with rate 1, and real line is the closest normal distribution N(0, 1/2).

Note that the approximated distribution is employed only when we obtain a model selection criterion, and that the Laplace distribution is employed when we estimate the coefficient parameters.

### 4.2.3 Bayesian predictive distribution for Bayesian lasso

Using the approximated prior distribution $\tilde{\pi}(\boldsymbol{\beta}|\sigma^2)$ in (4.10) on $\boldsymbol{\beta}$ and assuming an inverse gamma distribution $\pi(\sigma^2) = \mathrm{IG}(\nu_0/2, \eta_0/2)$ on $\sigma^2$, we derive the joint prior distribution $\pi(\boldsymbol{\beta}, \sigma^2)$ in the form

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2) \approx \tilde{\pi}(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2) = \tilde{\pi}(\boldsymbol{\beta}, \sigma^2). \tag{4.11}$$

From the approximated prior distribution and Bayes' rule, the approximated joint posterior distribution can be expressed as

$$\tilde{p}(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \tilde{p}(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})\ \tilde{p}(\sigma^2|\boldsymbol{y}), \tag{4.12}$$

where each approximated posterior distribution is given by

$$\tilde{p}(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) = N_p(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, \sigma^2 A), \qquad \tilde{p}(\sigma^2|\boldsymbol{y}) = \mathrm{IG}\left(\sigma^2|\frac{\nu_1}{2}, \frac{\eta_1}{2}\right). \tag{4.13}$$

Here,

$$\begin{aligned} A &= \left(X^T X + \frac{\lambda^2}{2} I_p\right)^{-1}, \\ \tilde{\boldsymbol{\beta}} &= A X^T \boldsymbol{y}, \\ \nu_1 &= n + \nu_0, \\ \eta_1 &= \eta_0 + \boldsymbol{y}^T \boldsymbol{y} - \tilde{\boldsymbol{\beta}}^T A^{-1} \tilde{\boldsymbol{\beta}}. \end{aligned} \tag{4.14}$$

Note that if the prior distribution $\pi(\boldsymbol{\beta}|\sigma^2)$ in (4.5) is used instead of the approximated prior distribution $\tilde{\pi}(\boldsymbol{\beta}|\sigma^2)$, it is difficult to obtain the posterior distribution $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y})$.

Using the approximated posterior distributions, we obtain the Bayesian predic-

tive distribution for the Bayesian lasso given by

$$\begin{aligned} h(\boldsymbol{z}|\boldsymbol{y}) &= \int f(\boldsymbol{z}|\boldsymbol{\beta},\sigma^2)\ p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta}d\sigma^2 \\ &= \frac{\Gamma\left(\frac{n+\nu_1}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)(\pi\nu_1)^{n/2}}|\tilde{\Sigma}|^{-1/2}\left[1+\frac{1}{\nu_1}(\boldsymbol{z}-X\tilde{\boldsymbol{\beta}}_n)^T\tilde{\Sigma}^{-1}(\boldsymbol{z}-X\tilde{\boldsymbol{\beta}}_n)\right]^{-(n+\nu_1)/2}, \end{aligned} \tag{4.15}$$

where $\tilde{\Sigma}=(\eta_1/\nu_1)(XAX^T+I_n)$ and $\Gamma(\cdot)$ is the Gamma function. This predictive distribution is an $n$-dimensional $t$-distribution with $\nu_1$ degrees of freedom.

### 4.2.4 Proposed criterion: aPIC

To derive the PIC type criterion, we need to calculate the bias term (4.4) for the Bayesian predictive distribution $h(\boldsymbol{z}|\boldsymbol{y})$ (4.15). It is still difficult to calculate the bias term analytically, because the Bayesian predictive distribution $h(\boldsymbol{z}|\boldsymbol{y})$ in (4.15) is an $n$-dimensional $t$-distribution. Hence, we approximate the distribution $h(\boldsymbol{z}|\boldsymbol{y})$ by a normal distribution $f(\boldsymbol{z}|\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)$ in the form

$$h(\boldsymbol{z}|\boldsymbol{y}) = f(\boldsymbol{z}|\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)\left\{1+O_p(n^{-1})\right\}, \tag{4.16}$$

where $\tilde{\sigma}^2$ is given by

$$\tilde{\sigma}^2 = \frac{(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})+\dfrac{\lambda^2}{2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}}+\eta_0}{n+p+\nu_0+2}. \tag{4.17}$$

This approximation is based on the Laplace approximation (Tierney and Kanade, 1986). For details of this approximation, we refer to Konishi and Kitagawa (2008).

For the approximated predictive distribution $f(\boldsymbol{z}|\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)$ in (4.16), we define an approximated predictive information criterion (aPIC) as follows:

$$\text{aPIC} = -2\log h(\boldsymbol{y}|\boldsymbol{y}) + 2B_p^*, \tag{4.18}$$

where the approximated bias term $B_p^*$ is given by

$$\begin{aligned} B_p^* &= E_{q(\boldsymbol{y})}\left[\log f(\boldsymbol{y}|\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2) - E_{q(\boldsymbol{z})}\{\log f(\boldsymbol{z}|\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)\}\right] \\ &\approx -\frac{1}{2\tilde{\sigma}^2}\left[E_{q(\boldsymbol{y})}[(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}}) - E_{q(\boldsymbol{z})}\{(\boldsymbol{z}-X\tilde{\boldsymbol{\beta}})^T(\boldsymbol{z}-X\tilde{\boldsymbol{\beta}})\}]\right]. \end{aligned} \tag{4.19}$$

Using the results of Kitagawa (1997) and Kim *et al.* (2012), we can calculate the approximated bias term as

$$B_p^* \approx \left(\frac{\sigma^{*2}}{\tilde{\sigma}^2}\right) \mathrm{tr}\left[X\left(X^TX + \frac{n^2\lambda^2}{2}I_p\right)^{-1}X^T\right], \tag{4.20}$$

where $\sigma^{*2}$ is a specific value such that $q(\boldsymbol{z}) = f(\boldsymbol{z}|\boldsymbol{\beta}^*,\sigma^{*2})$.

Then we obtain aPIC in the form

$$\begin{aligned} \mathrm{aPIC} = &-2\log\Gamma\left(\frac{n+\nu_1}{2}\right) + 2\log\Gamma\left(\frac{\nu_1}{2}\right) + n\log(\pi\nu_1) + \log\left|\tilde{\Sigma}\right| \\ &+ (n+\nu_1)\log\left[1+\frac{1}{\nu_n}(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})^T\tilde{\Sigma}^{-1}(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})\right] \\ &+ 2\left(\frac{\sigma^{*2}}{\tilde{\sigma}^2}\right)\mathrm{tr}\left[X\left(X^TX+\frac{\lambda^2}{2}I_p\right)^{-1}X^T\right]. \end{aligned} \tag{4.21}$$

Since the value of $\sigma^{*2}$ is generally unknown, we replace $\sigma^{*2}$ by the mode of the posterior distribution $\tilde{\sigma}^2$, and have

$$\begin{aligned} \mathrm{aPIC} = &-2\log\Gamma\left(\frac{n+\nu_1}{2}\right) + 2\log\Gamma\left(\frac{\nu_1}{2}\right) + n\log(\pi\nu_1) + \log\left|\tilde{\Sigma}\right| \\ &+ (n+\nu_1)\log\left[1+\frac{1}{\nu_1}(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})^T\tilde{\Sigma}^{-1}(\boldsymbol{y}-X\tilde{\boldsymbol{\beta}})\right] \\ &+ 2\mathrm{tr}\left[X\left(X^TX+\frac{\lambda^2}{2}I_p\right)^{-1}X^T\right]. \end{aligned} \tag{4.22}$$

The value of the hyperparameter $\lambda$ is selected as the minimizer of aPIC in (4.22). Some numerical results about aPIC are reported in Section 6.1.

## 4.3 MAP Bayesian lasso

To obtain the sparse MAP estimates of $\boldsymbol{\beta}$, the optimization methods such as any gradient procedures are required. However, it is difficult to obtain the posterior density function for the Bayesian lasso, and it may not be differentiable at $\beta = 0$ since it includes the Laplace prior. To overcome these drawbacks, we approximate the posterior density by the Monte Carlo integration, and propose a procedure that enables us to obtain the MAP estimates of the Bayesian lasso by Newton's method.

### 4.3.1 Posterior distribution approximated by Monte Carlo integration

Since the Bayesian lasso gives us the estimates of $\sigma^2$ and $\lambda$, our procedure leverages these estimates. Let $\hat{\sigma}^2$ and $\hat{\lambda}$ be the MAP estimates of $\sigma^2$ and $\lambda$, respectively. Then the (conditional) posterior density of $\boldsymbol{\beta}$ given $\hat{\sigma}^2$ and $\hat{\lambda}$ is proportionate to

$$\begin{aligned}
&\int \cdots \int N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \hat{\sigma}^2 I_n) \cdot N_p(\boldsymbol{\beta}|\mathbf{0}_p, \hat{\sigma}^2 D) \left\{ \prod_{j=1}^{p} \mathrm{Exp}\left( \tau_j^2 \middle| \frac{\hat{\lambda}^2}{2} \right) \right\} d\tau_1^2 \cdots \tau_p^2 \\
&\quad \propto \int \cdots \int N_p(\boldsymbol{\beta}|A^{-1}X^T\boldsymbol{y}, \hat{\sigma}^2 A^{-1}) \cdot |D|^{-1/2} \cdot |A|^{-1/2} \\
&\qquad\qquad \cdot \exp\left\{ -\frac{1}{2\hat{\sigma}^2} \boldsymbol{y}^T (I_n - XA^{-1}X^T)\boldsymbol{y} \right\} \left\{ \prod_{j=1}^{p} \mathrm{Exp}\left( \tau_j^2 \middle| \frac{\hat{\lambda}^2}{2} \right) \right\} d\tau_1^2 \cdots \tau_p^2.
\end{aligned} \tag{4.23}$$

It is difficult to evaluate the integration in (4.23) because of complexity of integrand. In general, some approximation methods, such as the Laplace approximation (Tierny and Kadane, 1986), may be used to approximate it. We cannot, however, employ this procedure since the integrand in (4.23) is not differentiable at $\beta_j = 0$.

In contrast, the Monte Carlo integration is applicable for posterior approximation. The Monte Carlo integration is a well-known numerical technique to

approximate a integration in statistics.

Let $\{\tau^2_{1(m)}, \ldots, \tau^2_{p(m)} : m = 1, \ldots, M\}$ be a random sample generated from $\prod_{j=1}^{p} \mathrm{Exp}(\tau_j^2|\hat{\lambda}^2/2)$ artificially, where size $M$ is encouraged to determine sufficiently large number. Then, we have the following approximation of (4.23):

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M} & N_p(\boldsymbol{\beta}|A^{-1}_{(m)}X^T\boldsymbol{y}, \hat{\sigma}^2 A^{-1}_{(m)}) \\ & \cdot |D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2} \cdot \exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n - XA^{-1}_{(m)}X^T)\boldsymbol{y}\right\},\end{aligned} \tag{4.24}$$

where $D_{(m)} = \mathrm{diag}(\tau^2_{1(m)}, \ldots, \tau^2_{p(m)})$, $A_{(m)} = X^TX + D^{-1}_{(m)}$. Since (4.24) is formed as the sum of differentiable function, (4.24) is totally differentiable. Hence, the posterior mode of the Bayesian lasso regression coefficients are given by maximizing (4.24) using Newton's method.

Thus, the approximated posterior distribution $\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, X, \lambda, \sigma^2)$ and the approximated marginal likelihood $\tilde{p}(\boldsymbol{y}|X, \sigma^2, \lambda)$ of the lasso are respectively given by

$$\begin{aligned}\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, X, \hat{\sigma}^2, \hat{\lambda}) &= \frac{\frac{1}{M}\sum_{m=1}^{M} N_p(\boldsymbol{\beta}|A^{-1}_{(m)}X^T\boldsymbol{y}, \hat{\sigma}^2A^{-1}_{(m)}) \cdot \xi_{(m)}}{\int \frac{1}{M}\sum_{\ell=1}^{M} N_p(\boldsymbol{\beta}|A^{-1}_{(\ell)}X^T\boldsymbol{y}, \hat{\sigma}^2A^{-1}_{(\ell)}) \cdot \xi_{(\ell)} d\boldsymbol{\beta}} \\ &= \sum_{m=1}^{M} \gamma_{(m)} N_p(\boldsymbol{\beta}|A^{-1}_{(m)}X^T\boldsymbol{y}, \hat{\sigma}^2A^{-1}_{(m)}), \\ \tilde{p}(\boldsymbol{y}|X, \hat{\sigma}^2, \hat{\lambda}) &= \int \frac{1}{M}\sum_{m=1}^{M} N_p(\boldsymbol{\beta}|A^{-1}_{(m)}X^T\boldsymbol{y}, \hat{\sigma}^2A^{-1}_{(m)}) \cdot \xi_{(m)} d\boldsymbol{\beta} \\ &= \frac{1}{M}\sum_{m=1}^{M} |D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2} \\ &\qquad \cdot \exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n - XA^{-1}_{(m)}X^T)\boldsymbol{y}\right\},\end{aligned} \tag{4.25}$$

where

$$\xi_{(m)} = |D_{(m)}|^{-1/2}|A_{(m)}|^{-1/2} \cdot \exp\left\{-\frac{1}{2\hat{\sigma}^2}\boldsymbol{y}^T(I_n - XA_{(m)}^{-1}X^T)\boldsymbol{y}\right\},$$
$$\gamma_{(m)} = \frac{\xi_{(m)}}{\sum_{\ell=1}^{M}\xi_{(\ell)}}.$$

Note that, the approximated posterior of the lasso is given in the form of a mixture of normal distributions with mixture weights $\gamma_{(1)}, \ldots, \gamma_{(M)}$.

### 4.3.2 MAP estimation by Newton's method

Newton's method (e.g., Murphy, 2012) is one of the second order optimization methods that take the Hessian, i.e. the curvature of the space, into account. This iterative algorithm consists of updates of the following form:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta_k H_k^{-1}\boldsymbol{g}_k, \quad \boldsymbol{g}_k = \frac{\partial f(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}}, \quad H_k = \frac{\partial^2 f(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T},$$

where $\boldsymbol{\theta}_k$ $(k = 1, \ldots)$ is a sequence of variables which converges to the optimal value $\hat{\boldsymbol{\theta}}$, $f(\boldsymbol{\theta})$ is a function which is maximized, and $\eta_k$ is a step size for $k$-th update.

In our procedure, the resulting regression coefficients are given by maximizing (4.24) or $\tilde{p}(\boldsymbol{\beta}|\boldsymbol{y}, X, \hat{\sigma}^2, \hat{\lambda})$ of (4.25). We use (4.24) as the objective function of the maximization problem, and the gradient $\boldsymbol{g}_k$ and the Hessian $H_k$ for $k$-th update

are respectively given as follows:

$$\begin{aligned}
\boldsymbol{g}_k &= \frac{1}{M}(2\pi)^{-p/2}(\hat{\sigma}^2)^{-(p+2)/2}\sum_{m=1}^{M}|D_{(m)}|^{-1/2} \\
&\qquad \cdot \exp\left\{-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta}_k + \boldsymbol{\beta}_k D_{(m)}^{-1}\boldsymbol{\beta}_k)\right\}(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k), \\
H_k &= \frac{1}{M}(2\pi)^{-p/2}(\hat{\sigma}^2)^{-(p+2)/2} \\
&\qquad \cdot \sum_{m=1}^{M}|D_{(m)}|^{-1/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta}_k + \boldsymbol{\beta}_k D_{(m)}^{-1}\boldsymbol{\beta}_k)\right\} \\
&\qquad \cdot \left\{A_{(m)} + \frac{1}{\hat{\sigma}^2}(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k)(X^T\boldsymbol{y} - A_{(m)}\boldsymbol{\beta}_k)^T\right\}.
\end{aligned} \tag{4.26}$$

We choose the value of step size $\eta_k$ from candidate values $\{\eta_k^{(1)}, \ldots, \eta_k^{(\ell)}\}$ so that $\boldsymbol{\theta}_{k+1} = \boldsymbol{\beta}_{k+1}$ has the largest posterior density, and we substitute the following function for the posterior density of $\boldsymbol{\beta}$:

$$q(\boldsymbol{\beta}, \boldsymbol{y}, X, \sigma^2, \lambda) = \log N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n) + \sum_{j=1}^{p}\log\left\{\frac{\lambda}{\sqrt{2\sigma^2}}\exp\left(-\frac{\lambda}{\sqrt{\sigma^2}}|\beta_j|\right)\right\}. \tag{4.27}$$

We use this formula to obtain the MAP estimates of the Bayesian lasso. However, it is difficult to derive sparse solutions for regression coefficients since we use a numerical procedure. For this problem, we can apply the sparse algorithm (Hoshina, 2012), which sets some regression coefficients exactly zero so that a posterior probability becomes large.

Although this procedure enables us to obtain the sparse MAP estimates of the Bayesian lasso, the optimized solution of Newton's method depends on the initial value. Especially, since objective function of this optimization may be waggly, it seems that many local optimums exist as shown in Fig. 4.4. To avoid this problem, the initial value selection is very important. We employ the posterior means as the initial value of the Newton's method because of its estimation stability, as shown

in Fig. 4.5.

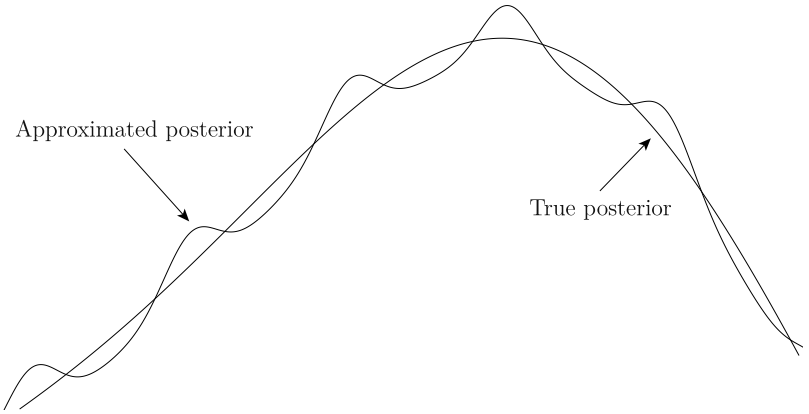



Fig. 4.4 Overview of the objective function of our procedure. Solid and dashed lines illustrate the approximated posterior and true posterior, respectively. Even if true posterior has no local maximum, the approximated posterior may have many local maximums. Thus, it is desired that the initial value of Newton's method is slightly near the global maximum.

The size of numerical integration $M$ may affect the result of our procedure. For this point, an empirical evidence shows that the size of $M$ also suffices at the relatively-small value. Figure 4.5 shows the solution paths in cases of $M = 50$, 500, 5000 respectively, and all solution paths are similar. From these results, we set $M$ to 500 in numerical studies of Chapter 6.2.

We call this procedure the "MAP Bayesian lasso" (Maximum a Approximated Posteriori with the Bayesian lasso). The details of the proposed procedure are given in Algorithm 3.

### 4.3.3 Other procedures

This section describes other sparse model building techniques which choose the value of a tuning parameter by model selection criteria.

#### Baysian lasso with model selection criteria

Suppose that $p(\boldsymbol{y}|\boldsymbol{\theta})$ is a likelihood of $n$-observation vector $\boldsymbol{y}$ on parameter $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is a posterior density of $\boldsymbol{\theta}$. Deviance information criterion (DIC) proposed

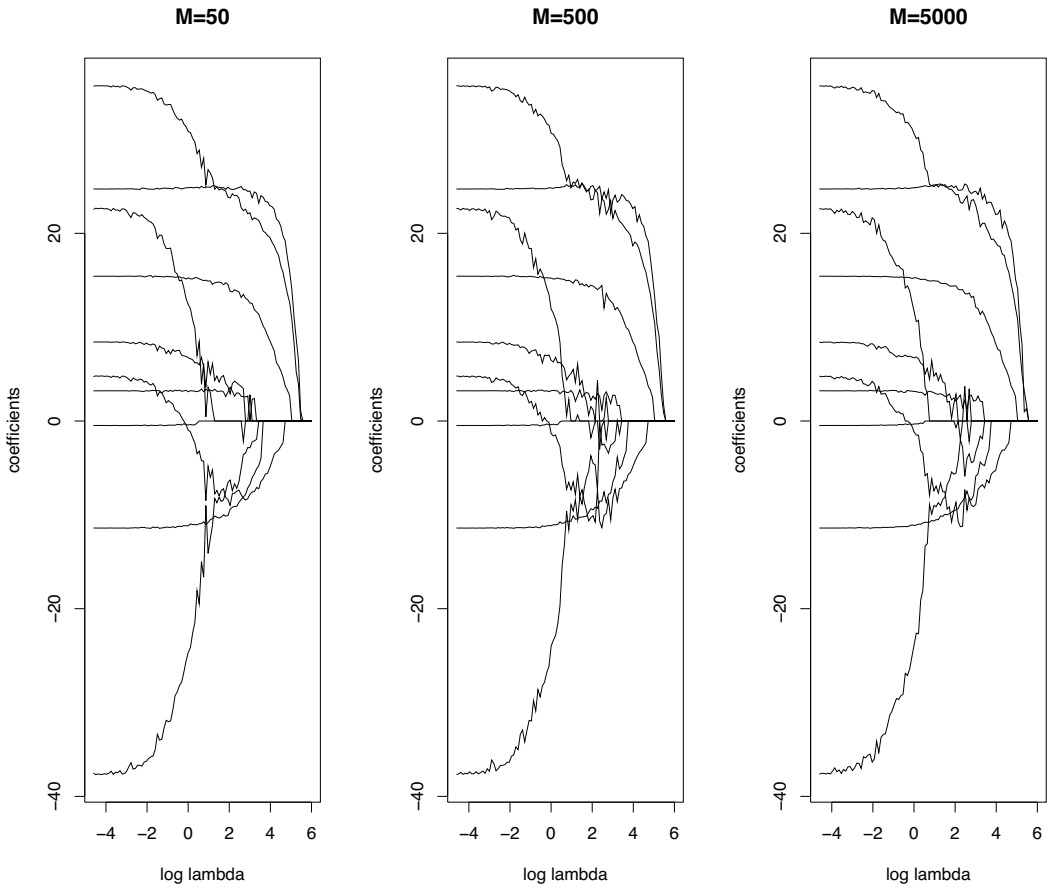



Fig. 4.5 Regularization paths for the diabetes data (Efron *et al.* , 2004) for $M = 50$ (left), $M = 500$ (center) and $M = 5000$ (right).

by Spiegelhalter *et al.* (2002) measures the effective number of parameters in a Bayesian model using an information theoretic argument. The measure $p_D$ for parameter $\boldsymbol{\theta}$ is defined by

$$p_D = -2E_{\boldsymbol{\theta}|\boldsymbol{y}}[\log p(\boldsymbol{y}|\boldsymbol{\theta})] + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}),$$

where $E_{\boldsymbol{\theta}|\boldsymbol{y}}(\cdot)$ denotes the expectation over posterior distribution of $\boldsymbol{\theta}$, and $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$.

Based on this measure, Spiegelhalter *et al.* (2002) proposed a deviance information criterion

$$\mathrm{DIC} = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 2p_D.$$

**Algorithm 3** MAP Bayesian lasso

$\sigma^2 \Leftarrow \hat{\sigma}^2$: posterior mode of $\sigma^2$;
$\lambda \Leftarrow \hat{\lambda}$: posterior mode of $\lambda$;
Initialize $\boldsymbol{\beta}_0 = \bar{\boldsymbol{\beta}}$ : posterior mean;
**for** $k = 1, 2, \ldots$ until convergence **do**
    Evaluate the gradient $\boldsymbol{g}_k$ of (4.26);
    Evaluate the Hessian $H_k$ of (4.26);
    Solve $\boldsymbol{z}_k = H_k^{-1}\boldsymbol{g}_k$;
    **for** $\ell = 1, 2, \ldots, L$, solve $\boldsymbol{\beta}_{k+1(\ell)} = \boldsymbol{\beta}_k + \eta_{k(\ell)}\boldsymbol{z}_k$ **do**
        Evaluate the value $q_{(\ell)} = q(\boldsymbol{\beta}_{k+1(\ell)}, \boldsymbol{y}, X, \sigma^2, \lambda)$ of (4.27);
    **end for**
    $\boldsymbol{\beta}_{k+1} \Leftarrow \underset{\boldsymbol{\beta}_{k+1(\ell)}}{\operatorname{argmax}}\{q_{(\ell)}\}$;
    $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T \Leftarrow \boldsymbol{\beta}_{k+1}$;
**end for**
$\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p) \Leftarrow \hat{\boldsymbol{\beta}}$;
**for** $j = 1, 2, \ldots, p$ **do**
    $\tilde{\beta}_j \Leftarrow 0$;
    **if** $q(\tilde{\boldsymbol{\beta}}, \boldsymbol{y}, X, \sigma^2, \lambda) > q(\hat{\boldsymbol{\beta}}, \boldsymbol{y}, X, \sigma^2, \lambda)$ **then**
        $\hat{\boldsymbol{\beta}} \Leftarrow \tilde{\boldsymbol{\beta}}$;
    **else** $\tilde{\boldsymbol{\beta}} \Leftarrow \hat{\boldsymbol{\beta}}$;
    **end if**
**end for**

Widely applicable or Watanabe-Akaike information criterion (WAIC) is proposed by Watanabe (2010a, 2010b). WAIC intends to evaluate the model accuracy by the Bayes or Gibbs generalization loss for singular or non-singular model. However, it is difficult to obtain these losses since we need to evaluate a expectation on predictive distribution. For this problem, Watanabe (2010a, 2010b) showed that the consistent estimator of the Bayes generalization loss is given by

$$\begin{aligned}\mathrm{WAIC} = &-\frac{1}{n}\sum_{i=1}^{n}\log E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[p(\boldsymbol{y}|\boldsymbol{\theta})\right] \\ &+\frac{1}{n}\sum_{i=1}^{n}\left\{E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\left(\log p(y_i|\boldsymbol{\theta})\right)^2\right] - E_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\log p(y_i|\boldsymbol{\theta})\right]^2\right\}.\end{aligned}$$

DIC and WAIC need to evaluate the posterior and predictive distribution respectively. The Gibbs sampler enables us to derive these values, and the Bayesian lasso which gives us the Gibbs sample of the lasso can be applicable for these procedures.

### Lasso with model selection criteria

The degrees of freedom can lead to several model selection criteria (e.g. Hirose *et al.*, 2013) which may improve prediction accuracy in the lasso.

In the lasso, Zou *et al.* (2007) introduced the AIC (Akaike, 1973), the BIC (Schwarz, 1978) and the Mallows' $C_p$ (Mallows, 1973), respectively, given by

$$\begin{aligned}
\mathrm{AIC} &= n\log(2\pi\hat{\sigma}^2) + \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{2\hat{\sigma}^2} + 2\mathrm{DF}, \\
\mathrm{BIC} &= n\log(2\pi\hat{\sigma}^2) + \frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{2\hat{\sigma}^2} + \log n \cdot \mathrm{DF}, \\
C_p &= \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 + 2\hat{\sigma}^2\mathrm{DF},
\end{aligned}$$

where the likelihood of $\boldsymbol{y}$ is given by $N_n(\boldsymbol{y}|X\boldsymbol{\beta}, \sigma^2 I_n)$ and DF is the degrees of freedom of the lasso. Although true value of DF is unknown, Zou *et al.* (2007) showed that the number of non-zero coefficients of the lasso estimate is an unbiased estimator of DF. The AIC and $C_p$ yield the same results when the same estimated $\sigma^2$ is used.

Hirose *et al.* (2013) also introduced the generalized cross validation (GCV; Craven and Wahba, 1979)

$$\mathrm{GCV} = n\frac{\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2}{(n - \mathrm{DF})^2}.$$

Note that the GCV does not need estimate of $\sigma^2$.

# Chapter 5

# Model selection in elastic net via Bayes model

The Bayesian lasso and other Bayesian extensions of the $L_1$ regularizations use the relationship between the regularizations and the Bayes model. The Bayesian information criterion (BIC; Schwarz, 1978) and the generalized Bayesian information criterion (GBIC; Konishi *et al.*, 2004) have also been derived based on the same relationship, and they evaluate the posterior probability of the models.

The GBIC, which is extension of the BIC, is applicable for the regularization procedure, while the BIC is not. However, the GBIC depends on the sample size since it evaluates the posterior density using the Laplace approximation (Tierney and Kadane 1986).

In contrast to this, we propose a model selection criterion, which evaluates the prediction accuracy of resulting models based on the Bayes model.

## 5.1 Bayes model of the elastic net

We consider the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an $n$-dimensional response vector, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ design matrix, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the $p$-dimensional observations for predictor variables, the elements of $\boldsymbol{x}_i$ are given as $x_{i1}, \ldots, x_{ip}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p$-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an $n$-dimensional error vector which elements have independent and identically distributed according to a normal distribution with mean zero and unknown variance $\sigma^2$. Without loss of generality, we assume that the predictors and response are standardized:

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n, \quad j = 1, \ldots, p. \tag{5.2}$$

The elastic net (Zou and Hastie, 2005) for linear regression models is given by

$$\hat{\boldsymbol{\beta}} := (1 + \lambda_2) \operatorname*{argmin}_{\boldsymbol{\beta}} \left[ \frac{1}{2n} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \frac{\lambda_2}{2} \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right], \tag{5.3}$$

where $\lambda_1$ ($> 0$) and $\lambda_2$ ($> 0$) are the regularization parameters which control the $L_1$ and $L_2$ penalty.

We can transform the expression (5.3) as follows:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \operatorname*{argmin}_{\boldsymbol{\beta}} \left[ \frac{1}{2n} \|\boldsymbol{y} - c_{\lambda_2} X\boldsymbol{\beta}\|^2 + \frac{\lambda_2 c_{\lambda_2}^2}{2} \sum_{j=1}^{p} \beta_j^2 + \lambda_1 c_{\lambda_2} \sum_{j=1}^{p} |\beta_j| \right] \\
&= \operatorname*{argmax}_{\boldsymbol{\beta}} \left[ \exp\left( -\frac{1}{2\sigma^2} \|\boldsymbol{y} - c_{\lambda_2} X\boldsymbol{\beta}\|^2 \right) \right. \\
&\qquad \left. \cdot \exp\left( -\frac{n\lambda_2 c_{\lambda_2}^2}{2\sigma^2} \sum_{j=1}^{p} \beta_j^2 \right) \cdot \exp\left( -\frac{n\lambda_1 c_{\lambda_2}}{\sigma^2} \sum_{j=1}^{p} |\beta_j| \right) \right] \\
&= \operatorname*{argmax}_{\boldsymbol{\beta}} \left[ N_n(\boldsymbol{y} | c_{\lambda_2} X\boldsymbol{\beta}, \sigma^2 I_n) \right. \\
&\qquad \left. \cdot N_p\left( \boldsymbol{\beta} | \mathbf{0}_p, \frac{\sigma^2}{n\lambda_2 c_{\lambda_2}^2} I_p \right) \cdot \prod_{j=1}^{p} \frac{n\lambda_1 c_{\lambda_2}}{2\sigma^2} \exp\left( -\frac{n\lambda_1 c_{\lambda_2}}{\sigma^2} |\beta_j| \right) \right],
\end{aligned} \tag{5.4}$$

where $c_{\lambda_2} = 1/(1+\lambda_2)$. Laplace distribution $(\lambda/2)\exp(-\lambda x)$ has scale mixture normal representation with an exponential mixing density,

$$\frac{\lambda}{2}\exp(-\lambda x) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau}}\exp\left(-\frac{1}{2\tau}x^2\right)\cdot \mathrm{Exp}(\tau|\lambda^2), \tag{5.5}$$

where $\mathrm{Exp}(\tau|\lambda^2) = (\lambda^2/2)\exp(-\lambda^2\tau/2)$ is a exponential density function with rate parameter $\lambda^2$. We exploit this in (5.4),

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmax}} \left[ N_n(\boldsymbol{y}|c_{\lambda_2}X\boldsymbol{\beta}, \sigma^2 I_n)\cdot N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^2}{n\lambda_2 c_{\lambda_2}^2} I_p\right) \right.$$
$$\left. \cdot \int\cdots\int N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^4}{n^2\lambda_1^2 c_{\lambda_2}^2} D\right)\cdot \prod_{j=1}^{p} \mathrm{Exp}\left(\tau_j^2|\frac{1}{2}\right) d\tau_1^2\cdots d\tau_p^2 \right], \tag{5.6}$$

where $D = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$. In (5.6), since $\boldsymbol{\beta}$ has two normal priors, we set these priors to one.

$$N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^2}{n\lambda_2 c_{\lambda_2}^2} I_p\right)\cdot N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^4}{n^2\lambda_1^2 c_{\lambda_2}^2} D\right) \propto N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^2}{c_{\lambda_2}^2} A_1^{-1}\right), \tag{5.7}$$

where $A_1 = n\lambda_2 I_p + (n^2\lambda_1^2/\sigma^2)D^{-1}$.

Thus, the Bayes model of the elastic net of (5.3) is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ \int \cdots \int N_n(\boldsymbol{y}|c_{\lambda_2} X\boldsymbol{\beta}, \sigma^2 I_n) \cdot N_p\left(\boldsymbol{\beta}|\mathbf{0}_p, \frac{\sigma^2}{c_{\lambda_2}^2} A_1^{-1}\right) \right. \\ &\qquad \left. \cdot \prod_{j=1}^{p} \operatorname{Exp}\left(\tau_j^2|\frac{1}{2}\right) d\tau_1^2 \cdots d\tau_p^2 \right] \\ &= \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[ \int \cdots \int (2\pi)^{-(n+p)/2} (\sigma^2)^{-(n+p)/2} |c_{\lambda_2}^2 A_1|^{1/2} \right. \\ &\qquad \cdot \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{\beta} - c_{\lambda_2} A_2^{-1} X^T \boldsymbol{y})^T A_2 (\boldsymbol{\beta} - c_{\lambda_2} A_2^{-1} X^T \boldsymbol{y}) \right\} \\ &\qquad \cdot \exp\left\{ -\frac{1}{2\sigma^2} \boldsymbol{y}^T (I_n - c_{\lambda_2}^2 X A_2^{-1} X^T) \boldsymbol{y} \right\} \\ &\qquad \left. \cdot \prod_{j=1}^{p} \operatorname{Exp}\left(\tau_j^2|\frac{1}{2}\right) d\tau_1^2 \cdots d\tau_p^2 \right], \end{aligned} \tag{5.8}$$

where $A_2 = c_{\lambda_2}^2 (A_1 + X^T X)$.

## 5.2 Bayesian information criteria

### 5.2.1 BIC and GBIC

The BIC proposed by Schwarz (1978) is a traditional model selection criterion and it is known that the BIC is one of the effective criteria. The BIC is motivated in the Bayesian approach, and it selects a model from a set of candidate models by maximizing the posterior probability.

Suppose we have a set of candidate models $M_1, \ldots, M_\ell$, and each model is characterized by the unknown model parameter $\boldsymbol{\theta}_k$ $(k = 1, \ldots, \ell)$ and the probability density function $f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)$ ($\boldsymbol{y}$ is an $n$-dimensional response vector). Let $\pi_k(\boldsymbol{\theta}_k|\lambda_k)$ be the prior distribution for $\boldsymbol{\theta}_k$ under $M_k$, and $\lambda_k$ is a hyperparameter correspond-

ing $M_k$. Then, the posterior probability of the model $M_k$ is defined by

$$p_{\text{post}}(M_k|\boldsymbol{y}) = \frac{\pi(M_k)\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)d\boldsymbol{\theta}_k}{\sum_{k=1}^{\ell}\pi(M_k)\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)d\boldsymbol{\theta}_k}, \tag{5.9}$$

where $\pi(M_k)$ is the prior distribution for the model $M_k$. The model with the largest posterior probability is equivalent to the model that maximizes

$$\pi(M_k)\int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)d\boldsymbol{\theta}_k = \pi(M_k)\cdot \text{ML}(\lambda_k|\boldsymbol{y}), \tag{5.10}$$

where $\text{ML}(\lambda_k|\boldsymbol{y})$ is the marginal likelihood of the model $M_k$. Typically, it is assumed that the priors over models are uniform, so that $\pi(M_k)$ is constant. The BIC select the model from a set of candidate models with the largest marginal likelihood.

To derive the marginal likelihood, we need to evaluate the following integral:

$$\text{ML}(\lambda_k|\boldsymbol{y}) = \int f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)d\boldsymbol{\theta}_k. \tag{5.11}$$

For this problem, the following Laplace approximation (Tierney and Kadane, 1986) is applicable.

We consider the following integral;

$$\int \exp\{nq(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}, \tag{5.12}$$

where $\boldsymbol{\theta}$ is a $p$-dimensional vector, and $q(\boldsymbol{\theta})$ is a twice differentiable function. Then, this integral is approximated as

$$\int \exp\{nq(\boldsymbol{\theta})\}\, d\boldsymbol{\theta} \approx \exp\left\{nq(\hat{\boldsymbol{\theta}})\right\}\int \exp\left\{-\frac{n}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T J(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right\}, \tag{5.13}$$

where $\hat{\boldsymbol{\theta}}$ is a mode of $q(\boldsymbol{\theta})$, and

$$J(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}q(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{5.14}$$

This approximation is based on the Taylor expansion of $q(\boldsymbol{\theta})$ around its mode. The

first derivative term of the Taylor expansion becomes zero because $\partial q(\hat{\boldsymbol{\theta}})/\partial \boldsymbol{\theta} = 0$. Further, the integral of the right-hand side of (5.14) is known as the multivariate Gaussian integral, thus we obtain the following Laplace approximation:

$$\int \exp\{nq(\boldsymbol{\theta})\}\, d\boldsymbol{\theta} \approx \exp\left\{nq(\hat{\boldsymbol{\theta}})\right\} (2\pi)^{p/2} n^{-p/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2}. \tag{5.15}$$

Schwarz (1978) used the Laplace approximation for the approximate the marginal likelihood around the MLE of $\boldsymbol{\theta}_k$, and derived the following model selection criterion:

$$\begin{aligned} \mathrm{BIC}(\hat{\boldsymbol{\theta}}_k) &= -2 \log f_k(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_k) + p \log n \\ &\approx -2 \log \mathrm{ML}(\lambda_k|\boldsymbol{y}), \end{aligned} \tag{5.16}$$

where $\hat{\boldsymbol{\theta}}_k$ is a MLE of $\boldsymbol{\theta}_k$, all component of $\hat{\boldsymbol{\theta}}_k$ are not zero, and the dimension of $\boldsymbol{\theta}_k$ is $p$. Note that terms with order less than $O(1)$ with respect to $n$ are ignored in the elicitation process of the BIC.

The BIC requires that the parameter must be estimated by MLEs procedure, that is, the BIC is not applicable in models estimated by any regularization procedures including the $L_1$ regularizations. To overcome this drawback, Konishi *et al.* (2004) proposed the GBIC for regularization procedures. Suppose that $\hat{\boldsymbol{\theta}}_k$ is a mode of $f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)$. Then, the GBIC is given by

$$-p \log 2\pi + p \log n + \log |J(\hat{\boldsymbol{\theta}}_k)| - 2nq(\hat{\boldsymbol{\theta}}_k), \tag{5.17}$$

where

$$\begin{aligned} q(\boldsymbol{\theta}_k) &= \frac{1}{n} \log\left\{f_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k(\boldsymbol{\theta}_k|\lambda_k)\right\}, \\ J(\hat{\boldsymbol{\theta}}_k) &= -\left.\frac{\partial^2 q(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^T}\right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}. \end{aligned} \tag{5.18}$$

GBIC is widely applicable because it assumes that $\hat{\boldsymbol{\theta}_k}$ is the posterior mode, and a number of regularized estimate can be seen as it.

### 5.2.2 Bayesian information criterion for the elastic net using the Monte Carlo integration

In Bayes model of elastic net, it is, however, hard to apply the GBIC because integrand of the marginal likelihood is not differentiable at $\beta_j = 0$. Thus, it is difficult to directly apply the Laplace approximation in this case. Further, accuracy of approximation the Laplace approximation depends on the sample size $n$. To overcome these drawback, we propose the methods which approximate the marginal likelihood using the Monte Carlo integration.

When $n > p$, using the Monte Carlo integration, we can approximate the right-hand side of (5.8) by the following:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M}(2\pi)^{-(n+p)/2}(\sigma^2)^{-(n+p)/2}|c_{\lambda_2}^2 A_{1(m)}|^{1/2} \\ \cdot \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-c_{\lambda_2}A_{2(m)}^{-1}X^T\boldsymbol{y})^T A_{2(m)}(\boldsymbol{\beta}-c_{\lambda_2}A_{2(m)}^{-1}X^T\boldsymbol{y})\right\} \\ \cdot \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}^T(I_n - c_{\lambda_2}^2 X A_{2(m)}^{-1}X^T)\boldsymbol{y}\right\},\end{aligned} \tag{5.19}$$

where $A_{1(m)} = n\lambda_2 I_p + (n^2\lambda_1^2/\sigma^2)D_{(m)}^{-1}$, $D_{(m)} = \mathrm{diag}(\tau_{1(m)}^2,\ldots,\tau_{p(m)}^2)$, $\{\tau_{1(m)}^2,\ldots,\tau_{p(m)}^2|\ m = 1,\ldots,M\}$ is a set of random samples from a exponential distribution $\prod_{j=1}^{p}\mathrm{Exp}(\tau_j^2|1/2) = \prod_{j=1}^{p}(1/2)\cdot\exp(-\tau_j^2/2)$, and $A_{2(m)} = c_{\lambda_2}^2(A_{1(m)} + X^TX)$.

Hence, we have the approximated marginal likelihood by integrating (5.8) over $\boldsymbol{\beta}$

$$\begin{aligned}\mathrm{ML} = \frac{1}{M}(2\pi)^{-n/2}(\sigma^2)^{-n/2}\sum_{m=1}^{M}|c_{\lambda_2}^2 A_{1(m)}|^{1/2}|A_{2(m)}|^{-1/2} \\ \cdot \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}^T(I_n - c_{\lambda_2}^2 X A_{2(m)}^{-1}X^T)\boldsymbol{y}\right\}.\end{aligned} \tag{5.20}$$

Similar to the GBIC, we derive a model selection criterion by taking the logarithm of (5.20) and multiplying $-2$

$$-2\log \mathrm{ML} = 2\log M + n\log 2\pi + n\log\sigma^2 - 2\log\sum_{m=1}^{M} |c_{\lambda_2}^2 A_{1(m)}|^{1/2}|A_{2(m)}|^{-1/2}B, \tag{5.21}$$

where $B = \exp\{-\boldsymbol{y}^T(I_n - c_{\lambda_2}^2 X A_{2(m)}^{-1} X^T)\boldsymbol{y}/(2\sigma^2)\}$.

# Chapter 6

# Numerical results

## 6.1 Numerical results for aPIC criterion

### 6.1.1 Monte Carlo simulations

Monte Carlo simulations were conducted to investigate the efficiency of the proposed modeling procedure based on the aPIC criterion for the Bayesian lasso described in Chapter 4. We generated data according to the linear regression model

$$y = \boldsymbol{x}^T \boldsymbol{\beta}^* + \varepsilon, \tag{6.1}$$

where $\boldsymbol{\beta}^*$ is a $p$-dimensional true coefficient vector, $\varepsilon \sim N(0, \sigma^2)$, and $\boldsymbol{x}$ was generated from a multivariate normal distribution with mean vector $\mathbf{0}_p$ and covariance matrix $\Sigma$. The structure of the covariance matrix is given below. In this simulation, we considered four cases inspired by Tibshirani (1996) as follows:

- Case 1: In this case we simulated 200 data sets with 20, 50, or 100 observations. Here, we set $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\sigma = 3$, and the correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ was $0.5^{|i-j|}$.
- Case 2: The second case is the same as Case 1, but with $\boldsymbol{\beta}^* = 0.85 \cdot \mathbf{1}_8$.
- Case 3: The third case is the same as Case 1, but with $\boldsymbol{\beta}^* = (0.5, \mathbf{0}_7^T)^T$, and $\sigma = 2$.

- Case 4: In this case we simulated 200 data sets with 50, 100, or 200 observations. Here, we set $\boldsymbol{\beta}^* = (\mathbf{0}_5^T, \mathbf{2}_5^T, \mathbf{0}_5^T, \mathbf{0.5}_5^T)^T$, and $\sigma = 5$.

In all cases, 2,000 samples from the MCMC simulation were used for estimating the parameters, where the first 1,000 samples were discarded as burn-in. In addition, we confirmed the convergence of the Markov chain simulations by using R.hat (Gelman and Rubin, 1992); the values were close to one. The hyperparameter $\lambda$ was tested for 200 values; $\lambda_i = \lambda_{\min} \cdot \exp[(\log \lambda_{\max} - \log \lambda_{\min}) \cdot (i/200)]$ $(i = 1, \ldots, 200)$, where $\lambda_{\max}$ is such that all coefficient parameters are zero and $\lambda_{\min}$ is $10^{-4}$ when $n = 20$ and $10^{-4}/n$ when $n$ is larger than 50.

The performances of our proposed procedure were evaluated in terms of three accuracies; variable selection, estimation, and prediction accuracies. As the variable selection accuracy, we employed the true positive rate (TPR), true negative rate (TNR), and true sign rate (TSR), respectively, defined by

$$\begin{aligned}
\mathrm{TPR} &= \frac{1}{200}\sum_{k=1}^{200} \frac{\left|\left\{j : \hat{\beta}_j^{(k)} \neq 0 \ \wedge \ \beta_j^* \neq 0\right\}\right|}{\left|\left\{j : \beta_j^* \neq 0\right\}\right|}, \\
\mathrm{TNR} &= \frac{1}{200}\sum_{k=1}^{200} \frac{\left|\left\{j : \hat{\beta}_j^{(k)} = 0 \ \wedge \ \beta_j^* = 0\right\}\right|}{\left|\left\{j : \beta_j^* = 0\right\}\right|}, \\
\mathrm{TSR} &= \frac{1}{200}\sum_{k=1}^{200} \frac{\left|\left\{j : \mathrm{sign}(\hat{\beta}_j^{(k)}) = \mathrm{sign}(\beta_j^*)\right\}\right|}{p},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p^{(k)})^T$ is the estimated coefficient vector for the $k$-th data set, and $|\{*\}|$ is the number of elements included in a set $\{*\}$. The estimation and prediction accuracies are determined by MSE and PSE as follows;

$$\begin{aligned}
\mathrm{MSE} &= \frac{1}{200}\sum_{k=1}^{200} (\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)^T \Sigma (\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*), \\
\mathrm{PSE} &= \frac{1}{200}\sum_{k=1}^{200} \left\{ \frac{1}{n} \|\hat{\boldsymbol{y}}^{(k)} - \tilde{\boldsymbol{y}}^{(k)}\|^2 \right\},
\end{aligned}$$

where $\hat{\boldsymbol{y}}^{(k)} = \boldsymbol{x}^{(k)T}\hat{\boldsymbol{\beta}}^{(k)}$, $\boldsymbol{x}^{(k)}$ is the predictor for the $k$-th data set, and $\tilde{\boldsymbol{y}}^{(k)}$ is a future observation generated from the true model (6.1).

Table 6.1 The results for Case 1 and Case 2.

| | Case 1 $n = 20$ | | | | | Case 2 $n = 20$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.37 | 4.79 | 11.88 | 1.00 | — | 0.93 | 3.42 | 11.16 |
| aPIC+SA | 0.81 | 0.62 | 0.69 | 5.64 | 12.56 | 0.51 | — | 0.49 | 5.83 | 13.04 |
| DIC | 1.00 | 0.00 | 0.36 | 5.31 | 11.94 | 1.00 | — | 0.88 | 4.38 | 11.38 |
| DIC+SA | 0.90 | 0.43 | 0.60 | 5.61 | 12.17 | 0.70 | — | 0.64 | 5.32 | 12.05 |
| Blasso | 1.00 | 0.00 | 0.37 | 5.06 | 12.16 | 1.00 | — | 0.91 | 3.80 | 11.46 |
| Blasso+SA | 0.65 | 0.73 | 0.70 | 8.10 | 14.49 | 0.45 | — | 0.44 | 6.64 | 13.74 |
| WAIC | 1.00 | 0.00 | 0.37 | 4.60 | 11.46 | 1.00 | — | 0.88 | 4.04 | 11.20 |
| WAIC+SA | 0.96 | 0.30 | 0.55 | 4.61 | 11.45 | 0.78 | — | 0.82 | 4.43 | 11.46 |
| Lasso | 0.90 | 0.57 | 0.70 | 4.33 | 11.61 | 0.72 | — | 0.70 | 4.33 | 11.61 |

| | $n = 50$ | | | | | $n = 50$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.38 | 1.39 | 9.95 | 1.00 | — | 0.99 | 1.43 | 10.42 |
| aPIC+SA | 0.99 | 0.58 | 0.73 | 1.38 | 9.96 | 0.81 | — | 0.81 | 2.03 | 10.91 |
| DIC | 1.00 | 0.00 | 0.38 | 1.56 | 10.04 | 1.00 | — | 0.97 | 1.47 | 10.44 |
| DIC+SA | 0.99 | 0.42 | 0.63 | 1.57 | 10.06 | 0.90 | — | 0.89 | 1.59 | 10.55 |
| Blasso | 1.00 | 0.00 | 0.38 | 1.42 | 9.98 | 1.00 | — | 0.98 | 1.36 | 10.34 |
| Blasso+SA | 0.98 | 0.51 | 0.69 | 1.65 | 10.16 | 0.86 | — | 0.86 | 1.72 | 10.63 |
| WAIC | 1.00 | 0.00 | 0.38 | 1.43 | 9.99 | 1.00 | — | 0.97 | 1.41 | 10.00 |
| WAIC+SA | 1.00 | 0.42 | 0.64 | 1.42 | 10.00 | 0.88 | — | 0.93 | 1.56 | 10.11 |
| Lasso | 1.00 | 0.56 | 0.72 | 1.34 | 9.94 | 0.91 | — | 0.90 | 1.71 | 10.60 |

| | $n = 100$ | | | | | $n = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.38 | 0.63 | 9.71 | 1.00 | — | 1.00 | 0.85 | 9.76 |
| aPIC+SA | 1.00 | 0.50 | 0.69 | 0.61 | 9.70 | 0.96 | — | 0.96 | 0.94 | 9.84 |
| DIC | 1.00 | 0.00 | 0.38 | 0.66 | 9.74 | 1.00 | — | 0.99 | 0.79 | 9.69 |
| DIC+SA | 1.00 | 0.47 | 0.67 | 0.64 | 9.73 | 0.98 | — | 0.98 | 0.80 | 9.70 |
| Blasso | 1.00 | 0.00 | 0.38 | 0.65 | 9.73 | 1.00 | — | 1.00 | 0.76 | 9.67 |
| Blasso+SA | 1.00 | 0.41 | 0.63 | 0.73 | 9.81 | 0.98 | — | 0.97 | 0.79 | 9.69 |
| WAIC | 1.00 | 0.00 | 0.38 | 0.66 | 9.46 | 1.00 | — | 0.99 | 0.76 | 9.54 |
| WAIC+SA | 1.00 | 0.50 | 0.69 | 0.65 | 9.45 | 0.97 | — | 0.98 | 0.79 | 9.56 |
| Lasso | 1.00 | 0.56 | 0.73 | 0.62 | 9.70 | 0.98 | — | 0.98 | 0.82 | 9.71 |

For each case, we compared nine procedures; aPIC (proposed procedure), aPIC + SA (aPIC with the sparse algorithm proposed by Hoshina (2012)), DIC, DIC + SA, Blasso (fully Bayesian procedure for the Bayesian lasso proposed by Park and Casella (2008)), Blasso + SA, WAIC, WAIC + SA, and Lasso. Except for Lasso, the parameters were estimated by using the posterior means, and the values of the hyperparameters $\nu_0$ and $\eta_0$ involved in the prior distribution on $\sigma^2$ were set

Table 6.2 The results for Case 3 and Case 4.

| | Case 3 $n = 20$ | | | | | Case 4 $n = 50$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.10 | 0.51 | 4.31 | 1.00 | 0.00 | 0.45 | 7.44 | 30.93 |
| aPIC+SA | 0.12 | 0.90 | 0.80 | 0.46 | 4.32 | 0.60 | 0.82 | 0.71 | 8.86 | 32.20 |
| DIC | 1.00 | 0.00 | 0.10 | 1.17 | 4.63 | 1.00 | 0.00 | 0.43 | 9.41 | 31.79 |
| DIC+SA | 0.37 | 0.71 | 0.66 | 1.12 | 4.61 | 0.73 | 0.57 | 0.63 | 9.94 | 32.35 |
| Blasso | 1.00 | 0.00 | 0.10 | 0.30 | 4.23 | 1.00 | 0.00 | 0.44 | 8.02 | 31.41 |
| Blasso+SA | 0.02 | 0.99 | 0.87 | 0.26 | 4.20 | 0.61 | 0.75 | 0.67 | 10.51 | 33.78 |
| WAIC | 1.00 | 0.00 | 0.11 | 1.38 | 4.73 | 1.00 | 0.00 | 0.46 | 4.29 | 28.25 |
| WAIC+SA | 0.58 | 0.55 | 0.57 | 1.32 | 4.68 | 0.84 | 0.49 | 0.69 | 4.28 | 28.29 |
| Lasso | 0.77 | 0.39 | 0.43 | 1.14 | 4.70 | 0.71 | 0.65 | 0.67 | 7.94 | 31.05 |

| | $n = 50$ | | | | | $n = 100$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.11 | 0.18 | 4.06 | 1.00 | 0.00 | 0.47 | 3.61 | 28.02 |
| aPIC+SA | 0.23 | 0.95 | 0.86 | 0.22 | 4.10 | 0.74 | 0.78 | 0.76 | 3.77 | 28.11 |
| DIC | 1.00 | 0.00 | 0.11 | 0.27 | 4.12 | 1.00 | 0.00 | 0.46 | 3.98 | 28.31 |
| DIC+SA | 0.41 | 0.83 | 0.78 | 0.28 | 4.13 | 0.80 | 0.60 | 0.69 | 4.01 | 28.33 |
| Blasso | 1.00 | 0.00 | 0.11 | 0.21 | 4.09 | 1.00 | 0.00 | 0.46 | 5.21 | 29.46 |
| Blasso+SA | 0.04 | 0.99 | 0.87 | 0.25 | 4.14 | 0.72 | 0.64 | 0.67 | 7.68 | 31.67 |
| WAIC | 1.00 | 0.00 | 0.12 | 0.39 | 4.27 | 1.00 | 0.00 | 0.48 | 1.94 | 26.62 |
| WAIC+SA | 0.64 | 0.64 | 0.66 | 0.38 | 4.25 | 0.88 | 0.56 | 0.75 | 1.97 | 26.65 |
| Lasso | 0.93 | 0.31 | 0.38 | 0.44 | 4.24 | 0.81 | 0.61 | 0.71 | 3.70 | 28.03 |

| | $n = 100$ | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | TPR | TNR | TSR | MSE | PSE | TPR | TNR | TSR | MSE | PSE |
| aPIC | 1.00 | 0.00 | 0.12 | 0.13 | 4.12 | 1.00 | 0.00 | 0.48 | 1.90 | 26.76 |
| aPIC+SA | 0.35 | 0.95 | 0.88 | 0.18 | 4.18 | 0.85 | 0.71 | 0.78 | 1.94 | 26.84 |
| DIC | 1.00 | 0.00 | 0.12 | 0.15 | 4.13 | 1.00 | 0.00 | 0.48 | 2.01 | 26.82 |
| DIC+SA | 0.54 | 0.86 | 0.82 | 0.18 | 4.16 | 0.87 | 0.59 | 0.73 | 2.02 | 26.88 |
| Blasso | 1.00 | 0.00 | 0.12 | 0.18 | 4.17 | 1.00 | 0.00 | 0.47 | 3.63 | 28.43 |
| Blasso+SA | 0.14 | 0.97 | 0.87 | 0.23 | 4.22 | 0.77 | 0.57 | 0.66 | 8.15 | 33.11 |
| WAIC | 1.00 | 0.00 | 0.12 | 0.20 | 4.12 | 1.00 | 0.00 | 0.50 | 0.77 | 25.70 |
| WAIC+SA | 0.81 | 0.65 | 0.68 | 0.20 | 4.12 | 0.96 | 0.70 | 0.84 | 0.78 | 25.72 |
| Lasso | 0.98 | 0.26 | 0.35 | 0.20 | 4.18 | 0.89 | 0.59 | 0.74 | 1.92 | 26.80 |

to 0.001. The tuning parameter in Lasso was selected by 10-fold cross-validation.

Tables 6.1 and 6.2 summarize the simulation results. We observe that aPIC has smaller MSE and PSE than other methods in Case 2 when $n = 20$ and Case 3 when $n = 50, 100$, while aPIC+SA does larger TNR than other methods in Case 4. DIC or DIC+SA provides slightly smaller TSR than other methods in many cases. While Blasso or Blasso+SA outperforms other methods in Case 3 when $n = 20$

Table 6.3 The numbers of observations and predictors for real datasets.

| | diabetes | Boston housing | Parkinson | communities and crimes |
|---|---|---|---|---|
| # of observations | 442 | 506 | 5875 | 2195 |
| # of predictors | 10 | 13 | 19 | 102 |

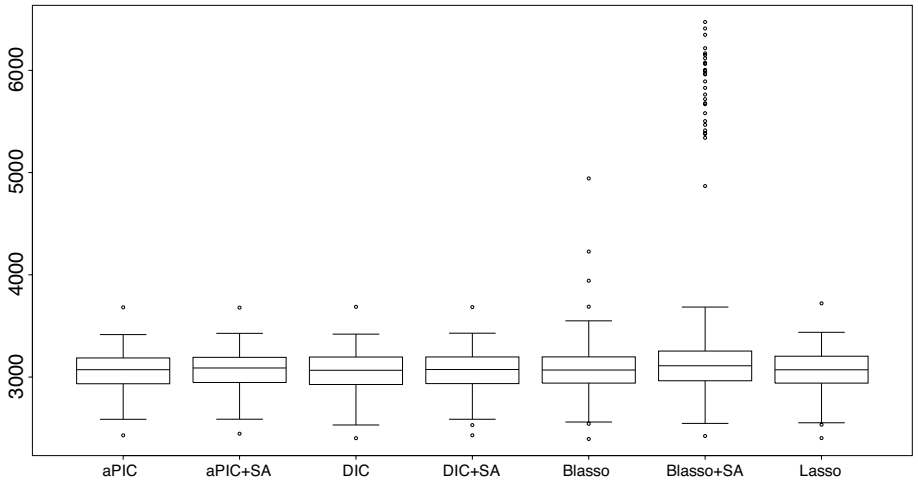
with respect to MSE or PSE, these methods tend to have poor performances in other cases. WAIC and WAIC+SA are better than other methods in terms of MSE and PSE in many cases, but WAIC provides the largest MSE and PSE in Case 3 when $n = 20$. Lasso provides the largest MSE in Case 3 when $n = 50$, although Lasso is competitive with other methods in many cases.

We also compared run-times of the methods; aPIC, DIC, Blasso, and WAIC. Case 1 when $n = 20$ was performed two times, and we averaged the computational times. The computational times of DIC, Blasso, and WAIC were 181.47 times, 0.93 times, and 203.74 times as much as aPIC, respectively. From this result, we observe that the computational time of aPIC is competitive with that of Bolasso, while DIC and WAIC require much computational times compared to aPIC or Bolasso.

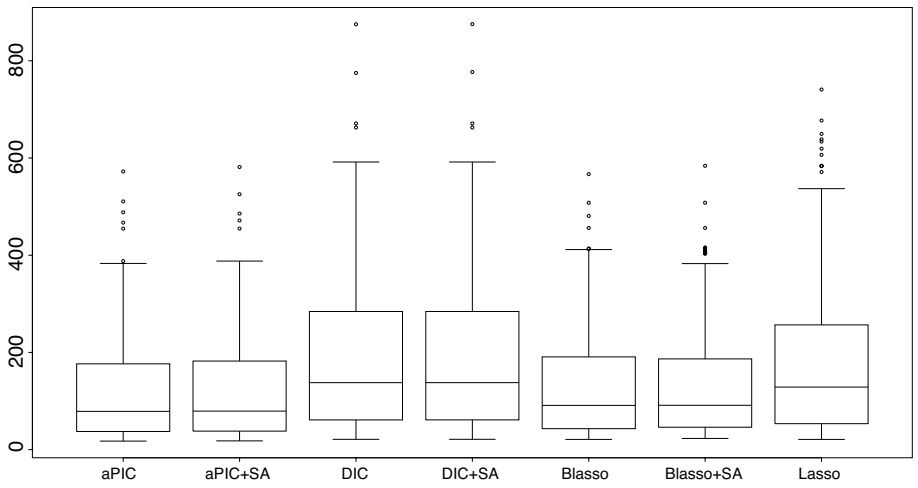
### 6.1.2 Real data examples

By applying our proposed method to real datasets, we examined the effectiveness of our proposed procedure. We used four benchmark datasets; diabetes, Boston housing, Parkinson's disease, and communities and crimes datasets. The diabetes dataset is available from the `lars` package in the software `R` (R Core Team, 2015). Remaining datasets are obtained from UCI database (*http://archive.ics.uci.edu/ml/index.html*). The numbers of observations and predictors for the four datasets are summarized in Table 6.3. Note that we deleted missing values for Parkinson's disease and communities and crimes datasets.

We randomly and equally divided each dataset into training data and test data. Using the training data, we implemented our proposed procedures (aPIC and aPIC+SA), and then computed PSEs by using the test data. We repeated this




(a)

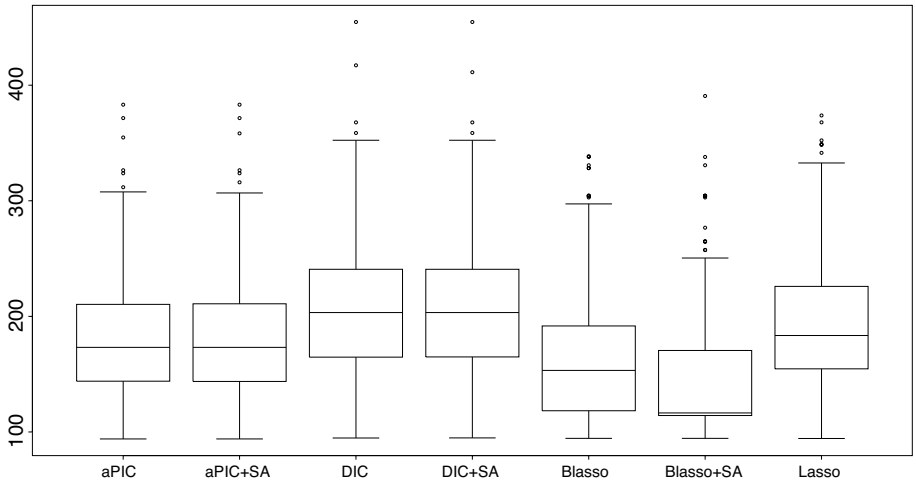



(b)

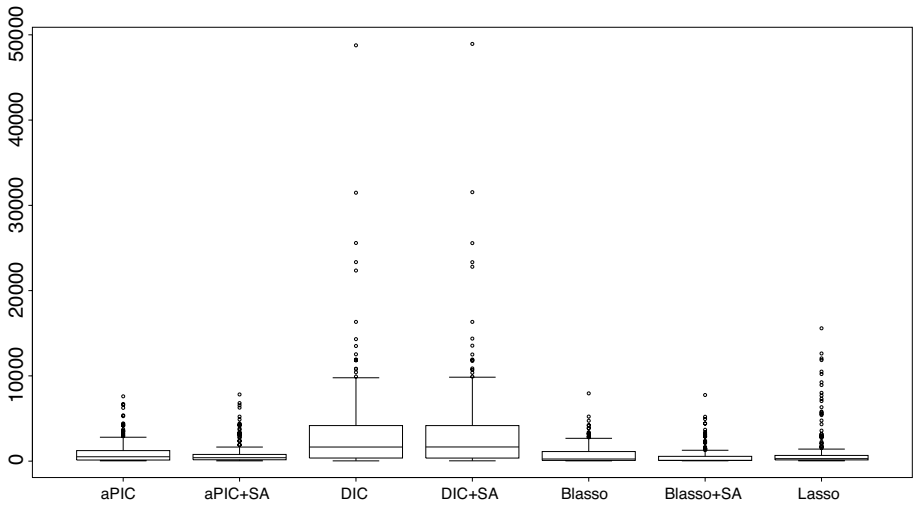
Fig. 6.1 Boxplots of the PSE. (a) shows the result for the diabetes, (b) that for the Boston housing.

procedure 200 times. In addition to our proposed procedures, we implemented DIC, DIC+SA, Blasso, Blasso+SA, and Lasso. WAIC and WAIC were not implemented owing to the computational problem (for details, memory shortage on our PC). For all datasets, we generated 4,000 MCMC samples, and then the first 1,000 samples were discarded as burn-in. We observed that the MCMC simulations converged, since the R.hat ratios were close to one.

Fig. 6.1 and 6.2 show the boxplots of the PSEs. Note that we eliminated




(c)




(d)

Fig. 6.2 Boxplots of the PSE. (c) shows the result for the Parkinson, (d) that for the communities and crimes.

one result for the communities and crimes dataset, since the result was clearly an outlier. From the figures, we observe that Blasso and Blasso+SA are often superior to other methods, although the Blasso has large variances in the diabetes dataset. Meanwhile, our proposed procedures, aPIC and aPIC+SA, produce small median values of PSEs similar to Blasso and Blasso+SA except for the Parkinson dataset, and have variances that are small and relatively stable. We conclude that aPIC and aPIC+SA may be useful in terms of yielding relatively small medians

with small variances.

## 6.2 Numerical results for the MAP Bayesian lasso

In order to examine the effectiveness of our proposed procedure, we conducted Monte Carlo simulations and real data analysis.

### 6.2.1 Simulated performance

Monte Carlo simulations were conducted to investigate the efficacy of our procedure. The data were generated from

$$y = \boldsymbol{x}^T \boldsymbol{\beta}^* + \varepsilon, \tag{6.2}$$

where $\boldsymbol{\beta}^*$ is a $p$-dimensional regression coefficients vector, $\varepsilon \sim N(0, \sigma^2)$, and $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ has the $p$-variate normal distribution with mean $\mathbf{0}_p$. We considered the following cases.

Example 1 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\sigma^2 = 3^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 2 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = 0.85 \cdot \mathbf{1}_p$, $\sigma^2 = 3^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 3 $n = 20$, $p = 8$, $\boldsymbol{\beta}^* = (5, \mathbf{0}_{p-1}^T)^T$, $\sigma^2 = 2^2$. $\mathrm{cor}(x_i, x_j) = \rho^{|i-j|}$, $\rho = 0.5$.

Example 4 $n = 200$, $p = 40$, $\boldsymbol{\beta}^* = (\mathbf{0}_{10}^T, \mathbf{2}_{10}^T, \mathbf{0}_{10}^T, \mathbf{2}_{10}^T)^T$, $\sigma^2 = 15^2$. $\mathrm{cor}(x_i, x_j) = \rho$ $(i \neq j)$, $\rho = 0.5$.

We computed the following four indicators; prediction squared error (PSE), mean squared error of the regression coefficients vector (MSE), false positive rate (FPR), and false negative rate (FNR) to evaluate the prediction and estimation accuracy of outcome model, and the simulation results were obtained by 200 Monte

Carlo trials.

$$\begin{aligned}
\text{PSE} &= \frac{1}{200}\left(\sum_{k=1}^{200} \|\hat{\boldsymbol{y}}^{(k)} - \tilde{\boldsymbol{y}}^{(k)}\|^2/n\right), \\
\text{MSE} &= \frac{1}{200}\left\{\sum_{k=1}^{200} (\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)^T R(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*)\right\}, \\
\text{FPR} &= \frac{1}{200}\left(\sum_{k=1}^{200} \#\{\hat{\beta}_j^{(k)} \neq 0 \ ; \ \beta_j^* = 0\}/\#\{\beta_j^* = 0\}\right), \\
\text{FNR} &= \frac{1}{200}\left(\sum_{k=1}^{200} \#\{\hat{\beta}_j^{(k)} = 0 \ ; \ \beta_j^* \neq 0\}/\#\{\beta_j^* \neq 0\}\right).
\end{aligned} \tag{6.3}$$

Where $\hat{\boldsymbol{y}}^{(k)}$ is a predicted vector of $k$-th data sets, $\tilde{\boldsymbol{y}}^{(k)}$ is a new response vector that independent from $\boldsymbol{y}$, $p \times p$ matrix $R$ is a correlation matrix of $\boldsymbol{x}$, and $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\beta}_1^{(k)}, \ldots, \hat{\beta}_p)^T$ is an estimated regression coefficients vector from $k$-th data set. We set $M$ of (4.25) to 500, shape and rate parameter $\nu_0, \eta_0$ of inverse-gamma prior on $\sigma^2$ are both 0.001, the tuning parameter $\lambda$ is estimated by the hierarchical Bayesian estimation with non-informative gamma prior on $\lambda^2$, and we use MLE for estimates of $\sigma^2$. In all examples, 3000 samples from the Gibbs sampler were used for estimating parameters after 1000 burn in.

We compared the indicators of our procedure with those of the other procedures described in Section 4.3.3 and the 10-fold Cross validation (CV). The full Bayesian approach (Mean) which estimates all parameters by posterior mean is also compared with our procedure. Table 6.4 and 6.5 show the comparison of these sparse regression modeling procedures. The result of AIC is not presented, since Mallows' $C_p$ criterion and AIC yield the same result when $\sigma^2$ is given. The Bayesian estimates derived by three procedures (Mean, DIC, and WAIC) were calculated by the sparse algorithm (Hoshina, 2012), since they have no sparse solution for the estimates of regression coefficients. The error variance $\sigma^2$ was estimated by the MLE in the lasso procedures with $C_p$ and BIC.

The simulation results are summarized as follows:

1. For Examples 1, 3, and 4, the Bayesian procedures except to a DIC have smaller errors than all lasso procedures in terms of PSE and MSE.

2. Our procedure has slightly large FPR in Examples 1, 3, 4, but all examples show that our procedure has smaller FNR. This may denotes that our procedure takes in more variables into the estimated model.
3. In Examples 1, 2, and 3, our procedure has the smallest value in terms of PSE, and has the smallest value in terms of MSE in Examples 1, 3.

From the summary of the Monte Carlo simulation, our procedure has better prediction and estimation accuracy. Moreover, our procedure hardly waste the important variables from model. Thus, we believe that our proposed methodology seems to be useful in terms of variable selection, parameter estimation and prediction. Note that WAIC needs the Gibbs sampling for each candidate value of $\lambda$.

### 6.2.2 Real data analysis

We explore our procedure by using two types of the diabetes datasets of Efron *et al.* (2004) which have been obtained from 442 diabetes patients. First, the proposed procedure was applied to low-dimensional dataset which are constructed ten baseline variable (age, sex, body mass index, average blood pressure and six blood serum measurements) and the response variable which is a quantitative measure of disease progression one year after baseline.

We compare 8 procedures, the proposed procedure (Proposed), posterior mean (Mean), DIC, WAIC, 10-fold Cross validation (CV), Mallows' $C_p$ ($C_p$), BIC, and Generalized Cross-validation (GCV). Table 6.6 reported the estimated standardized regression coefficients for this datasets.

In order to compare the prediction accuracy, the out-of-sample comparison was also conducted. We divided the datasets into 221 training and 221 test data. After the model building in training data, we computed the prediction error for test data. Table 6.6 showed the average prediction errors of 50 trials of this procedure.

Secondly, we studied high-dimensional diabetes dataset which has ten baseline predictor of first example and 54 certain interactions. Table 6.6 also reported the average prediction error of this dataset, and Fig. 6.3 and 6.4 reported the

Table 6.4 Comparison of sparse regression modeling procedures in Example 1 and 2. The values in parenthesis of PSE and MSE are their standard deviations.

Example 1.

| | PSE | | MSE | | FPR | FNR |
|---|---|---|---|---|---|---|
| Proposed | 6.17 | (2.71) | 3.83 | (2.89) | 0.53 | 0.09 |
| Mean | 8.04 | (4.66) | 5.57 | (5.35) | 0.27 | 0.25 |
| DIC | 15.18 | (6.28) | 10.29 | (5.28) | 0.04 | 0.50 |
| WAIC | 6.45 | (3.23) | 4.39 | (3.37) | 0.46 | 0.14 |
| CV | 7.49 | (4.60) | 4.25 | (4.02) | 0.47 | 0.12 |
| $C_p$ | 11.66 | (9.00) | 7.01 | (6.96) | 0.28 | 0.24 |
| BIC | 9.44 | (7.24) | 5.47 | (5.90) | 0.39 | 0.18 |
| GCV | 11.66 | (9.00) | 7.01 | (6.96) | 0.28 | 0.24 |

Example 2.

| | PSE | | MSE | | FPR | FNR |
|---|---|---|---|---|---|---|
| Proposed | 6.33 | (2.90) | 4.22 | (2.28) | – | 0.34 |
| Mean | 8.70 | (5.04) | 6.12 | (4.29) | – | 0.55 |
| DIC | 15.30 | (6.05) | 10.26 | (3.38) | – | 0.80 |
| WAIC | 7.06 | (3.80) | 4.86 | (3.21) | – | 0.45 |
| CV | 6.99 | (4.34) | 4.21 | (2.84) | – | 0.36 |
| $C_p$ | 10.71 | (7.57) | 6.49 | (4.40) | – | 0.50 |
| BIC | 9.48 | (6.97) | 5.76 | (4.16) | – | 0.45 |
| GCV | 10.71 | (7.57) | 6.49 | (4.40) | – | 0.50 |

estimated standardized regression coefficients.

The results of the real data analysis are summarized as follows:

1. In low-dimensional diabetes dataset, the resulting models of the Bayesian procedures except to a DIC have more variables than all lasso procedures. These procedures also have smaller values in terms of the value of the average prediction error.

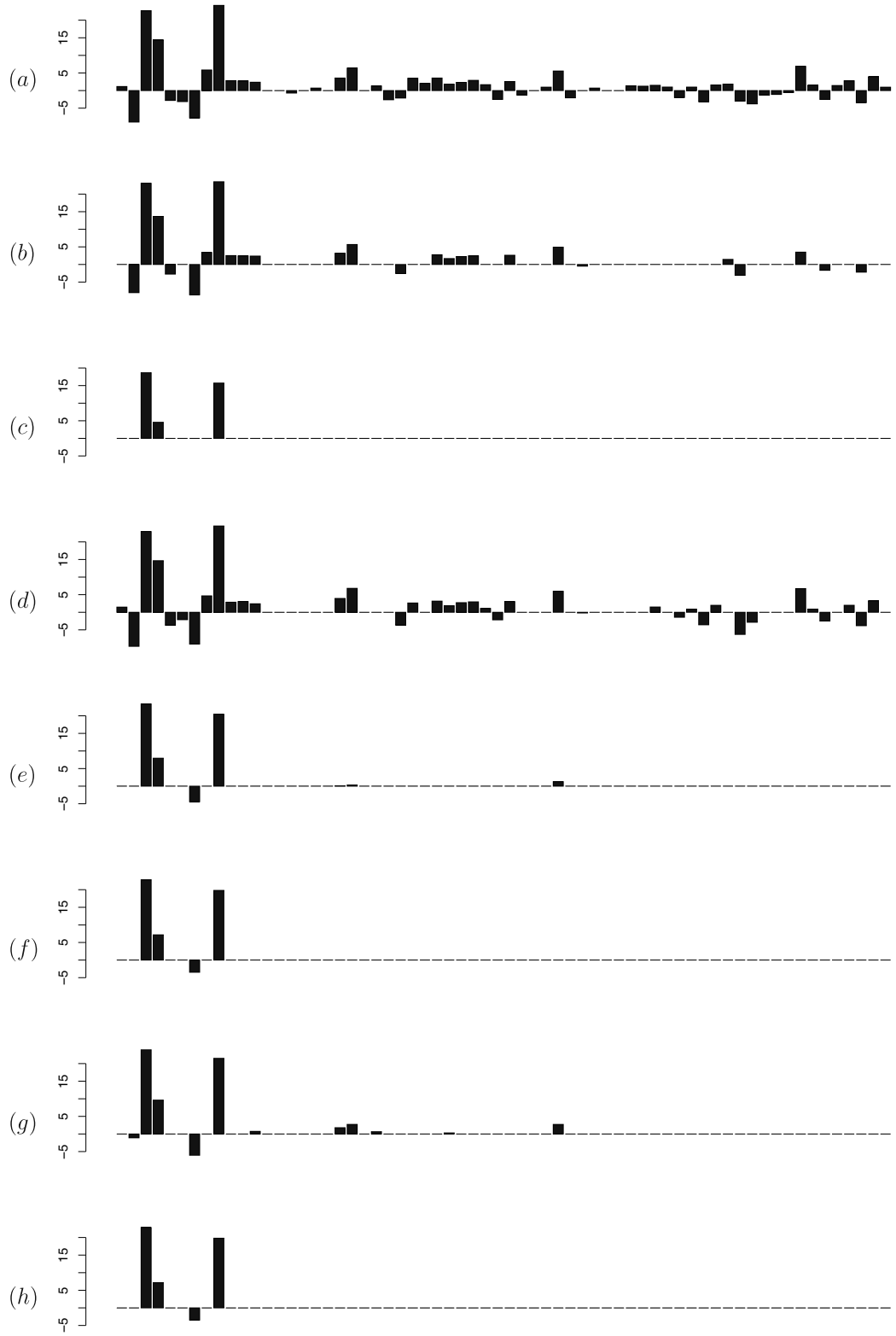



Fig. 6.3 Barplots of the estimated standardized regression coefficients for the high dimensional diabetes dataset: (a) shows the result for the proposed, (b) that for the Mean, (c) that for the DIC, (d) that for the WAIC, (e) that for the CV, (f) that for the $C_p$, (g) that for the BIC, and (h) that for the GCV.

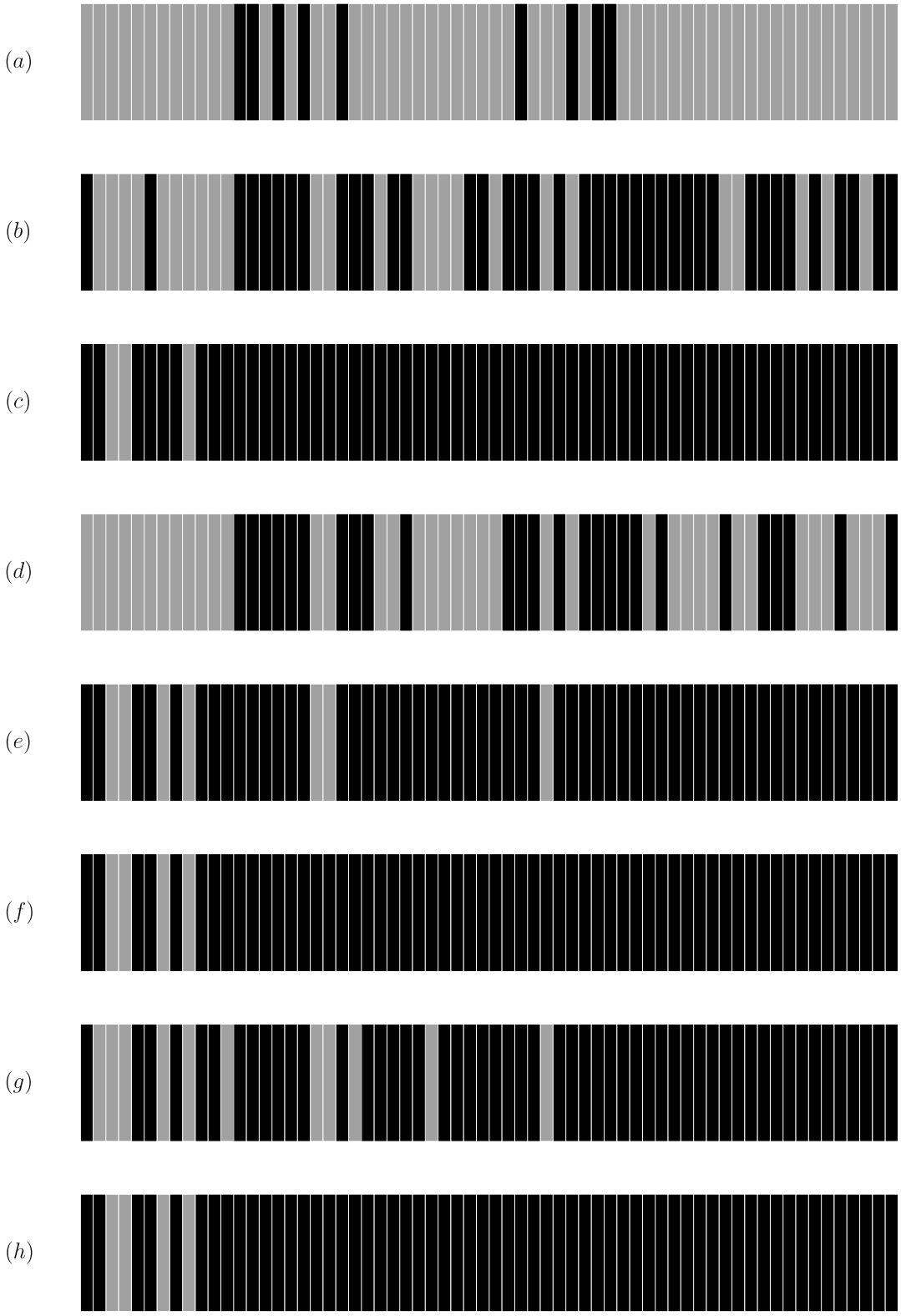



Fig. 6.4 The sparsity of the estimated standardized regression coefficients for the high dimensional diabetes dataset: (a) shows the result for the proposed, (b) that for the Mean, (c) that for the DIC, (d) that for the WAIC, (e) that for the CV, (f) that for the $C_p$, (g) that for the BIC, and (h) that for the GCV. Grey areas correspond to non-zero coefficients, and black areas correspond to zero coefficients.

Table 6.5 Comparison of sparse regression modeling procedures in Example 3 and 4. The values in parenthesis of PSE and MSE are their standard deviations.

Example 3.

| | PSE | | MSE | | FPR | FNR |
|---|---|---|---|---|---|---|
| Proposed | 2.59 | (1.11) | 1.34 | (1.07) | 0.62 | 0.00 |
| Mean | 2.76 | (1.23) | 1.36 | (1.16) | 0.44 | 0.00 |
| DIC | 6.56 | (2.23) | 3.37 | (2.00) | 0.01 | 0.00 |
| WAIC | 2.79 | (1.32) | 1.40 | (1.24) | 0.44 | 0.00 |
| CV | 3.73 | (4.17) | 1.53 | (3.64) | 0.42 | 0.02 |
| $C_p$ | 6.76 | (8.03) | 3.81 | (7.05) | 0.18 | 0.06 |
| BIC | 4.64 | (5.25) | 2.04 | (4.76) | 0.31 | 0.03 |
| GCV | 6.76 | (8.03) | 3.81 | (7.05) | 0.18 | 0.06 |

Example 4.

| | PSE | | MSE | | FPR | FNR |
|---|---|---|---|---|---|---|
| Proposed | 193.70 | (21.85) | 25.08 | (5.76) | 0.49 | 0.14 |
| Mean | 193.67 | (22.01) | 24.66 | (5.83) | 0.42 | 0.15 |
| DIC | 437.80 | (49.79) | 234.72 | (46.73) | 0.36 | 0.13 |
| WAIC | 202.37 | (24.00) | 24.23 | (7.11) | 0.46 | 0.09 |
| CV | 238.87 | (36.03) | 67.19 | (34.35) | 0.28 | 0.26 |
| $C_p$ | 315.58 | (144.96) | 140.80 | (137.76) | 0.23 | 0.34 |
| BIC | 220.94 | (33.12) | 50.27 | (27.49) | 0.31 | 0.23 |
| GCV | 315.58 | (144.96) | 140.80 | (137.76) | 0.23 | 0.34 |

2. In high-dimensional diabetes datasets, the resulting models of the Bayesian procedures except to a DIC have also more variables than all lasso procedures. Our procedure, posterior mean, and BIC have smaller values in terms of the value of the average prediction error though WAIC has larger value.

From the summary of the real data analysis, our procedure has better prediction accuracy.

Table 6.6 The estimated standardized regression coefficients for low-dimensional diabetes dataset. $*$s in table are expressive exactly zero values.

| | Proposed | Mean | DIC | WAIC | CV | $C_p$ | BIC | GCV |
|---|---|---|---|---|---|---|---|---|
| age | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| sex | -10.62 | -10.18 | $*$ | -9.77 | $*$ | $*$ | $*$ | $*$ |
| bmi | 24.94 | 24.96 | 18.47 | 24.88 | 17.44 | 14.65 | 14.65 | 14.65 |
| map | 14.94 | 14.65 | 3.68 | 14.34 | 0.24 | $*$ | $*$ | $*$ |
| tc | -13.07 | -9.80 | $*$ | -7.42 | $*$ | $*$ | $*$ | $*$ |
| ldl | 3.15 | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ | $*$ |
| hdl | -5.63 | -6.73 | $*$ | -7.73 | $*$ | $*$ | $*$ | $*$ |
| tch | 5.39 | 4.85 | $*$ | 4.34 | $*$ | $*$ | $*$ | $*$ |
| ltg | 26.65 | 25.35 | 14.71 | 24.48 | 14.58 | 11.79 | 11.79 | 11.79 |
| glu | 3.17 | 3.07 | $*$ | 2.94 | $*$ | $*$ | $*$ | $*$ |

Table 6.7 The average prediction error of the out-of-sample comparison. The number in parenthesis are the standard deviations.

Low-dimensional diabetes dataset

| Proposed | Mean | DIC | WAIC |
|---|---|---|---|
| 3025.39 | 3024.16 | 3856.53 | 3034.29 |
| (203.31) | (207.78) | (268.22) | (205.80) |

| CV | $C_p$ | BIC | GCV |
|---|---|---|---|
| 4212.28 | 4397.09 | 3430.46 | 4397.09 |
| (993.33) | (1096.13) | (651.82) | (1096.13) |

High-dimensional diabetes dataset

| Proposed | Mean | DIC | WAIC |
|---|---|---|---|
| 3095.19 | 3090.59 | 3933.00 | 3848.07 |
| (197.05) | (190.03) | (302.43) | (1597.29) |

| CV | $C_p$ | BIC | GCV |
|---|---|---|---|
| 3152.15 | 3259.43 | 3046.11 | 3259.43 |
| (259.40) | (386.14) | (184.40) | (386.14) |

# Chapter 7

# Concluding remarks

In the present thesis, we have proposed a number of new regularization procedures. We first proposed an algorithm which corrects the resulting regression coefficients of the Bayesian modeling to be sparse according to the posterior probability. This algorithm enables us to obtain sparse solutions from the Bayesian lasso, and it can be applied for several Bayes-type $L_1$ regularizations to perform simultaneously the parameter estimation and the variable selection.

Secondly, we proposed a new model selection criterion aPIC, for evaluating a Bayesian predictive distribution of the Bayesian lasso, for the selection of appropriate values of hyper-paramters included in a prior distribution. The proposed model selection criterion has been introduced by the approximated prior; the Laplace prior for the regression coefficients are approximated by a normal prior which is the closest distribution in terms of the the Kullback-Leibler information. Monte Carlo experiments showed that the proposed procedure is effective in terms of prediction, estimation, and model selection accuracies.

Further, we have proposed a new modeling procedure, the MAP Bayesian lasso, which derives the MAP estimates of the Bayesian lasso from an approximated posterior density. The posterior approximation is based on the Monte Carlo integration. Numerical examples showed that our procedure performs well in terms of variable selection, parameter estimation, and prediction. The real data analysis also showed the prediction efficiency of our procedure.

A model selection criterion for the elastic net have been proposed. This model

selection criterion evaluates the marginal likelihood. Although GBIC of Konishi *et al.* (2004) also evaluates the marginal likelihood using the Laplace approximation (it depends on the dimensionality and the sample size), our proposed procedure is derived by the Monte Carlo integration. It is expected that the our procedure does not depend on the dimensionality and the sample size, compared to analytical approaches. However, we have known that the estimation of the error variance affects the accuracy of our procedure from an empirical evidence. We leave this topics as future research.

Moreover, we have described some properties of the $L_1$ regularizations. The sparsities of the ridge, the lasso are compared by using the elementary differential geometry. The algorithms which calculate the estimates of the $L_1$ regularizations are introduced and a new algorithm which calculates the degrees of freedom of the LARS are described. The definition of the strength of the sparsity is given. We have introduced the relationships between the $L_1$ regularizations and the Bayes model, and the unimodalities of the Bayesian lasso, the Bayesian elastic net and the Bayesian adaptive lasso with hyper-priors have been shown. Although they are unpublished works, we believe that these results have academic values.

About the MAP Bayesian lasso, future studies will be required to consider the generalized sparse regression procedures such as the elastic net, the adaptive lasso, and the group lasso. The algorithm that calculates the degrees of freedom of the LARS requires further validation to publish its results.

# Bibliography


[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principles. *2nd International Symposium on Information Theory*, 267–281.

[2] Andrews, D.F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society*, Ser. B, 36, 99–102.

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*, New York: Springer.

[4] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24, 2350–2383.

[5] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory, and Applications*, New York: Springer.

[6] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31, 377–403.

[7] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81, 461–470.

[8] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross validation. *Journal of the American Statistical Association*, 99, 619–632.

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

[10] 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.

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

[12] Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 35, 2173–2192.

[13] Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 42, 819–849.

[14] Friedman, J. (2012). Fast sparse regression and classification. *International Journal of Forecasting*, 28, 722–738.

[15] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

[16] Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7, 397–416.

[17] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.

[18] Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96, 835–845.

[19] Hirose, K., Tateishi, S. and Konishi, S. (2013). Tuning parameter selection in sparse regression modeling. *Computational Statistics and Data Analysis*, 59, 28–40.

[20] Horel, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.

[21] Hoshina, I. (2012). Sparse regression modeling via the Bayesian lasso (in Japanese). *Bulletin of the Computational Statistics of Japan*, 25, 73–85.

[22] Hoshina, I. (2015). Sparse regression modeling via the MAP Bayesian lasso. *Bulletin of Informatics and Cybernetics*, 47, 37–58.

[23] Huang, J., Horowitz, J.L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36, 587–613.

[24] Kanba, M. and Naito, K. (2011). Selection of smoothing parameter for one-step sparse estimates with $L_q$ penalty. *Journal of Data Science*, 9, 549–564.

[25] Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, 100, 1338–1352.

[26] Kawano, S. (2014). Selection of tuning parameters in bridge regression models

via Bayesian information criterion. *Statistical Papers*, 55, 1207–1223

[27] Kawano, S., Hoshina, I., Shimamura, K., and Konishi, S. (2015). Predictive model selection criteria for Bayesian lasso regression. *Journal of the Japanese Society of Computational Statistics*, 28, 67–82.

[28] Kim, D., Kawano, S. and Konishi, S. (2012). Predictive information criteria for Bayesian nonlinear regression models. *Bulletin of Informatics and Cybernetics*, 44, 17–28.

[29] Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Communications in Statistics – Theory and Methods*, 26, 2223–2246.

[30] Konishi, S., Ando, T., and Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91, 27–43.

[31] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*, New York: Springer.

[32] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28, 1356–1378.

[33] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

[34] Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5, 369–412.

[35] Mallows, C. (1973). Some comments on $C_p$. *Technometrics*, 15, 661–675.

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

[37] Murphy, K. (2012). *Machine Learning – a Probabilistic Perspective*, MIT Press.

[38] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

[39] Polson, N.G., Scott, J.G. and Windle, J. (2014). The Bayesian bridge. *Journal of the Royal Statistical Society*, Ser. B, 76, 713–733.

[40] Reid, S., Tibshirani, R., and Friedman, J. (2014). A study of error variance estimation in lasso regression. *arXiv*, 1311.5274v2.

[41] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

[42] Sepehri, A. (2016). The Bayesian SLOPE. *arXive preprint*. arXive:1608.08968v2 [stat.ME].

[43] Simon, N., Friedman, J. Hasie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22, 231–245.

[44] Smith, A. F. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society*, Ser. B, 42, 213–220.

[45] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

[46] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

[47] Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics*, 40, 1198–1232.

[48] Tierney, L. and Kadane J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86.

[49] Tseng P. (2001). Convergence of a block coordinate descent method for non-differentiable minimization. *Journal of Optimization Theory and Applications*, 109, 474–494.

[50] Watanabe, S. (2010a). Equations of states in singular statistical estimation. *Neural Networks*. 23, 20–34.

[51] Watanabe, S. (2010b). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*. 11, 3571–3594.

[52] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93, 120–131.

[53] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression

with grouped variables. *Journal of the Royal Statistical Society*, Ser. B, 68, 49–67.

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

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

[56] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

[57] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.

[58] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36, 1509–1533.