SOME PERSPECTIVES ON THE PROBLEM OF MODEL SELECTION

TRAN MINH NGOC

(BSc and MSc, Vietnam National Uni.)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2011

Acknowledgements

I am deeply grateful to my supervisor, David John Nott, for his careful guidance and invaluable support. David has taught me so much about conducting academic research, academic writing and career planning. His confidence in me has encouraged me in building independent research skills. Having him as supervisor is my great fortune. I would also like to express my thanks to my former supervisor, Berwin Turlach - now at University of Western Australia, for his guidance and encouragement at the beginning period of my graduate program.

I would like to thank Marcus Hutter and Chenlei Leng for providing interesting research collaborations. It has been a great pleasure to work with them. Much of my academic research has been inspired and influenced through personal communication with Marcus. I would also like to acknowledge the financial support from NICTA and ANU for my two visits to Canberra which led to our joint works.

I would like to take this opportunity to say thank you to my mother for her endless love. To my late father: thank you for bringing me to science and for your absolute confidence in me. I would like to thank my wife Thu Hien and my daughter Ngoc Nhi for their endless love and understanding, thank my wife for her patience when I spent hours late at night sitting in front of the computer. You have always been my main inspiration for doing maths. I also thank my sisters for supporting me, both spiritually and financially.

Contents

Summary

Model selection in general and variable selection in particular are important parts of data analysis. This thesis makes some contributions to the model selection literature by introducing two general procedures for model selection and two novel algorithms for variable selection in very general frameworks. This thesis is based on a collection of my own works and joint works. Each chapter can be read separately.

After giving in Chapter 1 a brief literature review and motivation for the thesis, I shall discuss in Chapter 2 a general procedure for model selection, called the loss rank principle (LoRP). The main goal of the LoRP is to select a parsimonious model that fits the data well. General speaking, the LoRP consists in the so-called loss rank of a model defined as the number of other (fictitious) data that fit the model better than the actual data, and the model selected is the one with the smallest loss rank. By minimizing the loss rank, the LoRP selects a model by trading off between the empirical fit and the model complexity. LoRP seems to be a promising principle with a lot of potential, leading to a rich field. In this thesis, I have only scratched at the surface of the LoRP, and explored it as much as I can.

While a primary goal of model selection is to understand the underlying structure in the data, another important goal is to make accurate (out-of-sample) predictions on future observations. In Chapter 3, I describe a model selection procedure that has an explicit predictive motivation. The main idea is to select a model that is closest to the

full model in some sense. This results in selection of a parsimonious model with similar predictive performance to the full model. I shall then introduce a predictive variant of the Lasso - called the predictive Lasso. Like the Lasso, the predictive Lasso is a method for simultaneous variable selection and parameter estimation in generalized linear models. Unlike the Lasso, however, our approach has a more explicit predictive motivation, which aims at producing a useful model with high prediction accuracy.

Two novel algorithms for variable selection in very general frameworks are introduced in Chapter 4. The first algorithm, called the Bayesian adaptive Lasso, improves on the original Lasso in the sense that adaptive shinkages are used for different coefficients. The proposed Bayesian formulation offers a very convenient way to account for model uncertainty and for selection of tuning parameters, while overcoming the problems of model selection inconsistency and estimation biasedness in the Lasso. Extensions of the methodology to ordered and grouped variable selection are also discussed in detail. I then present the second algorithm which is for simultaneous fast variable selection and parameter estimation in high-dimensional heteroscedastic regression. The algorithm makes use of a Bayes variational approach which is an attractive alternative to Markov chain Monte Carlo methods in high-dimensional settings, and reduces to well-known matching pursuit algorithms in the homoscedastic case. This methodology has potential for extension to much more complicated frameworks such as simultaneous variable selection and component selection in flexible modeling with Gaussian mixture distributions.

List of Figures

List of Tables

List of Symbols and Abbreviations

AIC: Akaike's information criterion.

BIC: Bayesian information criterion or Schwarz's criterion.

BaLasso: Bayesian adaptive Lasso.

BLasso: Bayesian Lasso.

BMA: Bayesian model averaging.

BMS: Bayesian model selection.

CFR: correctly-fitted rate.

kNN: k nearest neighbors.

KL: Kullback-Leibler divergence.

Lasso: least absolute shrinkage and selection operator.

aLasso: adaptive Lasso.

pLasso: predictive Lasso.

LoRP: loss rank principle.

LR: loss rank.

MCMC: Markov chain Monte Carlo.

MDL: minimum description length.

ML: maximum likelihood.

MLE: maximum likelihood estimator.

MSE: mean squared error.

MUI: model uncertainty indicator.

NZE: number of zero-estimated coefficients.

OLS: ordinary least squares.

OP: optimal predictive model.

PELM: penalized empirical loss minimization.

PML: penalized maximum likelihood.

POPMOS: procedure for optimal predictive model selection.

PPS: partial prediction score.

VAR: variational approximation ranking algorithm.

$\mathcal{X}$: space of input values.

$\mathcal{Y}$: space of output values.

$D = \{(x_1,y_1),...,(x_n,y_n)\}$: observed data.

$\mathcal{D}$: set of all possible data D.

$\mathbf{x} = (x_1,...,x_n)^\top$: vector of x-observations, similarly $\mathbf{y}$.

$\mathbb{R}$: set of real numbers.

$\mathbb{N} = \{1,2,...\}$: set of natural numbers.

$\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

$\mathcal{F}$= ("small") class of functions.

$r : \mathcal{D} \to \mathcal{F}$= regressor/model.

$\mathcal{R}$= ("small") class of regressors/models.

$a \rightsquigarrow b$: a is replaced by b.

Chapter 1

Introduction

Model selection is a fundamental problem in statistics as well as in many other scientific fields such as machine learning and econometrics. According to R. A. Fisher, there are three aspects of a general problem of making inference and prediction: (1) model specification, (2) estimation of model parameters, and (3) estimation of precision. Before the 1970s, most of the published works were centered on the last two aspects where the underlying model was *assumed* to be known. Model selection has attracted significant attention in the statistical community mainly since the seminal work of Akaike [1973]. Since then, a large number of methods have been proposed. In this introductory chapter, we shall first give a brief review of the model selection literature, followed by motivation for, and a brief statement of the main contributions of, this thesis.

1.1 A brief review of the model selection literature

For expository purposes, we shall restrict here the discussion of the model selection problem to the regression and classification framework. Our later discussions are, however, by no

means limited to such a restriction.

Consider a data set $D=\{(x_1,y_1),...,(x_n,y_n)\}$ from a perturbed functional relationship

$$y = f_{\text{true}}(x) + \text{ noise.}$$

Given a family of function classes/models $\{\mathcal{F}_c, c\in\mathcal{C}\}$, we would like to choose the "best" to fit/interpret D and/or to make good predictions on future observations. Here $\mathcal{F}_c$ denotes a class of functions (which will also be referred to as a model) with the index c standing for its *c*omplexity. For example, it can be the class $\mathcal{F}_d$ of d-order polynomials or can be the kNN regression model $\mathcal{F}_k$ with k-nearest neighbors.

Many well-known procedures for model selection can be regarded as penalized versions of the maximum likelihood (ML) principle. One first has to assume a sampling distribution $\mathrm{P}(D|f)$ for D, e.g., the y_i have independent Gaussian distributions $N(f(x_i),\sigma^2)$. For estimation within a model, ML chooses

$$\hat{f}_D^c = \arg\max_{f\in\mathcal{F}_c} \mathrm{P}(D|f),$$

and for choice of model, penalized ML (PML) then chooses

$$\hat{c} = \arg\min_c \{-\log \mathrm{P}(D|\hat{f}_D^c) + \mathrm{pen}(\mathcal{F}_c)\},$$

where the penalty term $\mathrm{pen}(\mathcal{F}_c)$ depends on the used approach. For instance, $\mathrm{pen}(\mathcal{F}_c)$ might be $\frac{1}{2}k$ as in AIC [Akaike, 1973], or $\frac{\log n}{2}k$ as in BIC [Schwarz, 1978] where k is the number of free parameters in the model. From a practical point of view, AIC and BIC, especially AIC, are probably the most commonly used approaches to model selection. They are very easy to use and work satisfactorily in many cases. Some extension versions of AIC have also been proposed in the literature (see, e.g. Burnham and Anderson [2002]). All PML variants rely heavily on a proper sampling distribution (which may be difficult

to establish), ignore (or at least do not tell how to incorporate) a potentially given loss function, are based on distribution-free penalties (which may result in a bad performance for some specific distributions), and are typically limited to (semi)parametric models.

Related are penalized empirical loss minimization (PELM) methods (also known as structural risk minimization) originally introduced by Vapnik and Chervonenkis [1971]. Consider a bounded loss function $l(.,.)$, empirical loss $L_n(f)=\frac{1}{n}\sum_1^n l(f(x_i),y_i)$ and "true" loss $L(f)=\mathbf{E}l(f(X),Y)$. Let $\hat{f}_D^c=\text{argmin}_{f\in\mathcal{F}_c}L_n(f)$. Then PELM chooses

$$\hat{c}=\arg\min_c\{L_n(\hat{f}_D^c)+\text{pen}(\mathcal{F}_c)\}.$$

Unlike PML, the optimality properties of PELM are often studied in terms of *nonasymptotic theory*, in which concentration inequalities are used to obtain the so-called *oracle inequalities* which evaluate how close the estimator is to the optimal one (see Massart [2007] and Section 2.4 for a detailed review). The major question is what penalty function should be used. Koltchinskii [2001] and Bartlett et al. [2002] studied PELM based on Rademacher complexities which are estimates of $\mathbf{E}\text{sup}_{f\in\mathcal{F}_c}|L(f)-L_n(f)|$ which can be considered as an effective estimate of the complexity of $\mathcal{F}_c$. These methods have a solid mathematical basis and in particular their penalty terms are data-dependent, so one can expect better performance over model selection procedures based on distribution-free penalties. A main drawback is that they are intractable because they often involve unknown parameters that need to be estimated. Furthermore, from a practical point of view, PELM criteria are not easy to use.

The third class of model selection procedures are Bayesian model selection (BMS) methods which are very efficient and increasingly used. Typically, BMS consists in building a hierarchical Bayes formulation and using MCMC methods or some other computational algorithm to estimate posterior model probabilities. The model with the highest posterior

model probability will be selected; alternatively, inferences can be averaged over some models with highest posterior model probabilities. See O'Hagan and Forster [2004], George and McCulloch [1993], Smith and Kohn [1996] and Hoeting et al. [1999] for comprehensive introductions to BMS. BMS with MCMC methods may be computationally demanding in high-dimensional problems. A representative is the popular BIC of Schwarz [1978] which is an approximation of the minus logarithm of posterior model probability $-\log P(\mathcal{F}_c|D)$ (with a uniform prior on models). BIC possesses an optimality in terms of identification, i.e., it is able to identify the true model as $n \to \infty$ if the model collection contains the true one (see, e.g., Chambaz [2006]). However, BIC is not necessarily optimal in terms of prediction. Barbieri and Berger [2004] show, in the framework of normal linear models, that the model selected by BIC is not necessarily the optimal predictive one. Yang [2005] also show that BIC is sub-optimal compared to AIC in terms of mean squared error.

Another class of model selection procedures which are widely used in practice are empirical criteria, such as *hold-out* [Massart, 2007], bootstrap [Efron and Tibshirani, 1993], cross-validation and its variants [Allen, 1974, Stone, 1974, Geisser, 1975, Craven and Wahba, 1979]. A test set D' is used for selecting the c for which classifier/regressor $\hat{f}_D^c$ has smallest (test) error on D'. Typically D' is cut or resampled from D. Empirical criteria are easy to understand and use, but the reduced sample decreases accuracy, which can be a serious problem if n is small. Also, they are sometimes time consuming, especially in high-dimensional and complicated settings.

1.2 Motivations and contributions

Before the data analyst proceeds to select a model, he or she needs to know what kind of model needs to be selected. Phrased differently, the goal of the model selection problem

needs to be clearly specified. Different goals may lead to different models. An important goal in data analysis is to understand the underlying structure in the data. Suppose that we are given a collection of models that reflect a range of potential structures in the data and the task is to select among this given collection a model that best explains/fits the data. It is well-known that overfitting is a serious problem in structural learning from data, and model selection is typically regarded as the question of choosing the right model complexity. Regarding this, the goal of model selection amounts to selecting a model that fits the data well but is not too complex. Most of the procedures described in the previous section aim at addressing this goal. They have been well studied and/or widely used but are not without problems. PML and BMS need a proper sampling distribution (in some problems such as kNN classification, a sampling distribution may not be available) while PELM is not easy to use in practice and empirical criteria are sometimes time demanding. Moreover, some popular criteria, such as AIC and BIC, depend heavily on the *effective number of parameters* which is in some cases, such as ridge regression and kNN regression/classification, not well defined. The first contribution of the thesis is to develop a model selection procedure addressing this first goal, i.e., selecting a parsimonious model that fits the data well. We describe in Chapter 2 a general-purpose principle for deriving model selection criteria that can avoid overfitting. The method has many attractive properties such as always giving answers, not requiring insight into the inner structure of the problem, not requiring any assumption of sampling distribution and directly applying to any non-parametric regression like kNN. The principle also leads to a nice definition of model complexity which is both data-adaptive and loss-dependent - two desirable properties for any definition of model complexity.

Another important goal in model selection is to select models that have a good (out-of-sample) predictive ability, i.e., having an explicit predictive motivation. It is still not clear

whether or not a model selection rule satisfying the first goal discussed above can also satisfy this second goal. The second contribution of this thesis is the proposal of a method addressing this second goal: we propose in Chapter 3 a model selection procedure that has an explicit predictive motivation. An application of this procedure to the variable selection problem in the generalized linear regression models with l_1 constraints on the coefficients allows us to introduce a Lasso variant - the predictive Lasso - which improves predictive ability of the original Lasso [Tibshirani, 1996].

Variable selection is probably the most fundamental problem of model selection [Fan and Li, 2001]. Regularization algorithms such as the Lasso and greedy search algorithms such as the matching pursuit are very efficient and widely used. But they are not without problems such as producing biased estimates or involving extra tuning parameters [Friedman, 2008, Nott et al., 2010]. The third contribution of the thesis is the proposal of two novel algorithms for variable selection in very general frameworks that can improve upon these existing algorithms. We first propose in Chapter 4 the Bayesian adaptive Lasso which improves on the Lasso in the sense that adaptive shinkages are used for different coefficients. We also discuss extensions for ordered and grouped variable selection. We then consider a Bayes variational approach for fast variable selection in high-dimensional heteroscedastic regression. This methodology has potential for extension to much more complicated frameworks such as simultaneous variable selection and component selection in flexible modeling with Gaussian mixture distributions.

The materials presented in this thesis either have been published or are under submission for publication [Tran, 2009, Hutter and Tran, 2010, Tran, 2011b, Tran and Hutter, 2010, Tran et al., 2010, Nott et al., 2010, Leng et al., 2010, Tran, 2011a, Tran et al., 2011].

Chapter 2

The loss rank principle

In statistics and machine learning, model selection is typically regarded as the question of choosing the right model complexity. The maximum likelihood principle breaks down when one has to select among a set of nested models, and overfitting is a serious problem in structural learning from data. Much effort has been put into developing model selection criteria that can avoid overfitting. The loss rank principle, introduced recently in Hutter [2007], and further developed in Hutter and Tran [2010], is another contribution to the model selection literature. The loss rank principle (LoRP), whose main goal is to select a parsimonious model that fits the data well, is a general-purpose principle and can be regarded as a guiding principle for deriving model selection criteria that can avoid overfitting. General speaking, the LoRP consists in the so-called *loss rank* of a model defined as the number of other (fictitious) data that fit the model better than the actual data, and the model selected is the one with the smallest loss rank. The LoRP has close connections with many well-established model selection criteria such as AIC, BIC, MDL and has many attractive properties such as always giving answers, not requiring insight into the inner structure of the problem, not requiring any assumption of sampling distribution

and directly applying to any non-parametric regression like kNN.

The LoRP will be fully presented in Section 2.1 and investigated in detail for an important class of regression models in Sections 2.2 and 2.3. Section 2.4 discusses the LoRP for model selection in the classification framework. Some numerical examples are presented in Section 2.5. Section 2.6 presents applications of the LoRP to selecting the tuning parameters in regularization regression like the Lasso. Technical proofs are relegated to Section 2.7.

The materials presented in this chapter either have been published or are under submission for publication [Tran, 2009, Hutter and Tran, 2010, Tran, 2011b, Tran and Hutter, 2010].

2.1 The loss rank principle

After giving a brief introduction to regression and classification settings, we state the loss rank principle for model selection. We first state it for the case with discrete response values (Principle 3), then generalize it for continuous response values (Principle 5), and exemplify it on two (over-simplistic) artificial Examples 4 and 6. Thereafter we show how to regularize the LoRP for realistic problems.

We assume data $D=(\mathbf{x},\mathbf{y}):=\{(x_1,y_1),...,(x_n,y_n)\}\in(\mathcal{X}\times\mathcal{Y})^n=:\mathcal{D}$ has been observed. We think of the y as having an approximate functional dependence on x, i.e., $y_i\approx f_{\text{true}}(x_i)$, where $\approx$ means that the y_i are distorted by noise from the unknown "true" values $f_{\text{true}}(x_i)$. We will write (x,y) for generic data points, use vector notation $\mathbf{x}=(x_1,...,x_n)^\top$ and $\mathbf{y}=(y_1,...,y_n)^\top$, and $D'=(\mathbf{x}',\mathbf{y}')$ for generic (fictitious) data of size n.

In regression problems $\mathcal{Y}$ is typically (a subset of) the real set $\mathbb{R}$ or some more general measurable space like $\mathbb{R}^m$. In classification, $\mathcal{Y}$ is a finite set or at least discrete. We impose

no restrictions on $\mathcal{X}$. Indeed, $\mathbf{x}$ will essentially be fixed and plays only a spectator role, so we will often notationally suppress dependencies on $\mathbf{x}$. The goal of regression/classification is to find a function $f_D \in \mathcal{F} \subset \mathcal{X} \to \mathcal{Y}$ "close" to f_{true} based on the past observations D with $\mathcal{F}$ some class of functions. Or phrased in another way: we are interested in a regressor $r : \mathcal{D} \to \mathcal{F}$ such that $\hat{y} := r(D)(x) \equiv r(x|D) \equiv f_D(x) \approx f_{\text{true}}(x)$ for all $x \in \mathcal{X}$. The quality of fit to the data is usually measured by a loss function $\text{Loss}(\mathbf{y},\hat{\mathbf{y}})$, where $\hat{y}_i = f_D(x_i)$ is an estimate of y_i. Often the loss is additive (e.g., when observations are independent): $\text{Loss}(\mathbf{y},\hat{\mathbf{y}}) = \sum_{i=1}^{n} \text{Loss}(y_i,\hat{y}_i)$.

Example 1 (polynomial regression). For $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, consider the set $\mathcal{F}_d := \{f_{\mathbf{w}}(x) = w_d x^{d-1} + ... + w_2 x + w_1 : \mathbf{w} \in \mathbb{R}^d\}$ of polynomials of degree $d-1$. Fitting the polynomial to data D, e.g., by the least squares method, we estimate $\mathbf{w}$ with $\hat{\mathbf{w}}_D$. The regression function $\hat{y} = r_d(x|D) = f_{\hat{\mathbf{w}}_D}(x)$ can be written down in closed form. This is an example of parametric regression. Popular model selection criteria such as AIC [Akaike, 1973], BIC [Schwarz, 1978] and MDL [Rissanen, 1978] can be used to select a good d. $\diamondsuit$

Example 2 (k nearest neighbors). Let $\mathcal{Y}$ be some vector space like $\mathbb{R}$ and $\mathcal{X}$ be a metric space like $\mathbb{R}^m$ with some (e.g., Euclidian) metric $d(\cdot,\cdot)$. kNN estimates $f_{\text{true}}(x)$ by averaging the y values of the k nearest neighbors $\mathcal{N}_k(x)$ of x in D, i.e., $r_k(x|D) = \frac{1}{k}\sum_{i \in \mathcal{N}_k(x)} y_i$ with $|\mathcal{N}_k(x)| = k$ such that $d(x,x_i) \leq d(x,x_j)$ for all $i \in \mathcal{N}_k(x)$ and $j \notin \mathcal{N}_k(x)$. This is an example of non-parametric regression. Popular model selection criteria such as AIC and BIC need a proper probabilistic framework which is sometimes difficult to establish in the kNN context [Holmes and Adams, 2002]. $\diamondsuit$

In the following we assume a class of regressors $\mathcal{R}$ (whatever their origin), e.g., the kNN regressors $\{r_k : k \in \mathbb{N}\}$ or the least squares polynomial regressors $\{r_d : d \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}\}$. *Each regressor r can be thought of as a model.* Throughout this chapter, we use the terms "regressor" and "model" interchangeably. Note that unlike $f \in \mathcal{F}$, regressors $r \in \mathcal{R}$ are not

functions of x alone but depend on all observations D, in particular on $\mathbf{y}$. We can compute the empirical loss of each regressor $r \in \mathcal{R}$:

$$\text{Loss}_r(D) \;\equiv\; \text{Loss}_r(\mathbf{y}|\mathbf{x}) \;:=\; \text{Loss}(\mathbf{y},\hat{\mathbf{y}}) \;=\; \sum_{i=1}^{n} \text{Loss}(y_i, r(x_i|\mathbf{x},\mathbf{y}))$$

where $\hat{y}_i = r(x_i|D)$ in the third expression, and the last expression holds in case of additive loss.

Unfortunately, minimizing Loss_r w.r.t. r will typically *not* select the "best" overall regressor. This is the well-known overfitting problem. In case of polynomials, the classes $\mathcal{F}_d \subset \mathcal{F}_{d+1}$ are nested, hence Loss_{r_d} is monotone decreasing in d with $\text{Loss}_{r_n} \equiv 0$ perfectly fitting the data. In case of kNN, Loss_{r_k} is more or less an increasing function in k with perfect fit on D for $k=1$, since no averaging takes place. In general, $\mathcal{R}$ is often indexed by a flexibility or smoothness or complexity parameter, which has to be properly determined. The more flexible r is, the closer it can fit the data (i.e., having smaller empirical loss), but it is not necessarily better since it has higher variance. Our main motivation is to develop a general selection criterion that can select a parsimonious model that fits the data well.

Definition of loss rank

We first consider discrete $\mathcal{Y}$, fix $\mathbf{x}$, denote the observed data by $\mathbf{y}$ and fictitious replicate data by $\mathbf{y}'$. The key observation we exploit is that a more flexible r can fit more data $D' \in \mathcal{D}$ well than a more rigid one. The more flexible regressor r is, the smaller the empirical loss $\text{Loss}_r(\mathbf{y}|\mathbf{x})$ is. Instead of minimizing the unsuitable $\text{Loss}_r(\mathbf{y}|\mathbf{x})$ w.r.t. r, we could ask how many $\mathbf{y}' \in \mathcal{Y}^n$ lead to smaller Loss_r than $\mathbf{y}$. We define the loss rank of r (w.r.t. $\mathbf{y}$) as the number of $\mathbf{y}' \in \mathcal{Y}^n$ with smaller or equal empirical loss than $\mathbf{y}$:

$$\text{Rank}_r(\mathbf{y}|\mathbf{x}) \equiv \text{Rank}_r(L) := \#\{\mathbf{y}' \in \mathcal{Y}^n : \text{Loss}_r(\mathbf{y}'|\mathbf{x}) \leq L\} \;\text{ with }\; L := \text{Loss}_r(\mathbf{y}|\mathbf{x}). \quad (2.1)$$

We claim that the loss rank of r is a suitable model selection measure. For (2.1) to make sense, we have to assume (and will later assure) that $\mathrm{Rank}_r(L)<\infty$, i.e., there are only finitely many $\mathbf{y}'\in\mathcal{Y}^n$ having loss smaller than L.

Since the logarithm is a strictly monotone increasing function, we can also consider the logarithmic rank $\mathrm{LR}_r(\mathbf{y}|\mathbf{x}):=\log\mathrm{Rank}_r(\mathbf{y}|\mathbf{x})$, which will be more convenient.

Principle 3 (LoRP for discrete response). *For discrete $\mathcal{Y}$, the best classifier/regressor in some class $\mathcal{R}$ for data $D=(\mathbf{x},\mathbf{y})$ is the one with the smallest loss rank:*

$$r^{best} \;=\; \arg\min_{r\in\mathcal{R}}\mathrm{LR}_r(\mathbf{y}|\mathbf{x}) \;\equiv\; \arg\min_{r\in\mathcal{R}}\mathrm{Rank}_r(\mathbf{y}|\mathbf{x}) \tag{2.2}$$

where Rank_r *is defined in* (2.1).

We give now a simple example for which we can compute all ranks by hand to help the reader better grasp how the principle works.

Example 4 (simple discrete). Consider $\mathcal{X}=\{1,2\}$, $\mathcal{Y}=\{0,1,2\}$, and two points $D=\{(1,1),(2,2)\}$ lying on the diagonal $x=y$, with polynomial (zero, constant, linear) least squares regressors $\mathcal{R}=\{r_0,r_1,r_2\}$ (see Ex.1). r_0 is simply 0, r_1 the y-average, and r_2 the line through points $(1,y_1)$ and $(2,y_2)$. This, together with the quadratic Loss for generic $\mathbf{y}'$ and observed $\mathbf{y}=(1,2)$ and fixed $\mathbf{x}=(1,2)$, is summarized in the following table

d	$r_d(x\|\mathbf{x},\mathbf{y}')$	$\mathrm{Loss}_d(\mathbf{y}'\|\mathbf{x})$	$\mathrm{Loss}_d(D)$
0	0	$y_1'^2+y_2'^2$	5
1	$\frac{1}{2}(y_1'+y_2')$	$\frac{1}{2}(y_2'-y_1')^2$	$\frac{1}{2}$
2	$(y_2'-y_1')(x-1)+y_1'$	0	0

From the Loss we can easily compute the Rank for all nine $\mathbf{y}'\in\{0,1,2\}^2$. Equal rank due to equal loss is indicated by a "=" in the table below. Whole equality groups are

actually assigned the rank of their right-most member, e.g., for $d=1$ the ranks of $(y'_1,y'_2)=$ (0,1),(1,0),(2,1),(1,2) are all 7 (and not 4,5,6,7).

	$\text{Rank}_{r_d}(y'_1y'_2\|12)$									
d	1	2	3	4	5	6	7	8	9	$\text{Rank}_{r_d}(D)$
0	$y'_1y'_2=00<01=10<11<02=20<21=\mathbf{12}<22$									8
1	$y'_1y'_2=00=11=22<01=10=21=\mathbf{12}<02=20$									7
2	$y'_1y'_2=00=01=02=10=11=20=21=22=\mathbf{12}$									9

So the LoRP selects r_1 as best regressor, since it has minimal rank on D. r_0 fits D too badly and r_2 is too flexible (perfectly fits all D'). $\diamondsuit$

LoRP for continuous $\mathcal{Y}$. We now consider the case of continuous or measurable spaces $\mathcal{Y}$, i.e., usual regression problems. We assume $\mathcal{Y}=\mathbb{R}$ in the following exposition, but the idea and resulting principle hold for more general measurable spaces like $\mathbb{R}^m$. We simply reduce the model selection problem to the discrete case by considering the discretized space $\mathcal{Y}_\varepsilon=\varepsilon\mathbb{Z}$ for small $\varepsilon>0$ and discretize $\mathbf{y}\rightsquigarrow\mathbf{y}_\varepsilon\in\varepsilon\mathbb{Z}^n$ ("$\rightsquigarrow$" means "is replaced by"). Then $\text{Rank}^\varepsilon_r(L):=\#\{\mathbf{y}'_\varepsilon\in\mathcal{Y}^n_\varepsilon:\text{Loss}_r(\mathbf{y}'_\varepsilon|\mathbf{x})\le L\}$ with $L=\text{Loss}_r(\mathbf{y}_\varepsilon|\mathbf{x})$ counting the number of ε-grid points in the set

$$V_r(L) \;:=\; \{\mathbf{y}'\in\mathcal{Y}^n:\text{Loss}_r(\mathbf{y}'|\mathbf{x})\le L\} \tag{2.3}$$

which we assume (and later assure) to be finite, analogous to the discrete case. Hence $\text{Rank}^\varepsilon_r(L)\cdot\varepsilon^n$ is an approximation of the *loss volume* $|V_r(L)|$ of set $V_r(L)$, and typically $\text{Rank}^\varepsilon_r(L)\cdot\varepsilon^n=|V_r(L)|\cdot(1+O(\varepsilon))\to|V_r(L)|$ for $\varepsilon\to0$. Taking the logarithm we get $\text{LR}^\varepsilon_r(\mathbf{y}|\mathbf{x})=\log\text{Rank}^\varepsilon_r(L)=\log|V_r(L)|-n\log\varepsilon+O(\varepsilon)$. Since $n\log\varepsilon$ is independent of r, we can drop it in comparisons like (2.2). So for $\varepsilon\to0$ we can define the log-loss "rank" simply as the log-volume

$$\text{LR}_r(\mathbf{y}|\mathbf{x}) \;:=\; \log|V_r(L)|,\quad\text{where}\quad L:=\text{Loss}_r(\mathbf{y}|\mathbf{x}). \tag{2.4}$$

Principle 5 (LoRP for continuous response). *For measurable $\mathcal{Y}$, the best regressor in some class $\mathcal{R}$ for data $D=(\mathbf{x},\mathbf{y})$ is the one with the smallest loss volume:*

$$r^{best} \;=\; \arg\min_{r\in\mathcal{R}} \mathrm{LR}_r(\mathbf{y}|\mathbf{x}) \;\equiv\; \arg\min_{r\in\mathcal{R}} |V_r(L)|$$

where LR*, V_r, and L are defined in* (2.3) *and* (2.4)*, and $|V_r(L)|$ is the volume of $V_r(L)\subseteq\mathcal{Y}^n$.*

For discrete $\mathcal{Y}$ with counting measure we recover the discrete LoRP (Principle 3).

Example 6 (simple continuous). Consider Example 4 but with interval $\mathcal{Y}=[0,2]$. The first table remains unchanged, while the second table becomes

d	$V_d(L)=\{\mathbf{y}'\in[0,2]^2:...\}$	$\lvert V_d(L)\rvert$	$\mathrm{Loss}_d(D)$	$\lvert V_d(\mathrm{Loss}_d(D))\rvert$
0	$y_1'^2+y_2'^2\le L$	$\frac{\pi}{4}L$ if $L\le 4$; 4 if $L\ge 8$; $2\sqrt{L-4}+L(\frac{\pi}{4}-\cos^{-1}(\frac{2}{\sqrt{L}}))$ else	5	$\doteq 3.6$
1	$\frac{1}{2}(y_2'-y_1')^2\le L$	$4\sqrt{2L}-2L$ if $L\le 2$; 4 if $L\ge 2$	$\frac{1}{2}$	3
2	$0\le L$	4	0	4

So LoRP again selects r_1 as best regressor, since it has smallest loss volume on D. $\diamondsuit$

Often the loss rank/volume will be infinite, e.g., if we had chosen $\mathcal{Y}=\mathbb{Z}$ in Ex.4 or $\mathcal{Y}=\mathbb{R}$ in Ex.6. Regressors r with infinite rank might be rejected for philosophical or pragmatic reasons. The solution is to modify the Loss to make LR_r finite. A very simple modification is to add a small penalty term to the loss.

$$\mathrm{Loss}_r(\mathbf{y}|\mathbf{x}) \;\rightsquigarrow\; \mathrm{Loss}_r^\alpha(\mathbf{y}|\mathbf{x}) := \mathrm{Loss}_r(\mathbf{y}|\mathbf{x}) + \alpha\|\mathbf{y}\|^2, \quad \alpha>0 \text{ "small"}. \tag{2.5}$$

The Euclidian norm $\|\mathbf{y}\|^2:=\sum_{i=1}^n y_i^2$ is default, but other (non)norm regularizations are possible. The regularized $\mathrm{LR}_r^\alpha(\mathbf{y}|\mathbf{x})$ based on Loss_r^α is always finite, since $\{\mathbf{y}:\|\mathbf{y}\|^2\le L\}$ has finite volume. An alternative penalty $\alpha\hat{\mathbf{y}}^\top\hat{\mathbf{y}}$, quadratic in the regression estimates $\hat{y}_i=r(x_i|\mathbf{x},\mathbf{y})$ is possible if r is unbounded in every $\mathbf{y}\to\infty$ direction.

A scheme trying to determine a single (flexibility) parameter (like d and k in the above examples) would be of no use if it depended on one (or more) other unknown parameters

(α), since varying through the unknown parameter leads to any (non)desired result. Since the LoRP seeks the r of smallest rank, it is natural to also determine $\alpha{=}\alpha_{\min}$ by minimizing LR_r^α w.r.t. α. The good news is that this leads to meaningful results. Interestingly, as we will see later, a clever choice of α may also result in alternative optimalities of the selection procedure.

Related ideas

There are various other ideas that somehow count fictitious data. In normalized ML [Grünwald, 2007], the complexity of a stochastic model class is defined as the log sum over all D' of maximum likelihood probabilities. The empirical Rademacher complexity [Koltchinskii, 2001, Bartlett et al., 2002] averages over all possible relabeled instances. Instead of considering all D' one could consider only the set of all permutations of $\{y_1,...,y_n\}$, like in permutation tests [Efron and Tibshirani, 1993]. Finally, instead of defining the loss rank based on fictitious $\mathbf{y}'$, if we define the loss rank based on the future observations $\mathbf{y}^f$ generated from the posterior predictive distribution $p(\mathbf{y}^f|\mathbf{y})$, then the loss rank of a model is nothing but proportional to minus posterior predictive p-value [Meng, 1994, Gelman et al., 1996] (exactly, the loss rank then = 1 - Bayesian p-value). While Gelman et al. [1996] suggest to discard models with too small (smaller than 5%, say) Bayesian p-values, the LoRP suggests to select the model with smallest loss rank (i.e., highest Bayesian p-value).

2.2 LoRP for y-Linear Models

In this section we consider the important class of *y-linear* regressions with quadratic loss function. By "y-linear regression", we mean the fitted vector is only assumed to be linear

in $\mathbf{y}$ and its dependence on $\mathbf{x}$ can be arbitrary. This class is richer than it may appear. It includes the normal linear regression model, kNN, kernel regression and many other regression models. For y-linear regression and $\mathcal{Y}=\mathbb{R}$, the loss rank is the volume of an n-dimensional ellipsoid, which can efficiently be computed in closed form (Theorem 7). For the special case of projective regression, e.g., the classical linear regression, we can even determine the regularization parameter α analytically (Theorem 8).

We assume $\mathcal{Y}=\mathbb{R}$ in this section, generalization to $\mathbb{R}^m$ is straightforward. A y-linear regressor r can be written in the form

$$\hat{y} \;=\; r(x|\mathbf{x},\mathbf{y}) \;=\; \sum_{j=1}^{n} m_j(x,\mathbf{x})y_j \quad \forall x\in\mathcal{X} \quad \text{and some} \quad m_j : \mathcal{X}\times\mathcal{X}^n \to \mathbb{R}. \tag{2.6}$$

Particularly interesting is r for $x=x_1,...,x_n$

$$\hat{y}_i \;=\; r(x_i|\mathbf{x},\mathbf{y}) \;=\; \sum_{j} M_{ij}(\mathbf{x})y_j \quad \text{with} \quad M : \mathcal{X}^n \to \mathbb{R}^{n\times n},\;\; M_{ij}(\mathbf{x}) = m_j(x_i,\mathbf{x}), \tag{2.7}$$

i.e., the fitted vector can be written in the form $\hat{\mathbf{y}}=M\mathbf{y}$. For example, in kNN regression we have $m_j(x,\mathbf{x})=\frac{1}{k}$ if $j\in\mathcal{N}_k(x)$ and 0 else, and $M_{ij}(\mathbf{x})=\frac{1}{k}$ if $j\in\mathcal{N}_k(x_i)$ and 0 else. Another example is kernel regression which takes a weighted average over $\mathbf{y}$, where the weight of y_j to y is proportional to the similarity of x_j to x, measured by a kernel $K(x,x_j)$, i.e., $m_j(x,\mathbf{x})=K(x,x_j)/\sum_{j=1}^{n}K(x,x_j)$.

Consider now a general linear regressor M with quadratic loss and quadratic penalty as in (2.5)

$$\text{Loss}_M^{\alpha}(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{n}\Big(y_i - \textstyle\sum_{j=1}^{n} M_{ij}y_j\Big)^2 + \alpha\|\mathbf{y}\|^2 \;=\; \mathbf{y}^\top S_\alpha \mathbf{y}$$

where

$$S_\alpha = (\mathbb{1}-M)^\top(\mathbb{1}-M)+\alpha\mathbb{1} \tag{2.8}$$

($\mathbb{1}$ is the identity matrix). S_α is a symmetric matrix, for $\alpha>0$ it is positive definite and for $\alpha=0$ positive semidefinite. If $\lambda_1,...,\lambda_n\geq 0$ are the eigenvalues of S_0, then $\lambda_i+\alpha$ are the

eigenvalues of S_α. $V(L)=\{\mathbf{y}'\in\mathbb{R}^n:\mathbf{y}'^\top S_\alpha\mathbf{y}'\leq L\}$ is an ellipsoid with the eigenvectors of S_α being the main axes and $\sqrt{L/(\lambda_i+\alpha)}$ being their length. Hence the volume is

$$|V(L)| \;=\; v_n\prod_{i=1}^{n}\sqrt{\frac{L}{\lambda_i+\alpha}} \;=\; \frac{v_n L^{n/2}}{\sqrt{\det S_\alpha}}$$

where $v_n=\pi^{n/2}/\Gamma(\frac{n}{2}+1)$ is the volume of the n-dimensional unit sphere, and det is the determinant. Taking the logarithm we get

$$\mathrm{LR}^\alpha_M(\mathbf{y}|\mathbf{x}) \;=\; \log|V(\mathrm{Loss}^\alpha_M(\mathbf{y}|\mathbf{x}))| \;=\; \frac{n}{2}\log(\mathbf{y}^\top S_\alpha\mathbf{y})-\frac{1}{2}\log\det S_\alpha+\log v_n. \qquad (2.9)$$

Since v_n is independent of α and M it is possible to drop v_n. Consider now a class of y-linear regressors $\mathcal{M}=\{M\}$, e.g., the kNN regressors $\{M_k:k\in\mathbb{N}\}$ or the d-order polynomial regressors $\{M_d:d\in\mathbb{N}_0\}$.

Theorem 7 (LoRP for y-linear regression). *For $\mathcal{Y}=\mathbb{R}$, the best linear regressor $M:\mathcal{X}^n\to\mathbb{R}^{n\times n}$ in some class $\mathcal{M}$ for data $D=(\mathbf{x},\mathbf{y})$ is*

$$M^{best} \;=\; \mathop{\arg\min}_{M\in\mathcal{M},\alpha\geq 0}\{\frac{n}{2}\log(\mathbf{y}^\top S_\alpha\mathbf{y})-\frac{1}{2}\log\det S_\alpha\} \;=\; \mathop{\arg\min}_{M\in\mathcal{M}\;\alpha\geq 0}\Big\{\frac{\mathbf{y}^\top S_\alpha\mathbf{y}}{(\det S_\alpha)^{1/n}}\Big\} \qquad (2.10)$$

where $S_\alpha=S_\alpha(M)$ is defined in (2.8).

Note that M^{best} depends on $\mathbf{y}$ unlike the $M\in\mathcal{M}$. In general we need to find the optimal α numerically, however, it can be found analytically when M is a projection (Theorem 8). For each α and candidate model, the determinant of S_α in the general case can be computed in time $O(n^3)$. Often M is a very sparse matrix (like in kNN) or can be well approximated by a sparse matrix (like for kernel regression), which allows us to approximate $\det S_\alpha$ sometimes in linear time [Reusken, 2002]. To search the optimal α and M, the computational cost depends on the range of α we search and the number of candidate models we have.

Projective regression. Consider a projection matrix $M=P=P^2$ with $d(=\mathrm{tr}P)$ eigenvalues 1, and $n-d$ zero eigenvalues. This implies that S_α has d eigenvalues α and $n-d$ eigenvalues $1+\alpha$, thus $\det S_\alpha=\alpha^d(1+\alpha)^{n-d}$. Let $\rho=\|\mathbf{y}-\hat{\mathbf{y}}\|^2/\|\mathbf{y}\|^2$, then $\mathbf{y}^\top S_\alpha\mathbf{y}=(\rho+\alpha)\mathbf{y}^\top\mathbf{y}$ and

$$\mathrm{LR}_P^\alpha = \frac{n}{2}\log\mathbf{y}^\top\mathbf{y} + \frac{n}{2}\log(\rho+\alpha) - \tfrac{d}{2}\log\alpha - \tfrac{n-d}{2}\log(1+\alpha). \tag{2.11}$$

Solving $\partial\mathrm{LR}_P^\alpha/\partial\alpha=0$ w.r.t. α we get a minimum at $\alpha=\alpha_m:=\frac{\rho d}{(1-\rho)n-d}$ provided that $1-\rho>d/n$. After some algebra we get

$$\mathrm{LR}_P^{\alpha_m} = \tfrac{n}{2}\log\mathbf{y}^\top\mathbf{y} - \tfrac{n}{2}\mathrm{KL}(\tfrac{d}{n}\|1-\rho), \tag{2.12}$$

where $\mathrm{KL}(p\|q):=p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$ is the relative entropy or the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q. Note that (2.12) is still valid without the condition $1-\rho>d/n$ (the term $\log((1-\rho)n-d)$ has been canceled in the derivation). What we need when using (2.12) is that $d<n$ and $\rho<1$, which are very reasonable in practice. Interestingly, if in (2.5) we use the penalty $\alpha\|\hat{\mathbf{y}}\|^2$ instead of $\alpha\|\mathbf{y}\|^2$, the loss rank then has the same expression as (2.12) without any condition[1].

Minimizing $\mathrm{LR}_P^{\alpha_m}$ w.r.t. P is equivalent to maximizing $\mathrm{KL}(\frac{d}{n}\|1-\rho)$. The term ρ is a measure of fit. If d increases, then ρ decreases and conversely. We are seeking a tradeoff between the model complexity d and the measure of fit ρ, and the LoRP suggests the optimal tradeoff by maximizing the KL.

Theorem 8 (LoRP for projective regression). *The best projective regressor $P:\mathcal{X}^n\to\mathbb{R}^{n\times n}$ with $P=P^2$ in some projective class $\mathcal{P}$ for data $D=(\mathbf{x},\mathbf{y})$ is*

$$P^{best} \;=\; \arg\max_{P\in\mathcal{P}} \mathrm{KL}(\tfrac{\mathrm{tr}P(\mathbf{x})}{n}\|\tfrac{\mathbf{y}^\top P(\mathbf{x})\mathbf{y}}{\mathbf{y}^\top\mathbf{y}}). \tag{2.13}$$

[1]Then $S_\alpha=(\mathbb{1}_n-P)^\top(\mathbb{1}_n-P)+\alpha P^\top P=\mathbb{1}_n+(\alpha-1)P$ has d eigenvalues α and $n-d$ eigenvalues 1, thus $\det(S_\alpha)=\alpha^d$. The loss rank $\mathrm{LR}_P^\alpha=\frac{n}{2}\log\mathbf{y}^\top\mathbf{y}+\frac{n}{2}\log(1+(\alpha-1)(1-\rho))-\frac{d}{2}\log\alpha$ is minimized at $\alpha_m=\frac{\rho d}{(1-\rho)(n-d)}$. After some algebra we get the same expression of $\mathrm{LR}_P^{\alpha_m}$ as (2.12).

2.3 Optimality properties of the LoRP for variable selection

In the previous sections, the LoRP was stated for general-purpose model selection. By restricting our attention to linear regression models, we will point out in this section some theoretical properties of the LoRP for variable (also called feature or attribute) selection.

Variable selection is a fundamental topic in linear regression analysis. At the initial stage of modeling, a large number of potential covariates are often introduced; one then has to select a smaller subset of the covariates to fit/interpret the data. There are two main goals of variable selection, one is model identification, the other is regression estimation. The former aims at identifying the true subset generating the data, while the latter aims at estimating efficiently the regression function, i.e., selecting a subset that has the minimum mean squared error loss. Note that whether or not there is a selection criterion achieving simultaneously these two goals is still an open question [Yang, 2005, Grünwald, 2007]. We show that with the optimal parameter α (defined as α_m that minimizes the loss rank LR_M^α in α), the LoRP satisfies the first goal, while with a suitable choice of α, the LoRP satisfies the second goal.

Given $d+1$ potential covariates $X_0 \equiv 1, X_1, ..., X_d$ and a response variable Y, let $X = \mathbf{x}$ be a non-random design matrix of size $n \times (d+1)$ and $\mathbf{y}$ be a response vector, respectively (if $\mathbf{y}$ and X are centered, then the covariate 1 can be omitted from the models). Denote by $\mathcal{S} = \{0, j_1, ... j_{|\mathcal{S}|-1}\}$ the candidate model that has covariates $X_0, X_{j_1}, ..., X_{j_{|\mathcal{S}|-1}}$. Under a proposed model $\mathcal{S}$, we can write

$$\mathbf{y} \;=\; X_{\mathcal{S}} \beta_{\mathcal{S}} + \sigma \epsilon$$

where ϵ is a vector of noise with expectation $\mathbf{E}[\epsilon] = 0$ and covariance $\mathrm{Cov}(\epsilon) = \mathbb{1}_n$, $\sigma > 0$,

$\beta_{\mathcal{S}}=(\beta_0,\beta_{j_1},...,\beta_{j_{|\mathcal{S}|-1}})^\top$, and $X_{\mathcal{S}}$ is the $n\times|\mathcal{S}|$ design matrix obtained from X by removing the $(j+1)$st column for all $j\notin\mathcal{S}$.

2.3.1 Model consistency of the LoRP for variable selection

The ordinary least squares (OLS) fitted vector under model $\mathcal{S}$ is $\hat{\mathbf{y}}_{\mathcal{S}}=M_{\mathcal{S}}\mathbf{y}$ with $M_{\mathcal{S}}=X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}X_{\mathcal{S}}^\top$ being a projection matrix. From Theorem 8 the best subset chosen by the LoRP is

$$\hat{\mathcal{S}}_n \;=\; \arg\min_{\mathcal{S}} \mathrm{LR}_{\mathcal{S}}^{\alpha_m} \;=\; \arg\max_{\mathcal{S}}\{\mathrm{KL}(\frac{|\mathcal{S}|}{n}\|1-\rho_{\mathcal{S}})\},\;\; \rho_{\mathcal{S}}=\frac{\|\mathbf{y}-\hat{\mathbf{y}}_{\mathcal{S}}\|^2}{\|\mathbf{y}\|^2}.$$

The term $\rho_{\mathcal{S}}$ is a measure of fit. It will be very close to 0 if model $\mathcal{S}$ is big, otherwise, it will be close to 1 if $\mathcal{S}$ is too small. Therefore, it is reasonable to consider only cases in which $\rho_{\mathcal{S}}$ is bounded away from 0 and 1. In order to prove the theoretical properties of the LoRP, we need the following technical assumption.

(A) For each candidate model $\mathcal{S}$, $\rho_{\mathcal{S}}$ is bounded away from 0 and 1, i.e., there are constants c_1 and c_2 such that $0<c_1\le\rho_{\mathcal{S}}\le c_2<1$ with probability 1 (w.p.1).

Let $\hat{\sigma}^2_{\mathcal{S}}=\|\mathbf{y}-\hat{\mathbf{y}}_{\mathcal{S}}\|^2/n$ and $\mathcal{S}_{\text{null}}=\{0\}$. It is easy to see that for every $\mathcal{S}$

$$1-\rho_{\mathcal{S}}=\|\hat{\mathbf{y}}_{\mathcal{S}}\|^2/\|\mathbf{y}\|^2,\qquad n\hat{\sigma}^2_{\mathcal{S}}=\rho_{\mathcal{S}}\|\mathbf{y}\|^2,\qquad n\,\bar{\mathbf{y}}^2=\|\hat{\mathbf{y}}_{\mathcal{S}_{\text{null}}}\|^2\le\|\hat{\mathbf{y}}_{\mathcal{S}}\|^2\le\|\mathbf{y}\|^2 \quad (2.14)$$

where $\bar{\mathbf{y}}$ denotes the arithmetic mean $\sum_{i=1}^n y_i/n$. Assumption (A) follows from

(A') $0<\liminf_{n\to\infty}(\bar{\mathbf{y}})^2\le\limsup_{n\to\infty}(\frac{1}{n}\|\mathbf{y}\|^2)<\infty$ and $\forall\mathcal{S}$: $\hat{\sigma}^2_{\mathcal{S}}\to\text{constant}>0$ w.p.1.

The first condition of (A') is obviously very mild and satisfied in almost all cases in practice. The second one is routinely used to derive asymptotic properties of model selection criteria (e.g., Theorem 2 of Shao [1997] and Condition 1 of Wang et al. [2007]).

Lemma 9. *The loss rank of model $\mathcal{S}$ is*

$$\mathrm{LR}_{\mathcal{S}} \equiv \mathrm{LR}_{\mathcal{S}}^{\alpha_m} = \frac{n}{2}\log(n\hat{\sigma}_{\mathcal{S}}^2) + \frac{n}{2}H(\frac{|\mathcal{S}|}{n}) + \frac{d}{2}\log\frac{1-\rho_{\mathcal{S}}}{\rho_{\mathcal{S}}} \tag{2.15}$$

where $\rho_{\mathcal{S}}$ and $\hat{\sigma}_{\mathcal{S}}^2$ are defined in (2.14)*, and $H(p) := -p\log p - (1-p)\log(1-p)$ is the entropy of p. Under Assumption (A) or (A'), after neglecting constants independent of $\mathcal{S}$, the loss rank of model $\mathcal{S}$ has the form*

$$\mathrm{LR}_{\mathcal{S}} = \frac{n}{2}\log\hat{\sigma}_{\mathcal{S}}^2 + \frac{|\mathcal{S}|}{2}\log n + O_{\mathrm{P}}(1), \tag{2.16}$$

where $O_{\mathrm{P}}(1)$ denotes a bounded random variable w.p.1.

The proof is relegated to Section 2.7. This lemma implies that the loss rank $\mathrm{LR}_{\mathcal{S}}$ here is asymptotically a BIC-type criterion, thus we immediately can state without proof the following theorem which is the well-known model consistency of BIC-type criteria (see, for example, Chambaz [2006]).

Theorem 10 (Model consistency). *Under Assumption (A) or (A'), the LoRP is model consistent for variable selection in the sense that the probability of selecting the true model goes to 1 for data size $n \to \infty$.*

2.3.2 The optimal regression estimation of the LoRP

The second goal of model selection is often measured by the (asymptotic) mean efficiency [Shibata, 1983] which is briefly defined as follows. Let $\mathcal{S}_T$ denote the true model (which may contain an infinite number of covariates). For a candidate model $\mathcal{S}$, let $L_n(\mathcal{S}) = \|X_{\mathcal{S}_T}\beta_{\mathcal{S}_T} - X_{\mathcal{S}}\hat{\beta}_{\mathcal{S}}\|^2$ be the squared loss where $\hat{\beta}_{\mathcal{S}}$ is the OLS estimate, and $R_n(\mathcal{S}) = \mathbf{E}[L_n(\mathcal{S})]$ be the risk. The mean efficiency of a selection criterion δ is defined by the ratio

$$\mathrm{eff}(\delta) = \frac{\inf_{\mathcal{S}} R_n(\mathcal{S})}{\mathbf{E}[L_n(\mathcal{S}_\delta)]} \tag{2.17}$$

where $\mathcal{S}_\delta$ is the model selected by the method δ. Note that $\mathrm{eff}(\delta) \leq 1$. δ is said to be asymptotically mean efficient if $\liminf_{n\to\infty} \mathrm{eff}(\delta) = 1$.

By minimizing the loss rank in α we have shown that the LoRP satisfies the first goal of model selection. We now show that with a suitable choice of α, the LoRP also satisfies the second goal.

From (2.11), we have that

$$\mathrm{LR}^{\alpha}_{\mathcal{S}}(\mathbf{y}|\mathbf{x}) = \frac{n}{2}\log(\hat{\sigma}^2_{\mathcal{S}} + \frac{\alpha}{n}\mathbf{y}^\top\mathbf{y}) + \frac{n}{2}\log n - \frac{|\mathcal{S}|}{2}\log(\alpha) - \frac{n-|\mathcal{S}|}{2}\log(1+\alpha).$$

By choosing $\alpha = \tilde{\alpha} = \exp(-\frac{n(n+|\mathcal{S}|)}{|\mathcal{S}|(n-|\mathcal{S}|-2)})$, under Assumption (A), the loss rank of model $\mathcal{S}$ (neglecting the common constant $\frac{n}{2}\log n$) is proportional to

$$\mathrm{LR}^{\tilde{\alpha}}_{\mathcal{S}}(\mathbf{y}|\mathbf{x}) = n\log\hat{\sigma}^2_{\mathcal{S}} + \frac{n(n+|\mathcal{S}|)}{n-|\mathcal{S}|-2} + o_{\mathrm{P}}(1),$$

which is the corrected AIC of Hurvich and Tsai [1989]. As a result, the LoRP($\tilde{\alpha}$) is optimal in terms of regression estimation, i.e., it is asymptotically mean efficient [Shibata, 1983, Shao, 1997].

Theorem 11 (Asymptotic mean efficiency). *Under Assumption (A) or (A'), with a suitable choice of α, the loss rank is proportional to the corrected AIC. As a result, the LoRP is asymptotically mean efficient.*

2.4 LoRP for classification

We consider in this section the model selection problem in a (binary) classification framework. Let $D = \{(X_1,Y_1),...,(X_n,Y_n)\}$ be n independent realizations of random variables (X,Y), where X takes on values in some space $\mathcal{X}$ and Y is a $\{0,1\}$-valued random variable. We assume that these pairs are defined on a probability space $(\Omega,\Sigma,\mathrm{P})$ with $\Omega = (\mathcal{X}\times\mathcal{Y})^n$.

We are interested in constructing a predictor $t : \mathcal{X} \to \{0,1\}$ that predicts Y based on X. The performance of the predictor t is ideally measured by the prediction loss

$$\mathrm{P}\gamma(t) = \mathrm{P}(I_{Y \neq t(X)}) = \mathrm{P}(Y \neq t(X)) \tag{2.18}$$

where $\gamma(t)(x,y) := I_{y \neq t(x)}$ is called the contrast function. Hereafter, for a measure μ and a μ-integrable function f, we denote the integral $\int f d\mu$ by μf or $\mu(f)$.

Ideally, we want to seek an optimal predictor s that minimizes $\mathrm{P}\gamma(t)$ over all measurable $t : \mathcal{X} \to \{0,1\}$. However, finding such a predictor is impossible in practice because the class of all measurable functions $t : \mathcal{X} \to \{0,1\}$ is huge and typically not specified. Instead, we have to restrict to some small class of predictors $\mathcal{F}$. A question arises immediately here: how small should the class $\mathcal{F}$ be? A too small $\mathcal{F}$ may lead to an unreasonable prediction loss, while finding an optimizer in a too large $\mathcal{F}$ may be an impossible task. Therefore the class/model $\mathcal{F}$ itself must be selected as well (the terms *class* and *model* will be used interchangeably). In this section, we are interested in the model selection problem in which we would like to find a good model (in a sense specified later on) in a given set of models $\{\mathcal{F}_m,\, m \in \mathcal{M}\}$.

The unknown prediction loss (2.18) is often estimated by the empirical risk

$$\mathrm{P}_n\gamma(t) = \frac{1}{n}\sum_{1}^{n} I_{Y_i \neq t(X_i)} \tag{2.19}$$

where P_n is the empirical measure based on data D, $\mathrm{P}_n = \frac{1}{n}\sum_1^n \delta_{(X_i,Y_i)}$, with δ_x denotes the Dirac measure at x. For a class $\mathcal{F}_m$, one may seek a function $\hat{t}_m$ minimizing $\mathrm{P}_n\gamma(t)$ over $t \in \mathcal{F}_m$ and then choose model $\hat{m} = \inf_m \mathrm{P}_n\gamma(\hat{t}_m)$. Unfortunately, it is well-known that such a method leads to overfitting: the larger $\mathcal{F}_m$, the smaller the empirical risk $\mathrm{P}_n\gamma(\hat{t}_m)$. Consequently, the selected model is always the biggest one if the classes $\mathcal{F}_m$ are nested. This leads to the idea of accounting for the model complexity, in which we select a model

$\hat{m}$ that minimizes the sum of the empirical risk and a penalty term taking the model complexity into account.

Because $\mathrm{P}_n\gamma(t)$ underestimates $\mathrm{P}\gamma(t)$, a well-known regularized criterion for model selection is to penalize the approximation on $\mathcal{F}_m$ of the prediction loss by the empirical risk (see, e.g., [Koltchinskii, 2001, Fromont, 2007, Arlot, 2009])

$$\mathrm{crit}_n(m) = \mathrm{P}_n\gamma(\hat{t}_m) + \sup_{t\in\mathcal{F}_m} (\mathrm{P} - \mathrm{P}_n)\gamma(t). \tag{2.20}$$

The second term, denoted by $\mathrm{pen}_n(m)$, is a natural measure of the complexity of class $\mathcal{F}_m$, which measures the accuracy of empirical approximation on class $\mathcal{F}_m$. Then, the model to be selected is $m_n = \mathrm{argmin}_m\{\mathrm{crit}_n(m)\}$. For simplicity, we assume that m_n is uniquely determined.

In practice, P is unknown and so is $\mathrm{pen}_n(m)$. One has to estimate $\mathrm{pen}_n(m)$. Many methods have been proposed to estimate this theoretical penalty: VC-dimension [Vapnik and Chervonenkis, 1971], Rademacher complexities [Koltchinskii, 2001, Bartlett et al., 2002], resampling penalties [Fromont, 2007, Arlot, 2009]. All of these methods give upper bounds for $\mathrm{pen}_n(m)$. The performances of the methods are measured in terms of oracle inequalities. The sharper the estimate is, the better the performance is. These methods often works well in practice but are not without problems. For example, the VC-dimension is often unknown and needs to be estimated by another upper bound, Rademacher complexities are often criticized to be too large (the local Rademacher complexities [Bartlett et al., 2005, Koltchinskii, 2006] have been introduced to overcome this drawback, however the latter still suffer from the hard-calibration problem because they involve unknown constants).

In this section, based on the LoRP, we obtain a criterion to estimate the model m_n directly, *not* pen_n. Instead of giving an upper bound for $\mathrm{pen}_n(m)$, we directly estimate m_n

by minimizing a criterion over models $m \in \mathcal{M}$. Minimizing the criterion is asymptotically equivalent to minimizing $\text{crit}_n(m)$ with probability 1 (Theorem 12).

The criterion is derived in Section 2.4.1, and its optimality property is given in Section 2.4.2. A numerical example to demonstrate the criterion is given in Section 2.5.

2.4.1 The loss rank criterion

Let us recall the basic idea of the LoRP. Let $D = (\mathbf{x},\mathbf{y}) = \{(x_1,y_1),...,(x_n,y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ be the (actual) training data set with $\mathbf{x} = (x_1,...,x_n)$ are inputs and $\mathbf{y} = (y_1,...,y_n)$ are (perturbed) outputs. Let $\mathbf{y}'$ be other (fictitious) outputs (imagine that in experiment situations we can conduct the experiment many times with fixed design points $\mathbf{x}$, we then would get many other $\mathbf{y}'$). Suppose that we are using a model $M \in \mathcal{M}$ to fit the data D. Let $\text{Loss}_M(\mathbf{y}|\mathbf{x})$ be the empirical loss associated with a certain loss function when using a model $M \in \mathcal{M}$ to fit the data set $(\mathbf{x},\mathbf{y})$. The loss rank of model M then is defined as

$$\text{LR}_M(D) := \mu\{\mathbf{y}' \in \mathcal{Y}^n : \text{Loss}_M(\mathbf{y}'|\mathbf{x}) \leq \text{Loss}_M(\mathbf{y}|\mathbf{x})\} \tag{2.21}$$

with some measure μ on $\mathcal{Y}^n$. For example, μ can be the counting measure if $\mathcal{Y}$ is discrete, the usual Lebesgue measure on $\mathbb{R}^n$ if $\mathcal{Y} = \mathbb{R}$. As seen in the previous sections, for continuous data cases, using the usual Lebesgue measure leads to a closed form of loss rank and meaningful results.

The LoRP, as it is named, is a guiding principle rather than a specific selection criterion. When it comes to apply in a specific context, a suitable choice of measure μ in (2.21) is needed. In our current context of the binary classification, some suitable probability measure on $\mathcal{Y}^n = \{0,1\}^n$ should be used to define the loss rank. To formalize this, we define the loss rank of a model as the probability that a randomly resampled sample fit the model better than the actual sample. This definition of the loss rank makes it not only

possible to estimate the loss rank but also makes use of the available theory of resampling to justify the method.

We now formally define the loss rank. Let r_i, $i=1,...,n$ be n independent Rademacher random variables, i.e., r_i takes on values either -1 or 1 with probability $1/2$. The r_i's are assumed to be independent of D. Let $Y_i' := \frac{1+r_i}{2} - r_i Y_i$, i.e., we flip the value/label of Y_i with probability $1/2$. The loss rank of a model m is defined as

$$\mathrm{LR}_n(m) \equiv \mathrm{LR}_n(\mathcal{F}_m) := \mathrm{P}_R(\inf_{t\in\mathcal{F}_m} \frac{1}{n}\sum_1^n I_{Y_i'\neq t(X_i)} \leq \mathrm{P}_n\gamma(\hat{t}_m)|D) \tag{2.22}$$

where $\mathrm{P}_R(.|D)$ denotes the conditional probability w.r.t. the Rademacher sequence given data D. The selected model will be $\hat{m}_{\mathrm{LR}} = \mathrm{argmin}_{m\in\mathcal{M}}\mathrm{LR}_n(m)$. We name this method the loss rank (LR) criterion.

Intuitively, the empirical risk based on the actual D would be small for a too flexible class $\mathcal{F}_m$, but many resamples D' would then also result in small empirical risk, which leads to a large loss rank $\mathrm{LR}_n(m)$. Therefore, minimizing the loss rank helps avoid overfitting. Also, a too rigid $\mathcal{F}_m$ fitting D not well would lead to a large loss rank as well. Thus, the loss rank defined in (2.22) is a suitable criterion for model selection which trades off between the fit (empirical risk) and the model complexity.

The loss rank $\mathrm{LR}_n(m)$ (2.22) can be easily estimated by a simple Monte Carlo algorithm as follows:

1. $\hat{\mathrm{LR}}_n(m) \leftarrow 0$.

2. Toss a fair coin n times and define

$$Y_i' = \begin{cases} Y_i, & \text{head occurs at } i\text{-th time} \\ 1-Y_i, & \text{tail occurs at } i\text{-th time} \end{cases}, \ i = 1, 2, ..., n.$$

   If $\inf_{t\in\mathcal{F}_m}\frac{1}{n}\sum_1^n I_{Y_i'\neq t(X_i)} \leq \mathrm{P}_n\gamma(\hat{t}_m)$ then $\hat{\mathrm{LR}}_n(m) \leftarrow \hat{\mathrm{LR}}_n(m) + 1/B$.

3. Repeat step 2, B times.

The theoretical justification for this algorithm is the law of large numbers: $\hat{\mathrm{LR}}_n(m) \to \mathrm{LR}_n(m)$ *a.s.* as $B \to \infty$.

2.4.2 Optimality property

We now discuss the model consistency of the LR criterion by using the modern theory of empirical processes (see, e.g., van der Vaart and Wellner [1996]). To avoid dealing with difficulties of non-measurability in empirical process theory, we as usual assume that for each $m \in \mathcal{M}$, class $\mathcal{F}_m$ is countable. We need the following regularity condition:

(C) $\mathcal{D}_m = \{\gamma(t), t \in \mathcal{F}_m\}$, $m \in \mathcal{M}$ are Donsker classes.

Recall that a function class $\mathcal{F}$ is called a Donsker class if $\sqrt{n}(\mathrm{P}_n - \mathrm{P})f$ converges in probability to $N(0, \mathrm{P}(f - \mathrm{P}f)^2)$ uniformly in $f \in \mathcal{F}$. This, together with another condition that $\mathrm{P}\big(\sup_{f \in \mathcal{F}} |f - \mathrm{P}f|^2\big) < \infty$ (which is automatically satisfied in our context because $\gamma(t) \leq 1$ for every predictor t) are essential in order for the weak convergence of empirical processes to hold [van der Vaart and Wellner, 1996, Chapter 3]. These are also two essential conditions in order for Efron's bootstrap to be asymptotically valid [Gine and Zinn, 1990].

Theorem 12. *Under Assumption (C), minimizing* $\mathrm{LR}_n(m)$ *in* (2.22) *over* $m \in \mathcal{M}$ *is asymptotically equivalent to minimizing the ideal criterion* $\mathrm{crit}_n(m)$ *in* (2.20) *with probability 1, i.e.,* $\hat{m}_{\mathrm{LR}}$ *is a strong consistent estimate of* m_n.

On one hand, the LR criterion is closely related to penalized model selection based on Rademacher complexities. As realized by Lozano [2000], a very large model which generally contains a predictor predicting correctly most randomly generated labels results in a large Rademacher penalty. While a very large model will result in a large loss rank

which is the probability that a randomly relabeled sample behaves better than the actual sample. On the other hand, LR criterion is quite different from model selection based on Rademacher complexities. While Rademacher complexities give upper bounds for the ideal penalty $\text{pen}_n(m)$, the LR criterion offers a way to directly estimate the ideal model m_n. The proof of the theorem can be found in Section 2.7.

2.5 Numerical examples

In this section we present a number of numerical examples to demonstrate how the LoRP works in various model selection problems.

2.5.1 Comparison to AIC and BIC for model identification

Samples are generated from the model

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_d X_d + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{2.23}$$

where β is the vector of coefficients with some zero entries. Without loss of generality, we assume that $\beta_0 = 0$, otherwise, we can center the response vector $\mathbf{y}$ and standardize the design matrix X to exclude β_0 from the model. We shall compare the performance of LoRP to that of BIC and AIC with various factors n, d and signal-to-noise ratio (SNR) which is $\|\beta\|^2/\sigma^2$ ($\|\beta\|^2$ is often called the length of the signal).

For a given set of factors (n, d, SNR), the way we simulate a dataset from model (2.23) is as follows. Entries of X are sampled from a uniform distribution on $[-1,1]$. To generate β, we first create a vector $\mathbf{u} = (u_1,...,u_d)^\top$ whose entries are sampled from a uniform distribution on $[-1,1]$. The number of true covariates d^* is randomly selected from $\{1,2,...,d\}$, the last $d-d^*$ entries of $\mathbf{u}$ are set to zero, then coefficient vector β is computed

by $\beta_i = \{\text{length of signal}\} * u_i / \|\mathbf{u}\|$. In our simulation, the length of signal was fixed to be 10. n observation errors $\epsilon_1,...,\epsilon_n$ are sampled from a normal distribution with mean 0 and variance $\sigma^2 = \|\beta\|^2/\text{SNR}$. Finally, the response vector is computed by $\mathbf{y} = X\beta + \epsilon$. For each set of factors (n, d, SNR), 1000 datasets are simulated in the same manner to assess the average performance of the methods. For simplicity, a candidate model is specified by its order, i.e., we search the best model among only d models $\{1\},\{1,2\}...,\{1,2,...,d\}$. For the general case, an efficient branch-and-bound algorithm [see, e.g., Miller, 2002, Chapter 3] can be used to exhaustively search for the best subsets.

Table 2.1 presents percentages of correctly-fitted models with various factors n, d and SNR. As shown, LoRP outperforms the others. The better performance of LoRP over BIC, which is the most popular criterion for model identification, is very encouraging. This is probably because LoRP is a selection criterion with a data-dependent penalty. This improvement needs a theoretical justification which we intend to do in the future. Note that the equivalence between LoRP and BIC as shown in Lemma 9 is only asymptotic.

2.5.2 Comparison to AIC and BIC for regression estimation

Consider the following model which is taken from Shibata [1983]

$$y = y(x) = \log \frac{1}{1-x} + \epsilon, \;\; \epsilon \sim N(0, \sigma^2), \; x \in [0, 1). \tag{2.24}$$

We approximate the true function by a Fourier series and consider the problem of choosing a good order among models

$$y = \beta_0 + \sum_{l=1}^{k-1} \frac{\cos(\pi l x/\delta)}{l+1} \beta_l + \epsilon, \;\; k = 1, ..., K.$$

In the present context, a model in is completely specified by the order k of the Fourier series. Samples are created from (2.24) at the points $x_i = \delta \frac{i}{n+1}$, $i = 1,...,n$. As in Shibata

Table 2.1: Comparison of LoRP to AIC and BIC for model identification: Percentage of correctly-fitted models over 1000 replications with various factors n, d and signal-to-noise ratio (SNR).

n	d	SNR	AIC	BIC	LoRP	n	d	SNR	AIC	BIC	LoRP
100	5	1	62	62	69	300	5	1	74	82	83
		5	85	85	86			5	78	90	91
		10	80	90	91			10	81	94	94
	10	1	52	42	54		10	1	63	67	71
		5	63	77	77			5	70	85	86
		10	68	84	85			10	74	90	90
	20	1	32	22	36		20	1	54	45	61
		5	55	63	65			5	64	79	80
		10	56	73	74			10	67	85	85

[1983], we take $\delta = .99$, and $K = 163$ with various n and σ. The performance is measured by the estimate of mean efficiency (2.17) over 1000 replications.

Table 2.2 represents the simulation results. In general, LoRP (with $\alpha = \tilde{\alpha}$ as in Section 2.3) outperforms the others, except for cases with unrealistically high noise level. For cases with high noise, mean efficiency of BIC is often larger than that of AIC and LoRP. This was also shown in the simulation study of Shibata [1983], Table 1. This phenomenon can be explained as follows.

The risk of model k (the model specified by its order k) is $R_n(k) = \|(\mathbb{1} - M_k)\mathbf{y}_{\text{true}}\|^2 + k\sigma^2$ where M_k is the regression matrix under model k and $\mathbf{y}_{\text{true}}$ is the vector of true values $y(x_i)$. When $\sigma \to \infty$, the ideal $k^\star = \text{arginf}_k R_n(k) \to 1$. Because BIC penalizes the model complexity more strongly than AIC and LoRP do, the order chosen by BIC is closer to

$k^\star = 1$ than the ones chosen by AIC and LoRP. As a result, mean efficiency of BIC is larger than that of the others.

Table 2.2: Comparison of LoRP to AIC and BIC for regression estimation: Estimates of mean efficiency over 1000 replications with various factors n, d and signal-to-noise ratio (SNR).

n	σ	AIC	BIC	LoRP	n	σ	AIC	BIC	LoRP
400	.001	1.00	.98	.99	600	.001	1.00	.98	1.00
	.01	.93	.68	.90		.01	.99	.67	.92
	.05	.88	.67	.95		.05	.90	.66	.94
	.1	.88	.67	.92		.1	.90	.67	.93
	.5	.81	.66	.85		.5	.82	.66	.83
	1	.79	.63	.82		1	.79	.65	.82
	5	.67	.65	.70		5	.65	.67	.66
	10	.54	.67	.59		10	.54	.59	.54
	100	.31	.89	.33		100	.40	.90	.41

2.5.3 Selection of number of neighbors in kNN regression

Let us now see how the LoRP can be applied to select a good parameter k in kNN regression. We create a dataset of $n = 100$ observations (x_i, y_i) from the model:

$$y = f(x) + \varepsilon, \text{ with } f(x) = \frac{\sin(12(x+0.2))}{x+0.2}, \ x \in [0,1] \tag{2.25}$$

where $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 0.5$. The regression matrix $M^{(k)}$ for kNN regression is determined by $M^{(k)}_{ij} = \frac{1}{k}$ if $j \in \mathcal{N}_k(x_i)$ and 0 else. Then, the loss rank is

$$\mathrm{LR}(k) = \inf_{\alpha \geq 0} \{ \frac{n}{2} \log(\mathbf{y}^\top S^{(k)}_\alpha \mathbf{y}) - \frac{1}{2} \log\det S^{(k)}_\alpha \},$$

where $S_\alpha^{(k)} = (\mathbb{1} - M^{(k)})^\top (\mathbb{1} - M^{(k)}) + \alpha \mathbb{1}$. The most widely-used method for selecting k is probably the generalized cross-validation (GCV) criterion [Craven and Wahba, 1979]: $\mathrm{GCV}(k) = n\|(\mathbb{1} - M^{(k)})\mathbf{y}\|^2 / [\mathrm{tr}(\mathbb{1} - M^{(k)})]^2$. To judge how well GCV and LoRP work, we compare them to an "ideal" criterion based on the expected prediction error defined as

$$\mathrm{EPE}(k) \;=\; \sum_{i=1}^{n} \mathbf{E}(y_i - \hat{y}_i)^2 \;=\; \sum_{i=1}^{n} \Big[\sigma^2 + (f(x_i) - \frac{1}{k} \sum_{j \in \mathcal{N}_k(x_i)} f(x_j))^2 + \frac{\sigma^2}{k}\Big].$$

Of course this criterion is not available in real data applications. Figure 2.1(a) shows the curves $\mathrm{LR}(k)$, $\mathrm{GCV}(k)$, $\mathrm{EPE}(k)$ for $k = 2,...,20$ (the trivial case $k = 1$ is omitted), in which $k = 7$-nearest neighbors is chosen by LoRP and $k = 8$ is chosen by GCV. The "ideal" k is 5. Both LoRP and GCV do a reasonable job. LoRP works slightly better than GCV in this particular simulated data set.

Repeating the experiment 50 times, we find that LoRP *always* select a smaller k than GCV. The averaged k over 50 values selected by LoRP (by GCV) is 7.1 (7.4, respectively). In comparison with the "ideal" $k = 5$, this simulation study suggests that LoRP works slightly better than GCV.

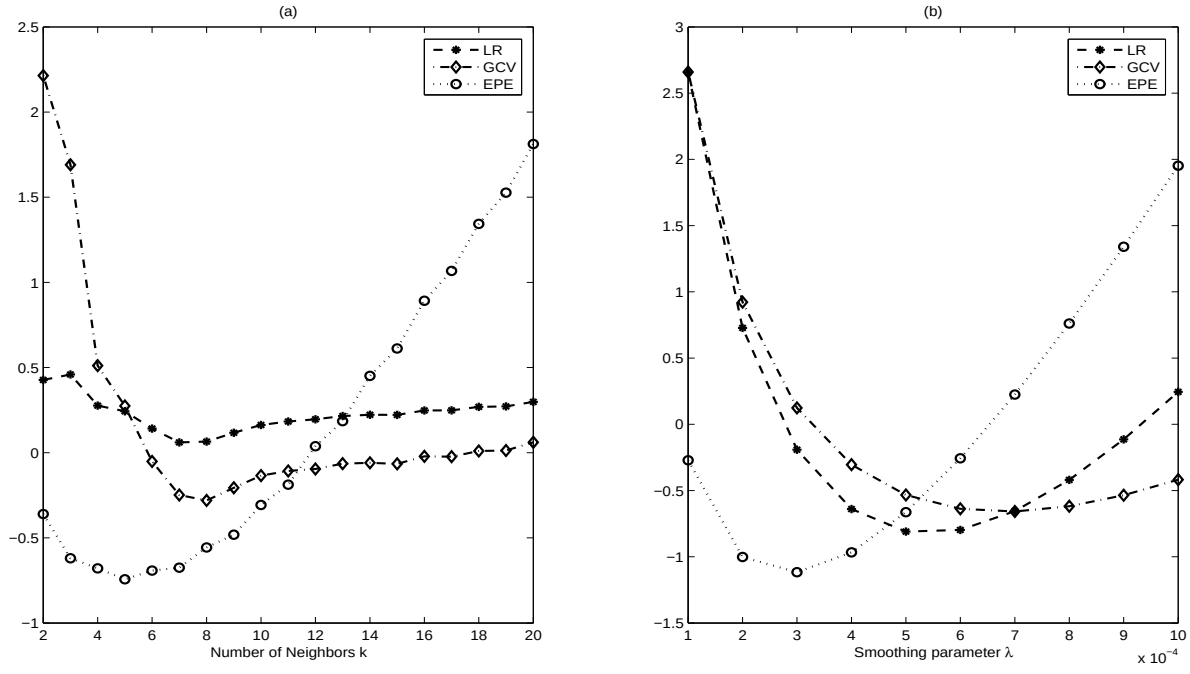
2.5.4 Selection of smoothing parameter

We now further demonstrate the use of the LoRP in selecting a good smoothing parameter for spline regression. Consider the following problem: find a function belonging to the class of functions with continuous 2nd derivative that minimizes the following penalized residual sum of squares:

$$\mathrm{RSS}(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt,$$

where λ is called the smoothing parameter. The second term penalizes the curvature of function f and the smoothing parameter λ controls the amount of penalty. Our goal is to choose a good λ.

Figure 2.1: Choosing the tuning parameters in kNN and spline regression. The curves have been scaled by their standard deviations. Plotted are loss rank (LR), generalized cross-validation (GCV) and expected prediction error (EPE).



It is well-known (see, e.g., Hastie et al. [2001], Section 5.4) that the solution is a natural spline $f(x) = \sum_{j=1}^{n} N_j(x)\theta_j$ where $N_1(x)$,...,$N_n(x)$ are the basis functions of the natural cubic spline:

$$N_1(x) = 1,\ N_2(x) = x,\ N_{k+2}(x) = d_k(x) - d_{n-1}(x) \ \text{ with } \ d_k(x) = \frac{(x - x_k)_+^3 - (x - x_n)_+^3}{x_n - x_k}.$$

The problem thus reduces to finding a vector $\theta \in I\!R^n$ that minimizes

$$\mathrm{RSS}(\theta) = (\mathbf{y} - N\theta)^\top (\mathbf{y} - N\theta) + \lambda \theta^\top \Omega \theta$$

where $N_{ij} = N_j(x_i)$ and $\Omega_{ij} = \int N_i''(x) N_j''(x) dx$. It is easy to see that the solution is $\hat{\theta}_\lambda = (N^\top N + \lambda\Omega)^{-1} N^\top \mathbf{y}$, and the fitted vector is $\hat{\mathbf{y}} = N\hat{\theta}_\lambda = M_\lambda \mathbf{y}$ with $M_\lambda = N(N^\top N + \lambda\Omega)^{-1} N^\top \mathbf{y}$. The fitted vector is linear in $\mathbf{y}$, thus the loss rank is

$$\mathrm{LR}(\lambda) = \arg\min_{\alpha \geq 0} \{ \frac{n}{2} \log(\mathbf{y}^\top S_\lambda^\alpha \mathbf{y}) - \frac{1}{2} \log\det S_\lambda^\alpha \}$$

where $S_\lambda^\alpha = (\mathbb{1} - M_\lambda)^\top (\mathbb{1} - M_\lambda) + \alpha \mathbb{1}$.

Let us consider again the dataset generated from model (2.25). Figure 2.1(b) shows the curves LR(λ), GCV(λ) and EPE(λ). The derivation of expressions for GCV(λ) and EPE(λ) is similar to the previous example. $\lambda \approx 3 \times 10^{-4}$ is the optimal value selected by the "ideal" criterion EPE. $\lambda \approx 5 \times 10^{-4}$ and $\lambda \approx 7 \times 10^{-4}$ are selected by LoRP and GCV, respectively. Averaged λ over 20 replications are 5.1×10^{-4} for LoRP and 7.2×10^{-4} for GCV. Once again, like the previous example, LoRP seems to work better than GCV.

2.5.5 Model selection by loss rank for classification

We now demonstrate the LR criterion for model selection in classification, developed in Section 2.4, by a simple example of a piecewise constant classifier and compare it to the model selection criterion based on Rademacher complexities. Consider the intervals model

selection problem which was described by Fromont [2007] (see also, Lozano [2000], Bartlett et al. [2002]). Given a number $N \in I\!N$, let $\mathcal{X} = \{1,2,...,2^N\}$. For $u,v \in I\!N, u \leq v$, denote by $I\!N[u,v]$ the set of integers in interval $[u,v]$. For an integer number m, $1 \leq m \leq N$, let

$$\mathcal{F}_m = \left\{ t : \mathcal{X} \to \{0,1\}, t = \sum_{k=1}^{2^m} c_k I_{I\!N[(k-1)2^{N-m}+1, k2^{N-m}]}, c_k \in \{0,1\}, k = 1, ...2^m \right\}$$

be the set of piecewise constant functions defined on $\mathcal{X}$ and taking on values $\{0,1\}$ with possible jumps at $k2^{N-m}$, $k = 1,...,2^m - 1$ (two functions of this kind are plotted in Figure 2.2).

For a given m_0, $1 \leq m_0 \leq N$, let S_0 be the set of odd-numbered segments:

$$S_0 = \bigcup_{k=2p+1,\ p=0,1,...,2^{m_0-1}-1} I\!N[(k-1)2^{N-m_0} + 1, k2^{N-m_0}].$$

Let X be a uniformly distributed random variable on $\mathcal{X}$ and Y be a $\{0,1\}$-valued random variable defined as

$$\mathrm{P}(Y = 1 | X \in S_0) = \frac{1}{2} + h, \text{ and } \mathrm{P}(Y = 1 | X \notin S_0) = \frac{1}{2} - h$$

where $h \in (1, \frac{1}{2})$ is called the margin parameter. We now have a model selection problem with N candidate models $\{\mathcal{F}_m,\ m \in \mathcal{M} = \{1,...,N\}\}$ and the optimal predictor $s(x) = I_{S_0}(x) \in \mathcal{F}_{m_0}$ belongs to one of them. See Figure 2.2 for plots of $s(x)$ and observations. We are interested in identifying the true model m_0. The advantage of the intervals model selection problem is that it is very easy to compute for each $m \in \mathcal{M}$

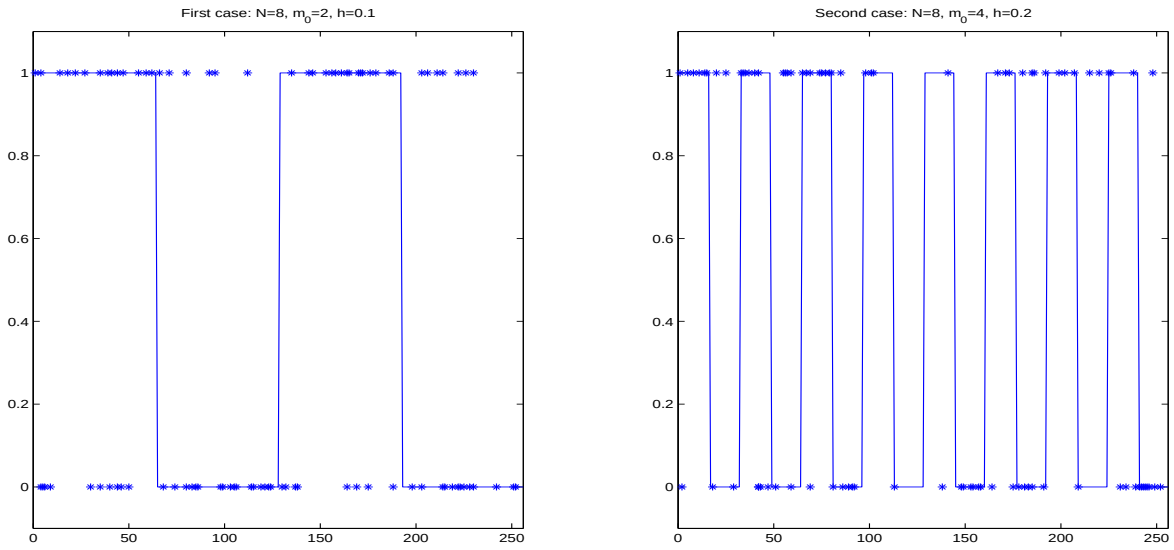
$$\mathrm{P}_n\gamma(\hat{t}_m) = \inf_{t \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^{n} I_{Y_i \neq t(X_i)} \text{ and } \sup_{t \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^{n} r_i I_{Y_i \neq t(X_i)}.$$

The reader is referred to Fromont [2007] for the details.

We compare the LR criterion to another criterion based on Rademacher complexities which is taken following Fromont [2007] to be

$$\mathrm{crit}_{\mathrm{RC}}(m) = \mathrm{P}_n\gamma(\hat{t}_m) + \mathrm{pen}_{\mathrm{RC}}(m) \text{ with } \mathrm{pen}_{\mathrm{RC}}(m) = E(\sup_{t \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^{n} r_i I_{Y_i \neq t(X_i)} | D).$$

Figure 2.2: Plots of the true functions and data for two cases.



We shall call this the Rademacher complexity (RC) criterion. In our experiment, the loss rank $\mathrm{LR}_n(m)$ and Rademacher complexities $\mathrm{pen}_{\mathrm{RC}}(m)$ are estimated by $B = 200$ Monte Carlo simulations.

Figure 2.2 plots true functions and observation data (with $n = 100$) for two cases: first with $N = 8$, $m_0 = 2$, $h = .1$, then $N = 8$, $m_0 = 4$, $h = .2$. These pictures show how hard it is to decide intuitively what the true model is. Figure 2.3 plots the LR criterion and RC criterion curves. Both criteria identify the true model in both cases.

Table 2.3 presents the proportions of correct identification over 100 replications for each of 16 cases with various sample sizes $n = 50$, 100, 200, 300 and noise levels $h = .05$, .1, .2, .3 ($m_0 = 4$). It suggests that both criteria are model selection consistent as the proportions increases to 1 as n and h increase. The simulation suggests that the LR criterion has a slight improvement over the RC criterion for large sample sizes.

Figure 2.3: Plots of the loss rank (LR) and Rademacher complexities (RC) vs complexity m.

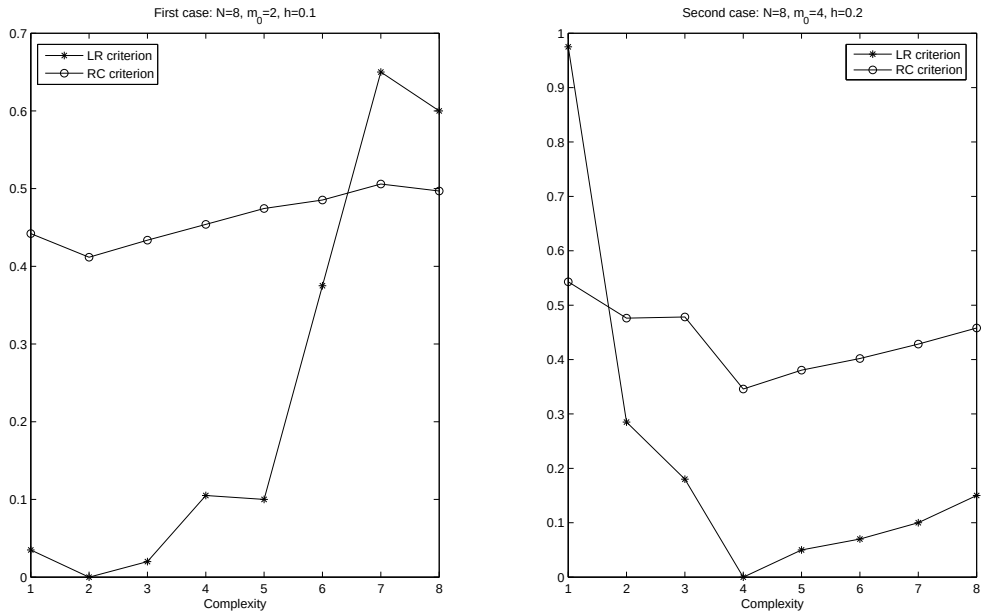


Table 2.3: Model selection by loss rank for classification: Proportions of correct identification of the loss rank (LR) and Redemacher complexities (RC) criteria for various n and h.

n	h	LR criterion	RC criterion	n	h	LR criterion	RC criterion
50	.05	.12	.13	200	.05	.23	.21
	.1	.35	.35		.1	.67	.66
	.2	.62	.64		.2	.99	.97
	.3	.95	.97		.3	1	1
100	.05	.15	.15	300	.05	.30	.28
	.1	.41	.41		.1	.78	.76
	.2	.89	.90		.2	1	.99
	.3	.98	.98		.3	1	1

2.6 Applications

We present in this section two well-studied applications of the LoRP, one is to selecting the ridge parameter in ridge regression and the other is to selecting shrinkage parameters in regularization procedures such as the Lasso for variable selection. Their full discussions can be found in Tran [2009, 2011b].

2.6.1 LoRP for choosing ridge parameter

Ridge regression

Consider the standard linear regression model

$$\mathbf{y} = X\beta + \epsilon \tag{2.26}$$

where $\mathbf{y}$ is an n-vector of responses, ϵ is an n-vector of noise, X is an $(n \times p)$ matrix standardized such that $X^\top X$ is in the form of a correlation matrix, $\mathbf{E}(\epsilon) = 0$, $\text{cov}(\epsilon) = \sigma^2 \mathbb{1}_n$, and $\beta = (\beta_1,...,\beta_p)^\top$ is the vector of regression coefficients. When X is full rank, it's well-known in the literature that the unbiased least squares estimator of β is $\hat{\beta}^{\text{OLS}} = (X^\top X)^{-1} X^\top \mathbf{y}$. When $X^\top X$ is nearly singular, however, the expected distance $\mathbf{E}\|\hat{\beta}^{\text{OLS}} - \beta\|^2 = \sigma^2 \text{tr}(X^\top X)^{-1}$ will be very large, and $\hat{\beta}^{\text{OLS}}$ is not stable (a small change in $\mathbf{y}$ may lead to a large change in $\hat{\beta}^{\text{OLS}}$ even in signs and some of its components may be extremely large in absolute value).

The fact that the OLS estimate $\hat{\beta}^{\text{OLS}}$ may explode when $X^\top X$ is ill-conditioned naturally leads to the idea of restricting coefficients β to a sphere by minimizing

$$\sum_{j=1}^{n} (y_j - \sum_{i=1}^{p} \beta_i x_{ji})^2, \quad \text{s.t.} \quad \sum_{i=1}^{p} \beta_i^2 \leq s, \tag{2.27}$$

where $s \geq 0$ is a complexity parameter of the model. This optimization problem is equivalent to the *penalized least square estimation*: minimizing w.r.t. β

$$\sum_{j=1}^{n} (y_j - \sum_{i=1}^{p} \beta_i x_{ji})^2 + \lambda \sum_{i=1}^{p} \beta_i^2, \tag{2.28}$$

where $\lambda > 0$ is called the *ridge parameter* that controls the amount of shrinkage of regression coefficients. There is a one-by-one correspondence between s and λ [Hastie et al., 2001, Chapter 3], an increase in s leads to a decrease in λ and otherwise.

The solution of (2.28) with a given λ is $\hat{\beta}(\lambda) = (X^\top X + \lambda \mathbb{1}_p)^{-1} X^\top \mathbf{y}$. This is often called the *ridge estimator*, and was originally introduced by Hoerl and Kennard [1970] in an attempt to deal with the ill-conditioned $X^\top X$. Although $\hat{\beta}(\lambda)$ is biased when $\lambda > 0$, there is a trade-off between the bias and the variance. Let $d_1^2 \geq ... \geq d_p^2$ be the eigenvalues of $X^\top X$, the expected distance between $\hat{\beta}(\lambda)$ and β [Hoerl and Kennard, 1970] is

$$\mathbf{E}\|\hat{\beta}(\lambda) - \beta\|^2 = \lambda^2 \beta^\top (X^\top X + \lambda \mathbb{1}_p)^{-2} \beta + \sigma^2 \sum_{1}^{p} \frac{d_i^2}{(d_i^2 + \lambda)^2}. \tag{2.29}$$

The first term is known as the squared bias, it equals 0 when $\lambda=0$, the second is the sum of variances $\mathrm{tr}[\mathrm{var}(\hat{\beta}(\lambda))]$. Hoerl and Kennard [1970] showed that there exists a $\lambda>0$ such that $\mathbf{E}\|\hat{\beta}(\lambda)-\beta\|^2<\mathbf{E}\|\hat{\beta}^{\mathrm{OLS}}-\beta\|^2$.

The remaining problem is how to choose a good ridge parameter. A large number of methods have been proposed: the ridge trace [Hoerl and Kennard, 1970], Hoerl-Kennard-Baldwin estimator (HKB) [Hoerl et al., 1975], PRESS, cross-validation and its variants [Allen, 1974, Stone, 1974, Geisser, 1975, Craven and Wahba, 1979, Golub et al., 1979], and the bootstrap [Delaney and Chatterjee, 1986].

In this section, based on the LoRP, we obtain a penalized maximum likelihood (PML) criterion for choosing λ. The criterion is of the form

$$-\sup(\text{log-likelihood}) + \text{ penalty of the complexity of model.}$$

This PML criterion can be considered as a "continuous" version of AIC whose penalty of the model complexity is the number of coefficients which is a discrete number. A simulation study is carried out to compare the suggested method to several competitors.

Penalized ML for choosing λ

Denote by M_λ the ridge regression model w.r.t. parameter λ, $\mathcal{M}=\{M_\lambda,\ \lambda>0\}$ is then the class of candidate models. The regression matrix w.r.t. model M_λ is $M_\lambda=X(X^\top X+\lambda\mathbb{1}_p)^{-1}X^\top$ (we use the same notations for both model and its regression matrix). The fitted vector $\hat{\mathbf{y}}_\lambda=M_\lambda\mathbf{y}$ is linear in $\mathbf{y}$ so Theorem 7 can be applied. The matrix S_α in (2.8) now is

$$S_\alpha = S_\alpha(\lambda) = (\mathbb{1}_n - M_\lambda)^\top(\mathbb{1}_n - M_\lambda) + \alpha\mathbb{1}_n.$$

Consider the singular value decomposition (SVD) of X, $X=UDV$, where U is an $(n\times n)$ orthogonal matrix, V is a $(p\times p)$ orthogonal matrix, D is an $(n\times p)$ diagonal matrix with

principal diagonal elements $d_1 \geq ... \geq d_p \geq 0$. By using the SVD of X, it is easy to see that the eigenvalues of S_α are

$$\alpha + (\frac{\lambda}{d_1^2+\lambda})^2, ..., \alpha + (\frac{\lambda}{d_p^2+\lambda})^2, 1+\alpha, ..., 1+\alpha.$$

Suppose at the moment that $\alpha = O_{\mathrm{P}}(1/n)$ (this will be justified later on, here α may be a random variable), where $a_n = O_{\mathrm{P}}(b_n)$ means random variables $|a_n/b_n| \leq C$ with some bounded constant C as $n \to \infty$ with probability 1. Then we get with probability 1 (w.p.1) that

$$\begin{aligned}
\det S_\alpha^\lambda &= (1+\alpha)^{n-p} \prod_{i=1}^{p} (\alpha + (\frac{\lambda}{d_i^2+\lambda})^2) \\
&= (1+\alpha)^{n-p} \left[\prod_{i=1}^{p} (1+\alpha(\frac{d_i^2+\lambda}{\lambda})^2)\right] \left[\prod_{i=1}^{p} (\frac{\lambda}{d_i^2+\lambda})^2\right] \\
&\approx [1+(n-p)\alpha] \left[1+\alpha \sum_1^p (\frac{d_i^2+\lambda}{\lambda})^2\right] \left[\prod_{i=1}^{p} (\frac{\lambda}{d_i^2+\lambda})^2\right] \\
&\approx \left[1+\alpha(n-p+\sum_1^p (\frac{d_i^2+\lambda}{\lambda})^2)\right] \left[\prod_{i=1}^{p} (\frac{\lambda}{d_i^2+\lambda})^2\right] \\
&= [1+\alpha\nu] \prod_{i=1}^{p} (\frac{\lambda}{d_i^2+\lambda})^2 \text{ where } \nu := n-p+\sum_1^p (\frac{d_i^2+\lambda}{\lambda})^2 = n + \frac{2p}{\lambda} + \frac{1}{\lambda^2}\sum_1^p d_i^4.
\end{aligned}$$

Let $\rho_\lambda = \|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2 / \|\mathbf{y}\|^2$, (2.9) becomes

$$\begin{aligned}
\mathrm{LR}_{M_\lambda}^\alpha(D) &= \frac{n}{2}\log(\mathbf{y}^\top S_\alpha^\lambda \mathbf{y}) - \frac{1}{2}\log\det S_\alpha^\lambda \\
&\approx \frac{n}{2}\log\|\mathbf{y}\|^2 + \frac{n}{2}\log(\rho_\lambda + \alpha) - \frac{1}{2}\log(1+\alpha\nu) - \frac{1}{2}\log[\prod_{i=1}^{p}(\frac{\lambda}{d_i^2+\lambda})^2].
\end{aligned}$$

Solving $\partial \mathrm{LR}_\lambda^\alpha(D)/\partial\alpha = 0$ with respect to α, we get a minimum at

$$\alpha = \alpha_m = \frac{\nu\rho_\lambda - n}{(n-1)\nu} \text{ provided } \nu\rho_\lambda > n \text{ w.p.1.} \tag{2.30}$$

ρ_λ can be considered as a measure of fit. Clearly, in the case of overfitting, ρ_λ will be very close to 0. The main point of LoRP is to avoid overfitting. Thus, it is reasonable to consider only λ such that ρ_λ is not so small in the sense of the following condition

Condition (C): $\nu\rho_\lambda = (n + \frac{2p}{\lambda} + \frac{1}{\lambda^2}\sum_1^p d_i^4)\rho_\lambda > n$ w.p.1.

Our experience to date shows that this condition is mostly satisfied in practice. Under the condition (C), $\alpha_m = O_{\mathrm{P}}(1/n)$ that justifies the assumption above about α. We then also get $\alpha_m/\rho_\lambda = O_{\mathrm{P}}(1/n)$ which leads to

$$\begin{aligned} \frac{n}{2}\log(\rho_\lambda + \alpha_m) &= \frac{n}{2}[\log\rho_\lambda + \frac{\alpha_m}{\rho_\lambda} - \frac{1}{2}(\frac{\alpha_m}{\rho_\lambda})^2 + \frac{1}{3}(\frac{\alpha_m}{\rho_\lambda})^3 + ...] \\ &= \frac{n}{2}\log\rho_\lambda + \frac{1}{2} + o_{\mathrm{P}}(1) \end{aligned}$$

where $a_n = o_{\mathrm{P}}(1)$ means $|a_n| \to 0$ as $n \to \infty$ w.p.1. Combine the last equalities and neglect the constants independent of model M_λ, we can finally write the loss rank of model M_λ as

$$\mathrm{LR}_\lambda(D) \equiv \mathrm{LR}_{M_\lambda}^{\alpha_m}(D) = \frac{n}{2}\log(\|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2) - \frac{1}{2}\log\big[\frac{\nu\rho_\lambda - 1}{n-1}\prod_{i=1}^{p}(\frac{\lambda}{d_i^2 + \lambda})^2\big]. \tag{2.31}$$

Assume now that the noise ϵ is Gaussian $N(0,\sigma^2\mathbb{1}_n)$, the log-likelihood of the observations from model (2.26) (neglecting constant $-\frac{n}{2}\log(2\pi)$) then is

$$l_n(\beta,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y} - X\beta\|^2.$$

Because of the equivalence between (2.27) and (2.28), the set $\Theta_\lambda = \{\theta = (\beta_1,...,\beta_p,\sigma^2) : \|\beta\|^2 \le s, \sigma^2 > 0\}$ (note that $s = s(\lambda)$ as there is a correspondence between s and λ) can be seen as the parameter space of regression model M_λ, thus, since ϵ is Gaussian

$$\sup_{\theta\in\Theta_\lambda} l_n(\beta,\sigma^2) = -\frac{n}{2}\log(\frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2) - \frac{n}{2} = -\frac{n}{2}\log(\|\mathbf{y} - \hat{\mathbf{y}}_\lambda\|^2) + \frac{n}{2}\log n - \frac{n}{2}.$$

Neglecting the constant term which is independent of model M_λ, (2.31) can be written as

$$\sup_{\theta\in\Theta_\lambda} l_n(\beta,\sigma^2) + \sum_1^p \log(1 + \frac{d_i^2}{\lambda}) - \frac{1}{2}\log\frac{\nu\rho_\lambda - 1}{n-1} \tag{2.32}$$

which has the form of a penalized maximum likelihood criterion

$$-\sup(\text{log-likelihood}) + \text{“penalty of the model complexity”}$$

where the penalty term is

$$\text{pen}(n,\lambda) = \sum_{1}^{p} \log(1+\frac{d_i^2}{\lambda}) - \frac{1}{2}\log\frac{\nu\rho_\lambda - 1}{n-1}. \tag{2.33}$$

Define $c{=}c(\lambda){=}1/\lambda$. We can see that c is a measure of the complexity of model M_λ: larger c (i.e., smaller λ) leads to bigger Θ_λ, thus M_λ is more complex/flexible; and otherwise. Noting that when n is sufficiently large

$$\text{pen}(n,\lambda) \approx \sum_{1}^{p} \log(1+\frac{d_i^2}{\lambda}) - \frac{1}{2}\log\rho_\lambda$$

that ρ_λ increases as λ increases, and that $\rho_\lambda \uparrow 1$ as $\lambda \uparrow \infty$, we have w.p.1 that (i) pen(n,λ) is an increasing monotone function of the complexity c, and (ii) pen(n,λ)$\to 0$ as $c \to 0$ and pen(n,λ)$\to\infty$ as $c\to\infty$. Therefore pen(n,λ) has the usual properties of a penalty function [Chambaz, 2006]. This penalty function depends on ρ_λ, so it is data-dependent. It has been widely criticised that PML criteria based on distribution-free penalties may sometimes work poorly for some specific distributions. PML based on data-dependent penalties may give better performance over based on distribution-free penalties.

A simulation study

We now conduct a systematic simulation study to evaluate the performance of the suggested criterion for choosing λ and compare it to other competitors including GCV [Golub et al., 1979]

$$\text{GCV}(\lambda) = \frac{1}{n}\|(\mathbb{1}_n - M_\lambda)\mathbf{y}\|^2/[\frac{1}{n}\text{tr}(\mathbb{1}_n - M_\lambda)]^2,$$

HKB estimator [Hoerl et al., 1975]

$$\lambda_{\text{HKB}} = ps^2/\|\hat\beta(0)\|^2, \quad s^2 = \|\mathbf{y} - X\hat\beta^{\text{OLS}}\|^2/(n-p)$$

and the ordinary least square (OLS). The HKB is introduced by the authors of the original papers on ridge regression, while GCV is the most widely used method.

Two factors that affect the ridge regression the most are degree of correlation between explanatory variables and signal-to-noise ratio (SNR). The degree of correlation is often measured by the *condition number* [Belsley et al., 1980] defined as $d_1/d_p \geq 1$ where $d_1 \geq ... \geq d_p > 0$ are singular values of design matrix X. The larger the condition number, the stronger the dependencies between explanatory variables. SNR is defined as $\|\beta\|^2/\sigma^2$.

In our study, four levels of correlation (very weak, weak, strong and very strong) w.r.t. condition numbers 5, 10, 50 and 100 (according to Belsley et al. [1980]) are studied. We considerer three levels of SNR: 1, 10 and 100 which can be considered as large, medium and small errors, respectively. Therefore 12 ridge regression models which represent various situations we would face in the real world are studied. For each model, a design matrix of size (50×4) and a response vector are generated. To search for the optimal ridge parameters, 1000 values of λ ranging from 0.001 to 1 in increments of .001 are used.

The performance of the methods is measured in terms of the average MSE in regression coefficients. For each of the 12 regression models, 100 replications are generated, the MSEs and the chosen λ's are taken average over the 100 replications. For a method δ, its average MSE is computed by

$$\text{MSE}(\delta) = \frac{1}{100} \sum_{j=1}^{100} \|\beta^{(j)} - \hat{\beta}^{(j)}(\delta)\|$$

where $\beta^{(j)}$ is the true coefficients of j-th replication and $\hat{\beta}^{(j)}(\delta)$ is the ridge estimator of $\beta^{(j)}$ with λ is chosen by method δ. Along with the average MSE(δ), the standard deviations sd(δ) are also computed.

Table 2.4 presents the average and standard deviation of MSE's over 100 replications for each of the 12 ridge regression models. The numbers in brackets are the means and standard deviations of selected λ's. LR outperforms the others, especially when there are at least weak dependencies (i.e., the condition number ≥ 10) between the explanatory

Table 2.4: LoRP for choosing ridge parameter in comparison with GCV, Hoerl-Kennard-Baldwin (HKB) estimator and ordinary least squares (OLS): Average MSE over 100 replications for various signal-to-noise ratio (SNR) and condition number (CN). Numbers in brackets are means and standard deviations of selected λ's.

SNR	CN	LR	GCV	HKB	OLS
	5	1.95±0.54	2.36±1.51	2.26±1.31	3.18±2.15
		(0.80±0.21)	(0.39±0.31)	(0.21±0.17)	(0)
	10	1.94±0.68	2.77±2.66	2.87±2.23	6.05±4.06
1		(0.79±0.21)	(0.38±.32)	(0.13±0.15)	(0)
	50	2.06±0.88	6.52±11.52	10.12±13.91	29.31±23.37
		(0.81±0.21)	(0.36±0.32)	(0.05±0.14)	(0)
	100	2.09±0.72	4.95±9.41	17.86±22.99	58.41±39.57
		(0.83±0.18)	(0.38±0.31)	(0.02±0.08)	(0)
	5	1.24±0.61	0.99±0.58	0.95±0.57	1.01±0.60
		(0.20±0.13)	(0.05±0.06)	(0.03±0.01)	(0)
	10	1.57±0.88	1.61±0.97	1.71±0.87	1.94±1.25
10		(0.20±0.12)	(0.05±0.07)	(0.04±0.01)	(0)
	50	1.44±0.95	3.47±4.03	4.26±5.61	9.91±8.38
		(0.21±0.14)	(0.04±0.08)	(0.01±0.01)	(0)
	100	1.42±0.83	2.95±2.96	6.27±8.06	18.85±13.39
		(0.20±0.13)	(0.03±0.07)	(0.01±0.01)	(0)
	5	0.49±0.31	0.32±0.20	0.32±0.20	0.31±0.19
		(0.04±0.01)	(0.001±0.003)	(0.003±0.001)	(0)
	10	1.328±0.88	1.327±0.95	1.40±0.95	2.02±1.37
100		(0.05±0.01)	(0.006±0.005)	(0.002±0.001)	(0)
	50	1.371±0.92	1.47±0.96	1.66±1.12	2.78±2.28
		(0.06±0.03)	(0.007±0.006)	(0.002±0.002)	(0)
	100	1.45±0.91	1.59±1.19	2.69±3.13	6.02±4.78
		(0.05±0.03)	(0.005±0.004)	(0.001±0.001)	(0)

variables. Also, as the condition number increases, the performance of LR increases, while that of GCV and HKB decreases.

We use the method of comparing means of two paired samples (see, e.g., [Rice, 1995, Chapter 11]) to test the hypothesis H_0:LR$=\delta$ (i.e., the overall average MSE of method δ is the same as that of LR) against the alternative H_1:LR$>\delta$ (i.e., LR is better than δ, or the overall average MSE of δ is larger than that of LR), where δ is each of the methods GCV, HKB and OLS. Table 2.5 shows the P-values of the tests, in which the P-values smaller than 0.01 are rounded down to 0. As shown, when there are dependencies between the explanatory variables, most of the P-values are smaller than significance level 0.05. Thus, we can conclude that the improvement of LR over the others is statistically significant. In general, we can rank the performance of the criteria as: LR$>$GCV$>$HKB$>$OLS. In summary, the simulation results strongly support the use of LR.

2.6.2 LoRP for choosing regularization parameters

The Lasso [Tibshirani, 1996] and other regularization procedures such as the SCAD [Fan and Li, 2001] are attractive methods for variable selection, subject to a proper choice of shrinkage parameter. We obtain in this section from the LoRP a criterion for choosing shrinkage parameters for variable selection purposes.

Let us consider the problem of variable selection in linear regression analysis. We consider the case where a large number (even larger than the sample size) of candidate covariates are introduced at the initial stage of modeling. One then has to select a smaller subset of the covariates to fit/interpret the data. If the number of potential covariates is not so large (as small as 30), one may use subset selection to select significant variables (Section 2.3). However, with a large number of covariates, searching on model space is computationally infeasible. Regularization procedures are successful methods to overcome

Table 2.5: P-values for testing LR$=\delta$/LR$>\delta$

SNR	CN	LR>GCV	LR>HKB	LR>OLS
1	5	0	0.01	0
	10	0	0	0
	50	0	0	0
	100	0	0	0
10	5	1	1	0.99
	10	0.35	0.02	0.01
	50	0	0	0
	100	0	0	0
100	5	1	1	1
	10	0.50	0.27	0
	50	0.14	0	0
	100	0.03	0	0

this problem. A Lasso-type procedure estimates the regression coefficient vector β by minimizing the sum of the squared error and a regularization term

$$\|\mathbf{y} - X\beta\|^2 + \lambda T(\beta), \tag{2.34}$$

where X is an $(n\times d)$ non-random design matrix, $\mathbf{y}$ is an $n-$vector of responses, and $\lambda\geq 0$ is a shrinkage parameter that controls the amount of regularization. The regularization function $T(\beta)$ can take different forms according to different regularization procedures. The original and most popular one used in the Lasso is the l_1 norm $T(\beta) = \sum_{j=1}^{d}|\beta_j|$. As λ increases, the coefficients are continuously shrunk towards 0. When λ is sufficiently large, some coefficients are shrunk to exact 0, thus leading to sparse solutions. This feature makes the Lasso-type procedures very attractive for variable selection. Indeed, their model

selection consistency has been shown [Zhao and Yu, 2006, Meinshausen and Buhlmann, 2006, Fan and Li, 2001]: Under some conditions, there exists a "proper" sequence of shrinkage parameters $\{\lambda_n\}$ under which

$$\{j : \hat{\beta}_j^{\lambda_n} \neq 0\} = \mathcal{S}_T \text{ w.p.1 when sample size } n \text{ is large enough,} \tag{2.35}$$

where $\hat{\beta}_{\lambda_n} = (\hat{\beta}_1^{\lambda_n}, ..., \hat{\beta}_d^{\lambda_n})^\top$ is the regularized estimator of β with shrinkage parameter λ_n, and $\mathcal{S}_T$ is the true model, i.e., $\mathcal{S}_T$ is the index set of true covariates. Therefore, it is convenient to use the Lasso-type procedures for variable selection purposes.

The remaining problem in practice is how to choose such proper λ_n. A widely-used criterion is the generalized cross-validation criterion (GCV) [Craven and Wahba, 1979, Tibshirani, 1996]. However, theoretical properties of GCV for choosing λ for the purpose of variable selection have not been investigated yet. Furthermore, for choosing shrinkage parameter for the SCAD method [Fan and Li, 2001], a regularization method closely related to the Lasso, GCV seems to be likely to choose shrinkage parameters that produce overfitted models [Wang et al., 2007]. Zou et al. [2007] showed that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of the Lasso. As a result, popular model selection criteria - like AIC, BIC and C_p - can be used for selecting λ. However, theoretical properties of the selected model remain unknown. We obtain in this section a criterion for selecting shrinkage parameters in order for regularization procedures to produce the true model.

Although regularization procedures can be used for simultaneous variable selection and estimation, it seems to be impossible to tune the shrinkage parameter to achieve both model selection consistency and optimal estimation at the same time. For an orthogonal design, Leng et al. [2006] showed that the Lasso estimator that is optimal in terms of estimation does not give consistent model selection. This fact was also shown by Poetscher and Leeb

[2009] for other regularized estimators. We are primarily concerned with the problem of variable selection, i.e., we use a Lasso-type procedure to produce a set of potential subsets and then select the best one among this preselected set using a model selection criterion. The preselected set consists of at most d subsets rather than 2^d possible subsets if using subset selection. After selecting the best subset, we of course can use an unpenalized procedure to estimate the coefficients in order to reduce estimation bias.

We shall derive from the LoRP a criterion, called the loss rank (LR) criterion, for selecting shrinkage parameters for variable selection purposes. As long as the regularization procedure in use has the consistency property (2.35), the shrinkage parameter selected by the LR criterion will produce the true model asymptotically with probability 1. This model selection consistency of the proposed criterion will be proven theoretically in the case where the number of covariates d is fixed and smaller than n. For cases with $d \gg n$, our simulation study suggests that this property still holds. The simulation also shows that our method for variable selection works surprisingly well. Benefiting from fast l_1-regularization algorithms, our method is able to correctly identify significant variables from thousands of candidates in several CPU seconds.

The LR criterion

Let $\hat{\beta}_\lambda = (\hat{\beta}_1^\lambda, ..., \hat{\beta}_d^\lambda)^\top$ be the regularized estimator of β w.r.t. a certain shrinkage parameter λ, i.e., $\hat{\beta}_\lambda$ is the solution of (2.34). Denote by $\mathcal{S}_\lambda = \{j : \hat{\beta}_j^\lambda \neq 0\}$ the index set corresponding to the non-zero coefficients, by $\mathrm{df}_\lambda = |\mathcal{S}_\lambda|$ the number of non-zero coefficients, and by $X_{\mathcal{S}_\lambda}$ the design matrix corresponding to the selected covariates. We assume at the moment that $\mathrm{df}_\lambda \leq n$ and further assume that matrices $X_{\mathcal{S}_\lambda}$ are full rank. The case where $\mathrm{df}_\lambda > n$ will be dealt with later on.

Fitting model $\mathcal{S}_\lambda$ by least squares, we denote the OLS estimator and the variance

estimator by

$$\hat{\beta}_{\mathcal{S}_\lambda} = (X_{\mathcal{S}_\lambda}^\top X_{\mathcal{S}_\lambda})^{-1} X_{\mathcal{S}_\lambda}^\top \mathbf{y} \; ; \; \hat{\sigma}_{\mathcal{S}_\lambda}^2 = \frac{1}{n}\|\mathbf{y} - X_{\mathcal{S}_\lambda}\hat{\beta}_{\mathcal{S}_\lambda}\|^2,$$

respectively. The fitted vector under model $\mathcal{S}_\lambda$

$$\hat{\mathbf{y}}_{\mathcal{S}_\lambda} = X_{\mathcal{S}_\lambda}\hat{\beta}_{\mathcal{S}_\lambda} = M_{\mathcal{S}_\lambda}\mathbf{y} \text{ with } M_{\mathcal{S}_\lambda} := X_{\mathcal{S}_\lambda}(X_{\mathcal{S}_\lambda}^\top X_{\mathcal{S}_\lambda})^{-1} X_{\mathcal{S}_\lambda}^\top$$

is, conditionally on $\mathcal{S}_\lambda$, linear[2] in $\mathbf{y}$. Then from (2.9), the loss rank of model $\mathcal{S}_\lambda$ with parameter α is

$$\mathrm{LR}_\lambda^\alpha \equiv \mathrm{LR}_{\mathcal{S}_\lambda}^\alpha = \frac{n}{2}\log(\mathbf{y}^\top S_\alpha^\lambda \mathbf{y}) - \frac{1}{2}\log\det(S_\alpha^\lambda)$$

where $S_\alpha^\lambda = (\mathbb{1} - M_{\mathcal{S}_\lambda})^\top(\mathbb{1} - M_{\mathcal{S}_\lambda}) + \alpha\mathbb{1} = (1+\alpha)\mathbb{1} - M_{\mathcal{S}_\lambda}$. Because projection matrix $M_{\mathcal{S}_\lambda}$ has df_λ eigenvalues 1 and $n - \mathrm{df}_\lambda$ eigenvalues 0, S_α^λ has df_λ eigenvalues α and $n - \mathrm{df}_\lambda$ eigenvalues $1+\alpha$. Thus, $\det S_\alpha^\lambda = \alpha^{\mathrm{df}_\lambda}(1+\alpha)^{n-\mathrm{df}_\lambda}$. Let $\rho_\lambda := \|\mathbf{y} - \hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2 / \|\mathbf{y}\|^2$, we have

$$\mathrm{LR}_\lambda^\alpha = \frac{n}{2}\log \mathbf{y}^\top\mathbf{y} + \frac{n}{2}\log(\rho_\lambda + \alpha) - \frac{\mathrm{df}_\lambda}{2}\log\alpha - \frac{n - \mathrm{df}_\lambda}{2}\log(1+\alpha).$$

Taking derivative w.r.t α, it is easy to see that $\mathrm{LR}_\lambda^\alpha$ is minimized at $\alpha_m = \frac{\rho_\lambda \mathrm{df}_\lambda}{(1-\rho_\lambda)n - \mathrm{df}_\lambda}$ provided that $1 - \rho_\lambda > \mathrm{df}_\lambda / n$. This condition is ensured by Assumption (A3) below. Finally, after some algebra, the loss rank of model $\mathcal{S}_\lambda$ can be explicitly expressed as

$$\mathrm{LR}_\lambda = \mathrm{LR}_\lambda^{\alpha_m} = \frac{n}{2}\log\|\mathbf{y}\|^2 - \frac{n}{2}\mathrm{KL}(\frac{\mathrm{df}_\lambda}{n}\|1 - \rho_\lambda). \tag{2.36}$$

where $\mathrm{KL}(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the Kullback-Leibler divergence between the Bernoulli distributions with parameters $p, q \in (0,1)$. The optimal shrinkage parameter(s) λ (for variable selection purposes) chosen by the LR criterion will be

$$\hat{\lambda}_{\mathrm{LR}} \in \mathrm{argmin}_{\lambda \geq 0}\mathrm{LR}_\lambda = \mathrm{argmax}_{\lambda \geq 0}\mathrm{KL}(\frac{\mathrm{df}_\lambda}{n}\|1 - \rho_\lambda). \tag{2.37}$$

[2]Strictly speaking, $\hat{\mathbf{y}}_{\mathcal{S}_\lambda}$ is not linear in $\mathbf{y}$ because $\mathcal{S}_\lambda$ depends on $\mathbf{y}$. However, we can consider preselected subsets $\mathcal{S}_\lambda$ as fixed models. If instead we first derive the LR criterion for a general fixed model $\mathcal{S}$ and then apply to $\mathcal{S}_\lambda$, we get the same results.

Often, LR_λ reaches its minimum in an interval $(\hat{\lambda}_l, \hat{\lambda}_u)$ (see Figure 2.4). Any λ in this interval produces the same model. This can be explained as follows. When λ increases from 0 to infinity, the number of non-zero coefficients of $\hat{\beta}_\lambda$ will be a non-increasing step function of λ [Efron et al., 2004]; in other words, the covariates are in turn removed from the models. As a result, by its definition LR_λ is also a step function. Note that our emphasis is on variable selection rather than on coefficient estimation.

Optimality property

In order to prove the model selection consistency of the LR criterion, we assume in this section that d is fixed and $d \leq n$. We need the following assumptions

(A1) There exists a deterministic sequence of reference shrinkage parameters λ_n such that $\mathcal{S}_{\lambda_n} \to \mathcal{S}_T$ w.p.1.

(A2) ϵ is Gaussian $N(0, \mathbb{1}_n)$.

(A3) For each candidate λ, ρ_λ is bounded away from 0 and 1, i.e., there are constants c_1, c_2 such that $0 < c_1 \leq \rho_\lambda \leq c_2 < 1$ w.p.1.

$\rho_\lambda = \|\mathbf{y} - \hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2 / \|\mathbf{y}\|^2$ is a measure of fit. In extreme cases where the resulting model $\mathcal{S}_\lambda$ is too big or too small, ρ_λ will be close to 0 and 1, respectively. Therefore, it is reasonable to consider only λ in which ρ_λ is bounded away from 0 and 1. Note that for every $\mathcal{S}_\lambda$ we have that

$$\rho_\lambda = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2 + \|\hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2} = \frac{\hat{\sigma}^2_{\mathcal{S}_\lambda}}{\hat{\sigma}^2_{\mathcal{S}_\lambda} + \|\hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2/n}.$$

For λ such that $\mathcal{S}_\lambda$ is the true model $\mathcal{S}_T$, (A3) follows from a mild sufficient condition

$$0 < \liminf_{n\to\infty} (\frac{1}{n}\|\hat{\mathbf{y}}_{\mathcal{S}_T}\|^2) \leq \limsup_{n\to\infty} (\frac{1}{n}\|\hat{\mathbf{y}}_{\mathcal{S}_T}\|^2) < \infty \ \text{ and } \ \hat{\sigma}^2_{\mathcal{S}_T} \to \sigma^2 > 0 \ \text{ w.p.1}$$

where $\hat{\mathbf{y}}_{\mathcal{S}_T}$ is the fitted vector under the true model. Moreover, if the intercept is included in the models, we have that $n(\bar{\mathbf{y}})^2 \leq \|\hat{\mathbf{y}}_{\mathcal{S}_\lambda}\|^2 \leq \|\mathbf{y}\|^2$. (A3) then follows from a very mild condition

$$0 < \lim\inf_{n\to\infty} (\bar{\mathbf{y}})^2 \leq \lim\sup_{n\to\infty} (\frac{1}{n}\|\mathbf{y}\|^2) < \infty \text{ and } \hat{\sigma}^2_{\mathcal{S}} \to \text{constant} > 0 \; \forall \mathcal{S} \text{ w.p.1.}$$

Assumption (A1) is satisfied by some regularization procedures, for example, Lasso [Zhao and Yu, 2006] and SCAD [Fan and Li, 2001]. Normality assumption (A2) is not a necessary condition for consistency. This assumption can be relaxed, but then a more complicated proof technique is needed.

We have the following lemma which is similar to Lemma 9.

Lemma 13. *Under Assumption (A3), the loss rank* LR_λ *has the form*

$$\mathrm{LR}_\lambda \;=\; \frac{n}{2}\log\hat{\sigma}^2_{\mathcal{S}_\lambda} + \frac{\mathrm{df}_\lambda}{2}\log n + O_{\mathrm{P}}(1), \tag{2.38}$$

where $O_{\mathrm{P}}(1)$ *denotes a bounded random variable w.p.1.*

The above lemma is used to prove model selection consistency of the LR criterion.

Theorem 14 (Model selection consistency of the LR criterion). *Assume that* d *is fixed. Under Assumptions (A1)-(A3), the shrinkage parameter selected by the LR criterion will produce the true model w.p.1 when* n *is large enough, i.e.,*

$$\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}_T) \to 1$$

where $\hat{\lambda}_{\mathrm{LR}}$ *is determined in* (2.37).

The idea of the proof is to bound the probabilities of picking under- and overfitted models. A model $\mathcal{S}$ is said to be underfitted if $\mathcal{S}$ misses at least one true covariate (i.e., $\mathcal{S} \not\supseteq \mathcal{S}_T$), overfitted if $\mathcal{S}$ contains all true covariates and at least one untrue (i.e., $\mathcal{S} \supsetneq \mathcal{S}_T$).

There is a finite number of such $\mathcal{S}$, so it is sufficient to prove that $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) \to 0$ for each of them. The detailed proof is relegated to Section 2.7.

We can of course use other model selection criteria such as AIC or BIC rather than LoRP for choosing the best subset among the preselected set produced by the regularization procedure. AIC is asymptotically optimal in terms of loss efficiency but likely to select overfitted models, while BIC is asymptotically optimal in terms of model selection consistency; see Shao [1997], Yang [2005]. Therefore one may use BIC as another stopping rule besides LoRP. The shrinkage parameter chosen by BIC will be

$$\hat{\lambda}_{\mathrm{BIC}} \in \mathrm{argmin}_{\lambda \geq 0} \mathrm{BIC}_\lambda \ \text{ where } \ \mathrm{BIC}_\lambda := \frac{n}{2} \log \hat{\sigma}^2_{\mathcal{S}_\lambda} + \frac{\mathrm{df}_\lambda}{2} \log n. \tag{2.39}$$

We see from Lemma 13 that, up to a constant, the LR criterion is asymptotically equivalent to BIC. It follows from the proof of Theorem 14 that using BIC also leads to the same model selection consistency, i.e., $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{BIC}}} = \mathcal{S}_T) \to 1$ as $n \to \infty$. However, finite-sample simulation studies below show that the LR criterion works better than BIC, especially when $d \gg n$.

High-dimensional variable selection problems in which $d \gg n$ are currently of great interest to scientists. In order for such a problem to be solvable, an essential assumption needed is that it is d^*-sparse [Candes and Tao, 2007], i.e., the number of true covariates d^* must be smaller than n. Under this solvability assumption, it is clear that we can safely ignore irrelevant cases in which the number of covariates df_λ under consideration is larger than n. Then the LR criterion (2.36) is still valid. In practice, therefore, we propose to ignore those λ under which $\mathrm{df}_\lambda > n$ and apply the LR criterion as usual. A theoretically rigorous treatment is beyond the scope of the thesis, which we intend to do in future research. However, a systematic simulation study below suggests that the LR criterion still works surprisingly well and enjoys model selection consistency.

Numerical examples

We present now a simulation study for the LR criterion, compare it to other methods, and also apply it to a real data set. The regularization procedure we use is the Lasso. The Lasso solution paths are computed by the LARS algorithm of Efron et al. [2004]. A widely-used method for choosing the Lasso parameter is GCV [Craven and Wahba, 1979, Tibshirani, 1996]

$$\mathrm{GCV}_\lambda = \frac{1}{n}\frac{\|\mathbf{y} - X\hat{\beta}_\lambda\|^2}{(1 - \frac{1}{n}\mathrm{DF}_\lambda)^2}$$

where $\mathrm{DF}_\lambda := \mathrm{tr}[X(X^\top X + \lambda W^-)^{-1}X^\top \mathbf{y}]$, $W = \mathrm{diag}(|\hat{\beta}_j^\lambda|)$ and W^- is a generalized inverse of W. Another one is the BIC-type criterion of Wang et al. [2007] (although its variable selection consistency requires the oracle property, a property not enjoyed by the Lasso)

$$\widetilde{\mathrm{BIC}}_\lambda = \log\frac{\|\mathbf{y} - X\hat{\beta}_\lambda\|^2}{n} + \mathrm{DF}_\lambda\frac{\log n}{n}.$$

Note that $\hat{\beta}_\lambda \neq \hat{\beta}_{\mathcal{S}_\lambda}$. The former is the Lasso estimator whereas the latter is the OLS estimator resulting from fitting model $\mathcal{S}_\lambda$ by least squares. Our proposed criteria (2.36) and (2.39) are constructed based on $\hat{\beta}_{\mathcal{S}_\lambda}$, not $\hat{\beta}_\lambda$. This is the essential difference between our approach and the others.

We consider the following example which is taken from Tibshirani [1996]:

$$y = x^\top\beta + \sigma\epsilon \tag{2.40}$$

where $\beta = (3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0)^\top$, x_i are marginally $N(0,1)$ with the correlation between x_i and x_j equal to $0.5^{|i-j|}$, $\epsilon \sim N(0,1)$. We compare the performance of LR and BIC criterion to that of GCV and $\widetilde{\mathrm{BIC}}$. The performance is measured by the frequency of underfitting, overfitting and correct fitting and average number of zero coefficients over 100 replications.

Table 2.6 summarizes the simulation results for various factors n and σ. Although $\widetilde{\mathrm{BIC}}$ works slightly better than GCV, it still produces overfitted models most of the time. BIC

does a good job and LR outperforms the others.

Table 2.6: LoRP for choosing regularization parameters: small-d case

σ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%)	Ave. No. of zeros
1	100	GCV	0	0	100	1.57
		$\widetilde{\text{BIC}}$	0	3	97	2.32
		BIC	0	89	11	4.88
		LR	0	97	3	4.97
	200	GCV	0	0	100	1.64
		$\widetilde{\text{BIC}}$	0	0	100	1.81
		BIC	0	94	6	4.93
		LR	0	100	0	5
3	100	GCV	0	0	100	1.34
		$\widetilde{\text{BIC}}$	0	0	100	1.53
		BIC	1	70	29	4.22
		LR	1	77	22	4.37
	200	GCV	0	0	100	1.69
		$\widetilde{\text{BIC}}$	0	0	100	2.09
		BIC	0	91	9	4.89
		LR	0	91	9	4.90

We now consider cases of large d with $d=300$ and $n=100$, 200, 500. We set up a *sparse recovery problem* in which most of the coefficients are zero except $\beta_{30}=\beta_{60}=...=\beta_{300}=10$. Table 2.7 summarizes the simulation results for various factors $n=100$, 200, 500 and $\sigma=1$, 3. The LR criterion works surprisingly well in comparison with BIC and the others.

Let us take a closer look at the simulation results in Tables 2.6 and 2.7. Although

the LR and BIC criteria are *asymptotically* equivalent to each other, the finite-sample simulation study shows that the LR criterion works better than BIC. A similar situation was also observed in Section 2.5 for subset selection. This is probably because, contrary to the BIC criterion, the penalty term of the LR criterion is data-adaptive. Some results in the model selection literature show that selection criteria with data-adaptive penalties are more encouraging in terms of performance than those with deterministic penalties; see Yang [2005] and references therein. We see that BIC seems to break down for the cases $d > n$ as it always produces overfitted models, but starts working well when $n > d$. The $O_{\mathrm{P}}(1)$ term in (2.38) plays an important role here: it serves as a "corrector" to BIC. Note that BIC is just an approximation to the logarithm of posterior model probability [Schwarz, 1978], the approximation might be inaccurate if n is not large enough relative to d.

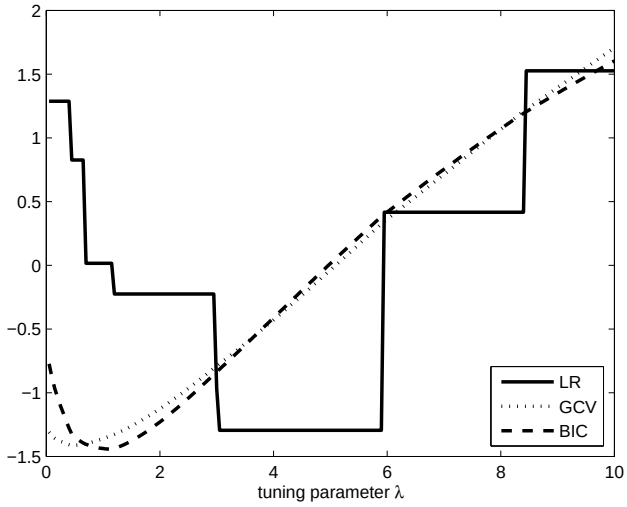
As another example, we consider a real data set. Stamey et al. [1989] studied the correlation between the level of prostate antigen (`lpsa`) and a number of clinical measures in men: log cancer volume (`lcavol`), log prostate weight (`lweight`), `age`, log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), log of capsular penetration (`lcp`), Gleason score (`gleason`), and percentage of Gleason scores 4 or 5 (`pgg45`). Following Tibshirani [1996], we assume a linear regression model between the response `lpsa` and the 8 covariates. We want to select a parsimonious model for the sake of scientific insight into the response-covariate relationship.

The data set of size 97 is standardized so that the intercept β_0 is excluded. Figure 2.4 presents the curves GCV_λ, $\widetilde{\mathrm{BIC}}_\lambda$, LR_λ (1000 values of λ ranging from 0.01 to 10 in increments of .01 were used to search for the optimal λ). The λ selected by GCV, $\widetilde{\mathrm{BIC}}$ are .5 and 1.1, and the corresponding models are {1, 2, 3, 4, 5, 7, 8}, {1, 2, 3, 4, 5, 8}, respectively. The LR criterion is minimized in the interval (3.1,5.9). Any value in this interval produces

Table 2.7: LoRP for choosing regularization parameters: large-d case

σ	n	Method	Under-fitted(%)	Correctly fitted(%)	Overfitted(%)	Ave. No. of zeros
1	100	GCV	0	0	100	90.20
		$\widetilde{\text{BIC}}$	0	0	100	95.8
		BIC	0	0	100	202.01
		LR	0	30	70	288.24
	200	GCV	0	0	100	87.51
		$\widetilde{\text{BIC}}$	0	0	100	89.45
		BIC	0	0	100	102.02
		LR	0	86	14	289.83
	500	GCV	0	0	100	97.51
		$\widetilde{\text{BIC}}$	0	0	100	104.45
		BIC	0	40	60	287.30
		LR	0	100	0	290
3	100	GCV	0	0	100	78.35
		$\widetilde{\text{BIC}}$	0	0	100	87.40
		BIC	0	0	100	202.04
		LR	0	18	82	287.51
	200	GCV	0	0	100	92.02
		$\widetilde{\text{BIC}}$	0	0	100	96.51
		BIC	0	0	100	102.01
		LR	0	58	42	289.29
	500	GCV	0	0	100	93.31
		$\widetilde{\text{BIC}}$	0	0	100	96.52
		BIC	0	35	65	288.35
		LR	0	80	20	289.75

Figure 2.4: Prostate cancer data: LR_λ, $\widetilde{\mathrm{BIC}}_\lambda$ and GCV_λ.



the same model $\mathcal{S}_{\mathrm{LR}} = \{1,\ 2,\ 5\}$. The BIC of these models are -19.20, -21.38, -25.19, respectively. That means the BIC also supports the choice of the LR criterion. (Note however that this does not mean that the BIC is an optimal criterion).

2.7 Proofs

Proof of Lemma 9. Inserting $\mathbf{y}^\top\mathbf{y} = n\hat{\sigma}^2_{\mathcal{S}}/\rho_{\mathcal{S}}$ into (2.12) and rearranging terms gives (2.15). By Assumption (A) the last term in (2.15) is bounded w.p.1. Taylor expansion $\log(1-p) = -p+O(p^2)$ implies $H(p)/p+\log p \to 1$, hence $\frac{n}{2}H(\frac{|\mathcal{S}|}{n}) = \frac{|\mathcal{S}|}{2}\log n + O(1)$. Finally, dropping the $\mathcal{S}$-independent term $\frac{n}{2}\log n$ from (2.15) gives (2.16). ■

Proof of Theorem 12. By $Y_i' := \frac{1+r_i}{2} - r_i Y_i$, it's easy to see that $I_{Y_i' \neq t(X_i)} = I_{r_i=1} - r_i I_{Y_i \neq t(X_i)}$, therefore

$$\inf_t \frac{1}{n}\sum_1^n I_{Y_i' \neq t(X_i)} = \frac{1}{n}\sum_1^n I_{r_i=1} - \sup_t \frac{1}{n}\sum_1^n r_i I_{Y_i \neq t(X_i)}. \tag{2.41}$$

Moreover,

$$\frac{1}{n}\sum_{1}^{n} r_i I_{Y_i \neq t(X_i)} = \frac{1}{n}\sum_{1}^{n} I_{Y_i \neq t(X_i)} - \frac{1}{n}\sum_{1}^{n}(1-r_i) I_{Y_i \neq t(X_i)} = \mathrm{P}_n\gamma(t) - \mathrm{P}_n^R\gamma(t) \qquad (2.42)$$

where $\mathrm{P}_n^R := \frac{1}{n}\sum W_i \delta_{(X_i,Y_i)}$ with $W_i := 1 - r_i \sim 2\mathrm{Binomial}(1,1/2)$ is the *weighted bootstrap empirical measure*. From (2.41)-(2.42) and (2.22), we have

$$\mathrm{LR}_n(m) = \mathrm{P}_R\Big(\sup_{t\in\mathcal{F}_m} (\mathrm{P}_n - \mathrm{P}_n^R)\gamma(t) \geq \frac{1}{n}\sum_{1}^{n} I_{r_i=1} - \mathrm{P}_n\gamma(\hat{t}_m) \big| D \Big).$$

The key point in the proof is the result of weak convergence of the weighted bootstrap empirical processes. The result states that, under Assumption (C), the difference between the conditional law of $\mathrm{P}_n - \mathrm{P}_n^R$ given data D and the law of $\mathrm{P} - \mathrm{P}_n$ converges to zero almost surely [van der Vaart and Wellner, 1996, p.346]. More formally, let $\hat{G}_n = \mathrm{P}_n - \mathrm{P}_n^R$ and $G_n = \mathrm{P} - \mathrm{P}_n$, and let $l^\infty(\mathcal{D}_m)$ be the space of all bounded functions from $\mathcal{D}_m$ to the real set $\mathbb{R}$ ($\hat{G}_n$ and G_n are random elements in $l^\infty(\mathcal{D}_m)$). Then

$$|\mathbf{E}_R h(\hat{G}_n) - \mathbf{E} h(G_n)| \to 0, \quad \mathrm{P} - \text{almost surely}$$

for every continuous, bounded function $h : l^\infty(\mathcal{D}_m) \to \mathbb{R}$.

Therefore, by the continuous mapping theorem with notice that $\frac{1}{n}\sum_1^n I_{r_i=1} \to 1/2$ a.s., we have P-almost surely

$$\Big| \mathrm{P}_R\Big(\sup_{t\in\mathcal{F}_m} (\mathrm{P}_n - \mathrm{P}_n^R)\gamma(t) \geq \tfrac{1}{n}\textstyle\sum_1^n I_{r_i=1} - \mathrm{P}_n\gamma(\hat{t}_m) \big| D \Big) \\ - \mathrm{P}\Big(\sup_{t\in\mathcal{F}_m} (\mathrm{P} - \mathrm{P}_n)\gamma(t) \geq \tfrac{1}{2} - \mathrm{P}_n\gamma(\hat{t}_m) \Big) \Big| \to 0.$$

Thus, as n is sufficiently large

$$\mathrm{LR}_n(m) = \mathrm{P}\left(\sup_{t\in\mathcal{F}_m} (\mathrm{P} - \mathrm{P}_n)\gamma(t) \geq \frac{1}{2} - \mathrm{P}_n\gamma(\hat{t}_m) \right) = \mathrm{P}(\mathrm{crit}_n(m) \geq \frac{1}{2}) \ \text{ w.p.1.}$$

For simplicity, suppose that $\mathrm{LR}_n(m)$ has a unique minimum at $\hat{m}_{\mathrm{LR}}$. If $\hat{m}_{\mathrm{LR}} \neq m_n$, $\mathrm{P}(\mathrm{crit}_n(m_n) \geq \frac{1}{2}) > \mathrm{P}(\mathrm{crit}_n(\hat{m}_{\mathrm{LR}}) \geq \frac{1}{2})$. On the other hand, $\mathrm{crit}_n(m_n) < \mathrm{crit}_n(\hat{m}_{\mathrm{LR}})$ by

the definition of m_n, so $\mathrm{P}(\mathrm{crit}_n(m_n) \geq \frac{1}{2}) \leq \mathrm{P}(\mathrm{crit}_n(\hat{m}_{\mathrm{LR}}) \geq \frac{1}{2})$. The contradiction implies $\hat{m}_{\mathrm{LR}} = m_n$ w.p.1. ■

Proof of Theorem 14. The main idea of the proof is taken from Chambaz [2006]. Let us denote by $\mathbf{z}_i = (x_{i1},...,x_{id},y_i)$ the i-th observation and by $\gamma(.,m,\sigma^2)$ the density of the Gaussian distribution with mean m and variance σ^2. Under model $\mathcal{S}$, the density of $\mathbf{z}_i$ is $p_{\theta_\mathcal{S}}(\mathbf{z}_i) = \gamma(y_i, \sum_{j\in\mathcal{S}} \beta_j x_{ij}, \sigma^2)$. The log-likelihood is

$$l_n(\theta_\mathcal{S}) = \sum_{i=1}^{n} \log p_{\theta_\mathcal{S}}(\mathbf{z}_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j\in\mathcal{S}}\beta_j x_{ij})^2.$$

It is easy to see that

$$\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) = -\frac{n}{2}\log\hat{\sigma}^2_\mathcal{S} - \frac{n}{2}(1+\log(2\pi)).$$

By (2.38), the loss rank of model $\mathcal{S}_\lambda$ now can be written as

$$\mathrm{LR}_\lambda = -\sup_{\theta\in\Theta(\mathcal{S}_\lambda)} l_n(\theta) + \frac{\mathrm{df}_\lambda}{2}\log n + C(n) + O_{\mathrm{P}}(1)$$

where the constant term $C(n) = \frac{n}{2}\log n - \frac{n}{2}(1+\log(2\pi))$ is independent of $\mathcal{S}_\lambda$.

No underestimation. It is sufficient to prove that $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) \to 0$ for each $\mathcal{S} \not\supseteq \mathcal{S}_T$, as there is only a finite number of such $\mathcal{S}$.

$$\begin{aligned}
\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) &= \mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}, \mathrm{LR}_{\hat{\lambda}_{\mathrm{LR}}} \leq \mathrm{LR}_{\lambda_n}) \\
&= \mathrm{P}\Big(\frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}})} l_n(\theta) - \frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \geq \frac{\log n}{2n}(\mathrm{df}_{\hat{\lambda}_{\mathrm{LR}}} - \mathrm{df}_{\lambda_n}) + o_{\mathrm{P}}(1),\ \mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}\Big) \\
&\leq \mathrm{P}\Big(\frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \geq \frac{\log n}{2n}(|\mathcal{S}| - \mathrm{df}_{\lambda_n}) + o_{\mathrm{P}}(1)\Big) \\
&\leq \mathrm{P}\Big(\frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S}_T)} l_n(\theta) \geq \frac{\log n}{2n}(|\mathcal{S}| - d^*) + o_{\mathrm{P}}(1)\Big) + P(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \\
&\leq \mathrm{P}\Big(\frac{1}{n}\sup_{\theta\in\Theta(\mathcal{S})} l_n(\theta) - \frac{1}{n} l_n(\theta^*) \geq \frac{\log n}{2n}(|\mathcal{S}| - d^*) + o_{\mathrm{P}}(1)\Big) + P(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \qquad (2.43)
\end{aligned}$$

where $\theta^* \in \mathcal{S}_T$ denotes the true parameter. By the law of large numbers for the supremum of the likelihood ratios (see, e.g., Lemma B1 of Chambaz [2006])

$$\frac{1}{n} \sup_{\theta \in \Theta(\mathcal{S})} l_n(\theta) - \frac{1}{n} l_n(\theta^*) \to - \inf_{\theta \in \Theta(\mathcal{S})} \mathrm{KL}(p_{\theta^*} \| p_\theta) \text{ w.p.1.}$$

Because $\mathcal{S} \not\supseteq \mathcal{S}_T$, $\inf_{\theta \in \Theta(\mathcal{S})} \mathrm{KL}(p_{\theta^*} \| p_\theta) > 0$. This, together with the fact that $\frac{\log n}{2n}(|\mathcal{S}| - d^*) \to 0$ and Assumption (A1), shows that the left-hand side term of (2.43) goes to 0 as $n \to \infty$.

No overestimation. Fix an overfitted model $\mathcal{S} \supsetneq \mathcal{S}_T$, let us denote by

$$H(\theta) := \mathrm{KL}(p_{\theta^*} \| p_\theta) = E[\frac{1}{n}(l_n(\theta^*) - l_n(\theta))] \geq 0 \ \forall \theta \in \Theta(\mathcal{S})$$

($H(\theta)$ is not necessarily positive) and $h_n(\theta) := \frac{l_n(\theta) - l_n(\theta^*)}{H(\theta)^{1/2}}$ with convention $\frac{0}{0} = 0$. For every $\theta \in \Theta(\mathcal{S})$

$$\begin{aligned} l_n(\theta) - l_n(\theta^*) + nH(\theta) &= l_n(\theta) - l_n(\theta^*) - E[l_n(\theta) - l_n(\theta^*)] \\ &= H(\theta)^{1/2}(h_n(\theta) - Eh_n(\theta)) \\ &\leq H(\theta)^{1/2} \sup_{\nu \in \Theta(\mathcal{S})} (h_n(\nu) - Eh_n(\nu)). \end{aligned} \tag{2.44}$$

By $\Theta(\mathcal{S}_T) \subset \Theta(\mathcal{S})$ and the property of supremum, for every $\epsilon > 0$ there exists $\theta_0 \in \Theta(\mathcal{S})$ such that

$$\sup_{\theta \in \Theta(\mathcal{S})} (l_n(\theta) - l_n(\theta^*)) \leq l_n(\theta_0) - l_n(\theta^*) + \epsilon \tag{2.45}$$

and also

$$l_n(\theta_0) - l_n(\theta^*) \geq 0. \tag{2.46}$$

From (2.45) and (2.44)

$$\sup_{\theta \in \Theta(\mathcal{S})} (l_n(\theta) - l_n(\theta^*)) \leq H(\theta_0)^{1/2} \sup_{\theta \in \Theta(\mathcal{S})} (h_n(\theta) - Eh_n(\theta)) + \epsilon. \tag{2.47}$$

From (2.46) and (2.44)

$$nH(\theta_0) \leq l_n(\theta_0) - l_n(\theta^*) + nH(\theta_0) \leq H(\theta_0)^{1/2} \sup_{\theta \in \Theta(\mathcal{S})} (h_n(\theta) - Eh_n(\theta))$$

or

$$nH(\theta_0)^{1/2} \leq \sup_{\theta\in\Theta(\mathcal{S})} (h_n(\theta) - Eh_n(\theta)). \tag{2.48}$$

Now, since $\epsilon > 0$ was chosen arbitrarily, (2.47) and (2.48) yield

$$\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_T)} l_n(\theta) \leq \sup_{\Theta(\mathcal{S})} \{l_n(\theta) - l_n(\theta^*)\} \leq \frac{1}{n} \left(\sup_{\theta\in\Theta(\mathcal{S})} (h_n(\theta) - Eh_n(\theta)) \right)^2. \tag{2.49}$$

We need the following bounded law of the iterated logarithm which is a consequence of Theorem 4.1, Dudley and Philipp [1983] or Lemma B2, Chambaz [2006].

Lemma 15. *There is a finite constant C so that*

$$\limsup_n \frac{\sup_{\theta\in\Theta(\mathcal{S})} |h_n(\theta) - Eh_n(\theta)|}{\sqrt{n \log\log n}} \leq C \quad w.p.1.$$

Now for every overfitted model $\mathcal{S} \supsetneq \mathcal{S}_T$, it is sufficient to prove that $\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) \to 0$. In fact,

$$\begin{aligned}
&\mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S}) = \mathrm{P}(\mathcal{S}_{\hat{\lambda}_{\mathrm{LR}}} = \mathcal{S},\ \mathrm{LR}_{\hat{\lambda}_{\mathrm{LR}}} \leq \mathrm{LR}_{\lambda_n}) \\
\leq\ & \mathrm{P}\left(\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_{\lambda_n})} l_n(\theta) \geq \frac{\log n}{2}(|\mathcal{S}| - \mathrm{df}_{\lambda_n}) + O_{\mathrm{P}}(1) \right) \\
\leq\ & \mathrm{P}\left(\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_T)} l_n(\theta) \geq \frac{\log n}{2}(|\mathcal{S}| - d^*) + O_{\mathrm{P}}(1) \right) + \mathrm{P}(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \\
=\ & \mathrm{P}\left(\left\lceil \frac{\log\log n}{\frac{d^*}{2}\log n} \right\rceil \left\lceil \frac{\sup_{\Theta(\mathcal{S})} l_n(\theta) - \sup_{\Theta(\mathcal{S}_T)} l_n(\theta)}{\log\log n} \right\rceil \geq \frac{|\mathcal{S}|}{d^*} - 1 + o_{\mathrm{P}}(1) \right) + \mathrm{P}(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \\
\leq\ & \mathrm{P}\left(\left\lceil \frac{\log\log n}{\frac{d^*}{2}\log n} \right\rceil \left\lceil \frac{\sup_{\Theta(\mathcal{S})} |h_n(\theta) - Eh_n(\theta)|}{\sqrt{n\log\log n}} \right\rceil^2 \geq \frac{|\mathcal{S}|}{d^*} - 1 + o_{\mathrm{P}}(1) \right) + \mathrm{P}(\mathcal{S}_{\lambda_n} \neq \mathcal{S}_T) \qquad (2.50)
\end{aligned}$$

where the last inequality follows from (2.49). Observe that $|\mathcal{S}| > d^*$ as $\mathcal{S} \supsetneq \mathcal{S}_T$. This, together with Lemma 15 and the fact that $\log\log n/(\frac{d^*}{2}\log n) \to 0$, implies that the first probability of (2.50) goes to zero. The second probability of (2.50) also goes to zero because of Assumption (A1). This completes the proof. ■

Chapter 3

Predictive model selection

As discussed in the introduction chapter, model selection has many different goals. We developed in the previous chapter a principle for model selection where the main motivation is to learn the underlying structure in data. In this chapter, we approach the problem from a different angle: consider the problem of model selection with an explicit predictive motivation. In other words, the primary goal in model selection now is to select useful models for predicting well future observations.

We present in Section 3.1 a procedure for optimal predictive model selection. Section 3.2 discusses a regularization version of this procedure for variable selection in generalized linear models, which leads to the proposal of a predictive version of the Lasso. The materials presented in this chapter have been published in Tran [2011a] and Tran et al. [2010].

3.1 A procedure for optimal predictive model selection

Let D be a given set of past observations and $\mathcal{M}=\{M_k,k\in\mathcal{K}\}$ be a set of candidate models from which we want to select a useful one in order to predict future observations. Denote by Δ a future observation. We assume that there exists a known (up to a normalization constant) predictive distribution $p(\Delta|D)$ which is the best one in terms of making predictions on future data but for some reasons (discussed later) should not be used. We will refer to $p(\Delta|D)$ as a reference distribution and give now two typical examples of it. The basic idea of the model selection method proposed in this section is applicable to both Bayesian and frequentists, but we will mainly take a Bayesian approach in this chapter.

Consider the model selection problem from a variable selection point of view, and consider the case in which it is believed that every covariate should have a nonzero but probably small coefficient. Then, from the Bayesian perspective, it is sometimes argued that the posterior predictive distribution based on the full model with a carefully elicited prior should be used to achieve the best prediction accuracy [Aitchison, 1975, Geisser, 1993], and that ignoring any covariate may lose some information for predicting the response. Another example of the reference distribution $p(\Delta|D)$ is the predictive distribution based on Bayesian model averaging (BMA) which has some optimalities for prediction, and works very well empirically (see Leamer [1978], Draper [1995], Raftery et al. [1997], Hoeting et al. [1999] and references therein).

Although the full model and BMA often have predictive optimalities, there are some reasons that may preclude their use. Their main drawback is *non-interpretability*. The full model does not tell us (clearly or in an easily accessible way) which and how predictors affect the response, while BMA does not produce an easily interpretable model because

it averages over all candidate models. For many reasons, analysts often prefer simple models. For example, given a large number of potential covariates, we would commonly like to select a smaller subset that predicts future responses as well as possible. This would give a model that can interpret which and how covariates affect the response. This drawback is somewhat similar to that of ridge regression (see Section 2.6.1). Although ridge regression produces stable estimates of coefficients and often has the optimal mean squared error, it does not give an interpretable model. In contrast to ridge regression, the Lasso (see Section 2.6.2) shrinks some coefficients to exact 0, so it produces an easily interpretable model. That is why the Lasso is somewhat prefered to ridge regression.

Another drawback of using the full model or BMA is that if there is a cost associated with data collection then it would be inadvisable to use all of the predictors or models.

There are some desirable properties that any statistical procedure should satisfy: prediction accuracy, simplicity (or parsimony) and interpretability. Our motivation is to look for a model that has all these desirable properties. The aim is to choose a single model that is interpretable (thus simple) and has the best predictive performance over any other single model that may have been reasonably selected. The idea is to trade-off between prediction accuracy and interpretability. To this end, we use a distance function to measure distances between the reference distribution $p(\Delta|D)$ and the predictive distributions $p(\Delta|D,M)$ of candidate models $M \in \mathcal{M}$, and seek a model that has the smallest distance. Then, the chosen model has obviously better predictive performance than any other single model - besides, it is *a model* (rather than a combination of models), thus it is interpretable. In addition, the predictive ability of the chosen model is similar to that of the reference distribution $p(\Delta|D)$. We will refer to this Procedure for Optimal Predictive MOdel Selection as POPMOS.

The POPMOS shall be fully described in Section 3.1.1, its implementation is discussed

in Section 3.1.2 and its application to a real data set is given in Section 3.1.5. We present in Section 3.1.3 two popular measures of predictive performance and introduce in Section 3.1.4 a model uncertainty indicator.

3.1.1 Setup of the POPMOS

Let $\int d(p,q)dx$ be a distance function that measures the distance (or pseudo-distance) between two density functions p and q (for simplicity, we assume that the Lebesgue measure is used, however the following procedure can be constructed similarly for the general case). We define the distance between the reference predictive distribution $p(\Delta|D)$ and the predictive distribution $p(\Delta|D,M_k)$ under model M_k by

$$\delta(M_k) \equiv \delta(M_k, D, d(.,.)) = \int d(p(\Delta|D), p(\Delta|D, M_k))d\Delta. \tag{3.1}$$

If a single model is preferred, it is natural to seek a model M_k that has the predictive distribution $p(\Delta|D,M_k)$ closest to $p(\Delta|D)$. Formally, the optimal predictive (OP) model (among a given collection of model $\mathcal{M}$) is determined as

$$\hat{M}_{\text{OP}} = \text{argmin}_{M_k \in \mathcal{M}} \delta(M_k). \tag{3.2}$$

This setup of the POPMOS is general enough to apply to various frameworks where the collection of single models involves linear regression models, generalized linear models, Cox models, graphical models, etc.

For the distance function, we will consider the Kullback-Leibler (KL) distance where $d(p,q) = p\log(p/q)$. The KL distance is widely used in statistics and information theory to measure the (pseudo) distance between two density functions and was used to derive two well-known model selection rules, namely AIC [Akaike, 1973] and MDL [Rissanen, 1978]. Besides, many other distance functions can be used as well. Some of them are the

Hellinger distance where $d(p,q)=(\sqrt{p}-\sqrt{q})^2$ and f-divergence where $d(p,q)=f(p/q)q$ for a convex function f such that $f(1)=0$. Using the KL distance, (3.1) becomes

$$\delta_{\mathrm{KL}}(M_k) = \mathbf{E}\left[\log \frac{p(\Delta|D)}{p(\Delta|D, M_k)}\right] \tag{3.3}$$

where the expectation is w.r.t. $p(\Delta|D)$.

3.1.2 Implementation of the POPMOS

We discuss here an implementation of the POPMOS in the general case where the BMA predictive distribution is used as the reference distribution $p(\Delta|D)$. The case of predictive variable selection in GLMs where the full model is used as the reference will be discussed in Section 3.2.

Let $p(M_k)$ be the prior probability of model $M_k \in \mathcal{M}$, $p(\theta_k|M_k)$ be the prior distribution of model parameter θ_k under model M_k. Then, the BMA predictive distribution of a future observation Δ is given by

$$p(\Delta|D) = \sum_{M_k \in \mathcal{M}} p(\Delta|M_k, D)p(M_k|D). \tag{3.4}$$

In this expression, $p(M_k|D)$ is the posterior probability of model M_k

$$p(M_k|D) = \frac{p(D|M_k)p(M_k)}{\sum_{M_l \in \mathcal{M}} p(D|M_l)p(M_l)}, \tag{3.5}$$

where

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \tag{3.6}$$

is the marginal likelihood and

$$p(\Delta|M_k, D) = \int p(\Delta|\theta_k, M_k, D)p(\theta_k|M_k, D)d\theta_k \tag{3.7}$$

is the posterior predictive distribution of Δ under model M_k. Expression (3.4) is a weighted average of the posterior predictive distributions of Δ under each model, the weights being the posterior model probabilities.

There is a sense in which BMA provides better predictive performance than any single model [Madigan and Raftery, 1994, Draper, 1995, Raftery et al., 1997, Hoeting et al., 1999, Clyde and George, 2004].

However, the implementation of BMA (and thus of the POPMOS) is often a difficult task. Fortunately, by virtue of recent computational advances and computational methodologies like Markov chain Monte Carlo (MCMC) methods, the computational burden of the integrals in (3.3) is greatly reduced. We discuss below an approach for implementing the POPMOS using the Occam's window idea of Madigan and Raftery [1994] (see also Hoeting et al. [1999]) and an MCMC algorithm for estimating integrals. Other methods such as variational Bayes could be used also.

Occam's window principle

The number of competing models under consideration is often huge and precludes the calculation of all distances $\delta(M_k)$. It's natural that if a model gets very little support from the data (i.e., its posterior model probability $p(M_k|D)$ is very small), it should be excluded from consideration. This is the Occam's window idea of Madigan and Raftery [1994]. More formally, we only consider models belonging to

$$\mathcal{A} = \left\{ M_k \in \mathcal{M} : \frac{\max_{M_l \in \mathcal{M}} p(M_l|D)}{p(M_k|D)} \leq C \right\} \tag{3.8}$$

where the cutoff parameter C is chosen by the data analyst, $C=20$ being often used [Madigan and Raftery, 1994, Raftery et al., 1997]. Then the reference predictive distribution

$p(\Delta|D)$ in (3.4) is approximated by

$$p(\Delta|D) = \sum_{M_k \in \mathcal{A}} p(\Delta|M_k, D)p(M_k|D) \tag{3.9}$$

and (3.2) reduces to

$$\hat{M}_{\text{OP}} = \text{argmin}_{M_k \in \mathcal{A}} \delta(M_k). \tag{3.10}$$

In most cases, the number of models in $\mathcal{A}$ is greatly reduced to fewer than 50 and often fewer than 25. Note that once $\mathcal{A}$ is determined, the posterior model probabilities must be normalized (so that $\sum_{M_k \in \mathcal{A}} p(M_k|D) = 1$).

MCMC for distance calculation

MCMC methods provide a very efficient way to estimate complicated integrals. A good reference book on MCMC in practice is Gilks et al. [1996]. The Metropolis-Hasting algorithm for estimating $\delta_{\text{KL}}(M_k)$ is as follows (for simplicity, we assume that the components of Δ are continuous):

1. Initialize a Markov chain to Δ_0, set $t \leftarrow 0$.
2. Sample a candidate point Δ from a multivariate normal distribution with mean Δ_t and covariance matrix $\sigma^2 I_p$ where p is the dimension of Δ.
3. Sample a point u from a uniform distribution $U(0,1)$.
4. If $u \leq \min(1, \frac{p(\Delta|D)}{p(\Delta_t|D)})$ then set $\Delta_{t+1} \leftarrow \Delta$, else set $\Delta_{t+1} \leftarrow \Delta_t$.
5. Set $t \leftarrow t+1$ and go back step 2 until $t > T$ - a prespecified length of the chain.

For selection of the scale parameter σ, in our following examples, σ is often set after a few trials by justifying the convergence of Markov chains graphically. σ can also be deter-

mined in an automatic and adaptive way to yield a desirable overall sampler acceptance probability [Garthwaite et al., 2010].

Let T be the length of the chain $\{\Delta_t\}$, and T_0 be the *burn-in* number. Then expectation (3.3) is approximated by

$$\delta_{\mathrm{KL}}(M_k) = \mathbf{E}\left[\log \frac{p(\Delta|D)}{p(\Delta|D, M_k)}\right] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \log \frac{p(\Delta_t|D)}{p(\Delta_t|D, M_k)}.$$

In order to get an accurate approximation, our experience shows that several chains with overdispersed starting points should be sampled so that the chains can run through the whole support of the target distribution.

Calculating integrals (3.6) **and** (3.7)**.** What remains in implementing the POPMOS is to compute integrals (3.6) and (3.7). In some special cases such as linear regression with conjugate priors (see Section 3.1.5) or discrete graphical models [Madigan and York, 1995], integrals (3.6) and (3.7) have closed forms. In general cases, the Laplace approximation is often used to estimate $p(D|M_k)$ [Schwarz, 1978, Tierney and Kadane, 1986, Raftery, 1996], and $p(\Delta|D,M_k)$ is often approximated by $p(\Delta|\hat{\theta}_k,M_k)$ where $\hat{\theta}_k$ is the maximum likelihood estimate of θ_k [Taplin, 1993, Draper, 1995]. The relative approximation error is $O(n^{-1})$ [Kass and Vaidyanathan, 1992].

3.1.3 Measures of predictive ability

As mentioned earlier, a primary goal of statistical analysis is to make predictions and inferences on future data. Many authors argue that a model is more impressive/preferable if it assigns higher probabilities to the actual (test) data. Thus, a good and widely-used measure of predictive ability is the *partial predictive score* (PPS) [Good, 1952, Geisser, 1980, Hoeting et al., 1999]. Suppose that the data is split into two parts, the *training set*

D^T and the *prediction set* D^P. Then the partial predictive score of model M is defined as

$$\text{PPS}(M) = -\frac{1}{|D^P|}\sum_{\Delta\in D^P}\log p(\Delta|M, D^T) \tag{3.11}$$

where $|D^P|$ is the cardinality of D^P and $p(\Delta|M,D^T)$ is the predictive distribution under model M given the training data D^T. The smaller the PPS, the better the predictive performance.

Another measure of predictive ability is the *predictive coverage* (PC). Consider the regression context in which $\Delta=(x,y)$ where x is the explanatory value and y is the response value. Let m and s be the mean and the standard deviation (which can be estimated by MCMC) of the predictive distribution $p(\Delta|M,D^T)=p(y|x,\mathcal{M},D^T)$ of the response y at predictor value x. The 90%, say, prediction interval for a future observation of response y at x is approximated by the interval $m\pm 1.645s$. The (90%) PC then is defined as the proportion of observations in D^P that fall in the 90% prediction interval. In the following examples, we use these two measures, PPS and PC, to assess the predictive performance of selected models.

3.1.4 Model uncertainty indicator

We now introduce an indicator to measure model uncertainty, which we call the *model uncertainty indicator* (MUI). It is defined as the ratio of the second highest posterior model probability to the highest. More formally, let $M_0=\text{argmax}_{M\in\mathcal{A}}p(M|D)$, then the MUI is defined as

$$\text{MUI} = \frac{\max_{M\in\mathcal{A}\setminus\{M_0\}}p(M|D)}{p(M_0|D)} \leq 1. \tag{3.12}$$

It is clear that the larger the MUI is, the more model uncertainty there is. A very small MUI indicates no model uncertainty. Our experience shows that when MUI is small enough (often, MUI $\leq .5$), the OP model (i.e., the model selected by (3.10)) coincides with the

highest posterior probability model. However, if the MUI is large, the OP model is often different to the highest posterior probability model and has better predictive ability.

3.1.5 An example

We now demonstrate the POPMOS for predictive variable selection in linear regression analysis where the BMA is used as the reference. We use the Bayesian framework used in Raftery et al. [1997]. Each model M under consideration is of the form

$$Y = \beta_0 + \beta_1 X_{i_1} + ... + \beta_k X_{i_k} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

where $\{X_{i_1},...,X_{i_k}\}$ is a subset of the set $\{X_1,...,X_p\}$ of all potential covariates. Let $\mathbf{y}$ and X (w.r.t. M) be the response vector and the corresponding design matrix, respectively. It is reasonable to assign a uniform prior to possible combinations of covariates, i.e., the prior information is "objective" between models. For model parameters, we assume priors

$$\beta|\sigma^2 \sim N_{k+1}(\mu, \sigma^2 V)\ \ ,\ \ \frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu.$$

Hyperparameters μ,V,ν,λ are chosen as follows (see Raftery et al. [1997] for the details)

$$\nu = 2.58,\ \lambda = .28,\ \mu = (\hat{\beta}_0, 0, ..., 0),\ V = \mathrm{diag}(s^2_Y, \frac{\phi^2}{s^2_{i_1}}, ..., \frac{\phi^2}{s^2_{i_k}})$$

where $\hat{\beta}_0$ is the OLS estimate of β_0, s^2_Y,$s^2_{i_1}$,...,$s^2_{i_k}$ are sample variances of Y,X_{i_1},...,X_{i_k}, respectively, and ϕ=2.85. Typically, in our experience, results are relatively insensitive to changes in values of the hyperparameters.

Then the marginal likelihood (3.6) under model M is

$$p(D|M) = \frac{\Gamma(\frac{\nu+n}{2})(\nu\lambda)^{\nu/2}[\lambda\nu + (\mathbf{y} - X\mu)^\top(\mathbb{1} + XVX^\top)^{-1}(\mathbf{y} - X\mu)]^{-(\nu+n)/2}}{\pi^{n/2}\Gamma(\nu/2)|\mathbb{1} + XVX^\top|^{1/2}} \qquad (3.13)$$

and the posterior predictive distribution (3.7) is

$$p(\Delta|D, M) = \frac{\Gamma(\frac{\nu+n+1}{2})}{\sqrt{\pi}\Gamma(\frac{n+\nu}{2})} \frac{1}{(1 + \mathbf{x}^\top(X^\top X + V^{-1})^{-1}\mathbf{x})^{1/2}} \frac{A^{(n+\nu)/2}}{B^{(n+\nu+1)/2}} \qquad (3.14)$$

where

$$A = \lambda\nu + \|\mathbf{y}\|^2 + \mu^\top V^{-1}\mu - (X^\top\mathbf{y} + V^{-1}\mu)^\top(X^\top X + V^{-1})^{-1}(X^\top\mathbf{y} + V^{-1}\mu)$$

and

$$B = \lambda\nu + \|\mathbf{y}\|^2 + y^2 + \mu^\top V^{-1}\mu - \\ -(\mathbf{x}y + X^\top\mathbf{y} + V^{-1}\mu)^\top(\mathbf{x}\mathbf{x}^\top + X^\top X + V^{-1})^{-1}(\mathbf{x}y + X^\top\mathbf{y} + V^{-1}\mu).$$

Analysis of the crime data

Criminal behavior has been argued to be strongly related to criminal activity's costs and benefits and to other legitimate opportunities. Ehrlich [1973] used the data from 47 U.S. states in 1960 to test this argument. The dependent variable was the crime rate. The costs of crime were measured by probability of imprisonment and average time served in prison. The benefits were related to wealth and income inequality in the community. The investigation also included other variables such as sex ratio, percentage of young males, etc. In summary, 15 potential covariates (Table 3.1) were considered.

This benchmark dataset has been analyzed by many authors. Previous diagnostic checkings (see, e.g., Draper and Smith [1981]) did not show any violation of the linear assumption. Ehrlich [1973] used the stepwise method to select significant variables. However, Raftery et al. [1997] reported evidence against Ehrlich's results and suggested using posterior probabilities to do variable selection. We now use this dataset to demonstrate the POPMOS and compare it to other model selection rules.

Table 3.1 summarizes the experimental results using the whole dataset. Models selected by different methods are listed in the corresponding columns. The third column is the overall posterior probability that the j-th covariate is in a model, i.e., $P(\beta_j \neq 0|D)$, calculated by summing the posterior probabilities of models that contain the j-th covariate,

Table 3.1: Crime data: Overall posterior probabilities and selected models

Number	Covariate	$P(\beta_j \neq 0 \mid D)$	AIC	BIC	OP	MP
1	% of males age 14-24	.78	$\star$	$\star$	$\star$	$\star$
2	Indicator for southern state	.18				
3	Mean years of schooling	.97	$\star$	$\star$	$\star$	$\star$
4	Police expenditure in 1960	.72	$\star$	$\star$	$\star$	$\star$
5	Police expenditure in 1959	.50			$\star$	$\star$
6	Labor force participation rate	.08				
7	No. males per 1000 females	.08				
8	State population	.24				
9	No. nonwhites per 1000 people	.61	$\star$	$\star$	$\star$	$\star$
10	Unemployment rate age 14-24	.11				
11	Unemployment rate age 35-39	.45	$\star$	$\star$		
12	Wealth	.31	$\star$			
13	Income inequality	1.00	$\star$	$\star$	$\star$	$\star$
14	Probability of imprisonment	.82	$\star$	$\star$	$\star$	$\star$
15	Ave. time in state prisons	.23	$\star$			
MUI=.71 suggests that there is high model uncertainty						

$j = 1,2,...,15$. The POPMOS selected the predictors with highest posterior probabilities ($\geq .5$). Raftery et al. recommended (from an empirical analysis) using posterior probabilities rather than p-values for variable selection. The last column presents the so-called *median probability model (MP)* introduced by Barbieri and Berger [2004]. The MP model is defined as the model consisting of those covariates which have overall posterior probability $P(\beta_j \neq 0|D) \geq .5$. In the framework of normal linear regression and under some conditions, Barbieri and Berger showed that the MP model has the optimal predictive

performance in terms of predictive expected squared loss (see Barbieri and Berger [2004] for the full definition). As shown in Table 3.1, the OP model is the same as the MP model.

Table 3.1 also shows the models selected by AIC and BIC (which were exhaustively searched by using the branch-and-bound algorithm [Miller, 2002]). AIC, BIC and POP-MOS produced three different models. This is not a surprise because these criteria have different goals. As we may expect, the AIC model is the "biggest" model among selected models: it contains 9 covariates versus 7 covariates for OP and BIC. As we will see next, AIC models sometimes have poor predictive performances.

We now use the crime data to assess the predictive ability of the selection rules. To this end, the dataset was randomly split into two parts. One with 24 observations was used as the training set, the other with 23 observations was used as the prediction set. Other splits can be adopted. Table 3.2 shows the PPS and PC of the selected models. With $C=20$ being used, model set $\mathcal{A}$ contains 29 models. The model uncertainty indicator MUI=.61 suggests that there is moderate model uncertainty. As shown, the OP model has a better predictive performance than the AIC and BIC models. AIC has a poor predictive performance.

Note that the models selected using half of the data are slightly different from the models selected using the full data (however, they both contain the most important covariates). This is not a surprise because of the small size of the dataset. If we had a large enough dataset, using either the full data or half of it would lead to the same results. The selected models summarized in Table 3.2 are used only to examine the methods, they are not the final chosen models.

Table 3.2: Crime data: Assessment of predictive ability

Method	Model							PPS	PC
AIC	1	3	4	5	9	13	14	.18	82.61%
BIC		3	4		9	13	14	.16	82.61%
MP	1	3	4	5	9	13		.12	86.96%
OP	1	3	4	5	9	13		.12	86.96%
BMA	all							.06	91.30%
MUI=.61 suggests that there is moderate model uncertainty									

3.2 The predictive Lasso

We present in this section an application of the POPMOS idea to the variable selection problem in generalized linear models (GLMs). The method described in the previous section may be challenging to implement in high-dimensional GLMs because searching over the whole model space is computationally infeasible. Like the idea of the Lasso, we overcome this problem by using l_1 constraints on the coefficients. By doing this, we can enjoy the computational advantages of algorithms for convex optimization with l_1 constraints. Unlike the Lasso, however, our approach has an explicit predictive motivation which aims at selecting a useful model with high prediction accuracy. We refer to this methodology as the predictive Lasso or pLasso for short. The pLasso will be fully developed in Section 3.2.1-3.2.2 and some examples will be presented in Section 3.2.3. This material was developed in Tran et al. [2010].

3.2.1 The predictive Lasso

We consider the problem of simultaneous coefficient estimation and variable selection for GLMs with potential covariates $\mathbf{x}=(x_0\equiv 1,x_1,...,x_p)^\top\in\mathcal{X}$ and the response $y\in\mathcal{Y}$. With a suitable link function g, $g(E(y|\mathbf{x}))$ is assumed to be a linear combination of $\mathbf{x}$

$$g(E(y|\mathbf{x})) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = \mathbf{x}^\top\boldsymbol{\beta}. \tag{3.15}$$

We assume that the covariates x_i are in their final forms, no further transformations are needed (i.e., for various reasons and in order to keep things simple, we restrict ourselves to the linear approximation (3.15)). The sampling distribution of an observation $\Delta_i=(\mathbf{x}_i,\, y_i)$ then is assumed to have the following form

$$p(\Delta_i|\boldsymbol{\beta},\phi) = p(\mathbf{x}_i)p(y_i|\mathbf{x}_i,\boldsymbol{\beta},\phi) \propto p(\mathbf{x}_i)\exp\left(\frac{1}{a(\phi)}\left[y_i\theta(\mathbf{x}_i^\top\boldsymbol{\beta}) - b(\theta(\mathbf{x}_i^\top\boldsymbol{\beta}))\right]\right),$$

where $\boldsymbol{\beta}\in I\!R^{p+1}$, $\phi>0$ are the coefficient vector and scale parameter, respectively, and θ, a and b are known functions. In order to discuss the methodology in a general setting, we consider predictors $\mathbf{x}$ as random. Bayesian variable selection with a random covariate has been considered in a decision theoretic framework where the main concern is prediction of a future observation for which the corresponding predictor is not yet observed (see, e.g., Lindley [1968]). The case with fixed design points can be considered as a special case, then the density $p(\mathbf{x}_i)$ in the above expression can be omitted.

We are concerned with the problem of simultaneous coefficient estimation and variable selection with the goal of prediction in mind. Like the Lasso, we would like to develop a method for simultaneous variable selection and parameter estimation. However, unlike the Lasso our approach has a more explicit predictive motivation, which aims at producing a useful model with high prediction accuracy.

Given the past dataset D and certain priors for parameters $(\boldsymbol{\beta},\, \phi)$ of the full model,

the predictive distribution $p(\Delta|D)$ for a future observation $\Delta=(\mathbf{x},y)$ is given by

$$p(\Delta|D) = p(\mathbf{x}|D)p(y|\mathbf{x},D) = p(\mathbf{x}|D)\int p(y|\mathbf{x},\boldsymbol{\beta},\phi)p(\boldsymbol{\beta},\phi|D)d\boldsymbol{\beta}d\phi. \tag{3.16}$$

We can assume that $p(\mathbf{x}|D)\equiv p(\mathbf{x})$, i.e., future design points are independent of past data. We propose to estimate the coefficient vector $\boldsymbol{\beta}$ by solving the following optimization problem:

$$\min_{\beta}\int \log\frac{p(\Delta|D)}{p(\Delta|\boldsymbol{\beta},\phi)}p(\Delta|D)d\Delta \quad \text{s.t.} \quad \sum_{j=1}^{p} w_j|\beta_j| \le \tau \tag{3.17}$$

where the tuning parameter $\tau\ge 0$ and weights $w_j\ge 0$ are chosen later. (As will become clear shortly, ϕ plays no role in this optimization problem, we can assume at the moment that ϕ is known). Note that the objective function is the Kullback-Leibler divergence from $p(\Delta|\boldsymbol{\beta},\phi)$ to the reference predictive distribution $p(\Delta|D)$. We refer to this procedure of estimating $\boldsymbol{\beta}$ through the optimization of (3.17) as the predictive Lasso (pLasso).

Let $\{\Delta_t=(\mathbf{x}_t,y_t),\, t=1,...,T\}$ be Markov chain Monte Carlo (MCMC) samples from the predictive distribution $p(\Delta|D)$. The integral in (3.17) then can be approximated by the average $(1/T)\sum_{t=1}^{T}\log[p(\Delta_t|D)/p(\Delta_t|\boldsymbol{\beta},\phi)]$, and (3.17) becomes

$$\min -\frac{1}{T}\sum_{t=1}^{T}\log p(\Delta_t|\boldsymbol{\beta},\phi) \quad \text{s.t.} \quad \sum_{j=1}^{p} w_j|\beta_j| \le \tau, \tag{3.18}$$

or more specifically

$$\min \frac{1}{T}\sum_{t=1}^{T}\left[b(\theta(\mathbf{x}_t^\top\boldsymbol{\beta})) - y_t\theta(\mathbf{x}_t^\top\boldsymbol{\beta})\right] \quad \text{s.t.} \quad \sum_{j=1}^{p} w_j|\beta_j| \le \tau. \tag{3.19}$$

This optimization problem is also equivalent to

$$\min \frac{1}{T}\sum_{t=1}^{T}\left[b(\theta(\mathbf{x}_t^\top\boldsymbol{\beta})) - y_t\theta(\mathbf{x}_t^\top\boldsymbol{\beta})\right] + \lambda\sum_{j=1}^{p} w_j|\beta_j| \tag{3.20}$$

where λ is a tuning parameter. Such an optimization problem is easier to deal with if the objective function is convex. The convexity of the objective function turns out to depend on the link function, and holds for most popular GLMs.

Often, the integral in $\mathbf{x}$ is approximated by a sum over N points $\mathbf{x}_1^f,...,\mathbf{x}_N^f$. These points might not coincide with the observed design points, they "come from the future" (hence the superscript "f" stands for "future"). For each $\mathbf{x}_i^f$, let $\bar{y}_i^f$ be the mean of MCMC samples $\{y_{it}, t=1,2,...\}$ drawn from $p(y_i^f|\mathbf{x}_i^f, D)$ - the predictive distribution of the future response y_i^f at design point $\mathbf{x}_i^f$ given past data D. Then, it is easy to see that (3.20) becomes

$$\min \frac{1}{N}\sum_{i=1}^{N}[b(\theta(\boldsymbol{\beta}^\top \mathbf{x}_i^f)) - \bar{y}_i^f \theta(\boldsymbol{\beta}^\top \mathbf{x}_i^f)] + \lambda \sum_{j=1}^{p} w_j|\beta_j|. \tag{3.21}$$

Note that, under the squared error loss, $\bar{y}_i^f$ is an estimate of the best prediction (w.r.t. the predictive distribution $p(y_i^f|\mathbf{x}_i^f, D)$) for the response at $\mathbf{x}_i^f$. As will be seen in Section 3.2.2, for linear regression with a convenient specification of priors there is no need to conduct MCMC because the predictions $\bar{y}_i^f = E(y_i^f|\mathbf{x}_i^f, D)$ have a closed form.

We have approximated the integral over $\mathbf{x}$ by a sum over N "future" points $\mathbf{x}_i^f$, $i = 1,...,N$. Typically, these points are specified depending on the context and/or on the distribution $p(\mathbf{x})$ over $\mathcal{X}$. As a default implementation of our procedure, however, we propose to identify the future points $\mathbf{x}_i^f$ with the observed training points $\mathbf{x}_i$, $i = 1,...,n$. The reason behind this is that if the sample size n is large enough and the observed training points $\mathbf{x}_i$ were randomly selected from $p(\mathbf{x})$, then by the law of large numbers the integral over $\mathbf{x}$ can be well approximated by the sum over $\mathbf{x}_i$. In what follows therefore, if not otherwise specified, we consider the pLasso for GLMs in the following form

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}[b(\theta(\mathbf{x}_i^\top \boldsymbol{\beta})) - \bar{y}_i^f \theta(\mathbf{x}_i^\top \boldsymbol{\beta})] + \lambda \sum_{j=1}^{p} w_j|\beta_j|. \tag{3.22}$$

Note that the original (adaptive) Lasso for GLMs is

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}[b(\theta(\mathbf{x}_i^\top \boldsymbol{\beta})) - y_i \theta(\mathbf{x}_i^\top \boldsymbol{\beta})] + \lambda \sum_{j=1}^{p} w_j|\beta_j|. \tag{3.23}$$

The pLasso in this form differs from the original Lasso only in the way it replaces the observed responses y_i by the predictions $\bar{y}_i^f = E(y_i^f|\mathbf{x}_i,D)$. Available routines to solve (3.23) then can be used for (3.22).

We have not yet considered the issue of choice of the tuning parameters in the pLasso. As the primary goal of the pLasso is to predict the future, cross-validation is a very natural choice for estimating λ. As in the adaptive Lasso, the weights w_j can be assigned as $1/|\tilde{\beta}_j|$ with $\tilde{\beta}_j$ the MLE of β_j or some others such as the Lasso estimate. In a Bayesian context it is also natural to consider $\tilde{\beta}_j$ as the posterior mode.

3.2.2 Some useful prior specifications

Given the available routines to solve the optimization problem of form (3.22), all what we need to implement the pLasso is to calculate the quantities $\bar{y}_i^f = E(y_i^f|\mathbf{x}_i,D)$. To do so, in general, we first need to specify a useful prior for parameters, determine posterior distributions and then estimate $\bar{y}_i^f = E(y_i^f|\mathbf{x}_i,D)$ by MCMC or some other method. However, in some cases there is no need to conduct MCMC. We first present in this section a prior specification for linear models in which the predictions $\bar{y}_i^f$ have closed form. For genalized linear models, we present here two prior specifications. The first is adapted from Chen and Ibrahim [2003] which is interpretable in terms of observables rather than parameters. The second one proposed recently by Gelman et al. [2008] is useful for routine applied use.

Prior specification for linear models

Consider the usual linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the n-vector of responses, X is an $n\times(p+1)$ design matrix and $\boldsymbol{\epsilon}$ is an n-vector of iid normal errors with mean zero and variance σ^2. The $(p+1)$-vector $\boldsymbol{\beta}$ consists of unknown parameters and we consider the situation where σ^2 is also unknown. Consider the conjugate prior specification [O'Hagan and Forster, 2004, Chapter 11] $p(\boldsymbol{\beta},\sigma^2)=p(\sigma^2)p(\boldsymbol{\beta}|\sigma^2)$ in which $p(\sigma^2)$ is inverse gamma

$$p(\sigma^2) = \frac{(a/2)^{(d/2)}}{\Gamma(d/2)}(\sigma^2)^{-d/2-1}\exp(-\frac{a}{2\sigma^2})$$

and $p(\boldsymbol{\beta}|\sigma^2)$ is multivariate normal, $N(\mathbf{m},\sigma^2 V)$. With these priors the predictive distribution of a new observation $\Delta=(\mathbf{x},y)$ is $p(\Delta|D)=p(\mathbf{x}|D)p(y|\mathbf{x},D)$ with $p(y|\mathbf{x},D)=t_{d+n}\Big(\mathbf{x}^\top\tilde{\boldsymbol{\beta}},s^2(1+\mathbf{x}^\top\hat{V}\mathbf{x})\Big)$ where

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (X^\top X+V^{-1})^{-1}(V^{-1}\mathbf{m}+X^\top\mathbf{y}),\\
\hat{V} &= (V^{-1}+X^\top X)^{-1},\\
s^2 &= \frac{a+\mathbf{m}^\top V^{-1}\mathbf{m}+\mathbf{y}^\top\mathbf{y}-(V^{-1}\mathbf{m}+X^\top\mathbf{y})^\top(V^{-1}+X^\top X)^{-1}(V^{-1}\mathbf{m}+X^\top\mathbf{y})}{n+d-2},\\
\hat{\boldsymbol{\beta}} &= (X^\top X)^{-1}X^\top\mathbf{y}.
\end{aligned}$$

We write $w(\mathbf{x})=1+\mathbf{x}^\top\hat{V}\mathbf{x}$.

Now consider the predictive Lasso (3.17) where as usual the integral over $\mathbf{x}$ is approximated by a sum over N "future" points $\mathbf{x}_i^f$. Then equivalently, we need to minimize (the scale ϕ is now re-denoted by σ^2)

$$\sum_{i=1}^N\int\Big[-\log p(y_i^f|\mathbf{x}_i^f,\boldsymbol{\beta},\sigma^2)\Big]\,p(y_i^f|\mathbf{x}_i^f,D)dy_i^f \quad\text{s.t.}\quad \sum_{j=1}^p w_j|\beta_j|\le\tau. \tag{3.24}$$

Noting that

$$\log p(y_i^f|\mathbf{x}_i^f,\boldsymbol{\beta},\sigma^2) = -\frac{1}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}(y_i^f-(\mathbf{x}_i^f)^\top\boldsymbol{\beta})^2,$$

minimizing (3.24) is equivalent to minimizing

$$\frac{N}{2}\log\sigma^2+\frac{1}{2\sigma^2}\sum_{i=1}^N E\Big((y_i^f-(\mathbf{x}_i^f)^\top\boldsymbol{\beta})^2|\mathbf{x}_i^f,D\Big) \quad\text{s.t.}\quad \sum_{j=1}^p w_j|\beta_j|\le\tau. \tag{3.25}$$

With the closed form of the predictive distribution as a t-distribution we have

$$E\left((y_i^f - (\mathbf{x}_i^f)^\top\boldsymbol{\beta})^2|\mathbf{x}_i^f, D\right) = s^2 w(\mathbf{x}_i^f) + \left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}} - (\mathbf{x}_i^f)^\top\boldsymbol{\beta}\right)^2.$$

Substituting this into (3.25) we must minimize

$$\frac{N}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N} s^2 w(\mathbf{x}_i^f) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}} - (\mathbf{x}_i^f)^\top\boldsymbol{\beta}\right)^2 \tag{3.26}$$

subject to the constraint. Minimizing this as a function of $\boldsymbol{\beta}$ amounts as before to an ordinary Lasso problem where the responses are replaced with the fitted values from the full model at the future design points $\mathbf{x}_i^f$, $i=1,...,N$. With a non-informative prior and with the $\mathbf{x}_i^f$ as the observed design points $\mathbf{x}_i$ this is the ordinary Lasso, since in this case $\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}$ and for the least squares estimator

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta})^2 = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}})^2 + \sum_{i=1}^{n}(\mathbf{x}_i^\top\hat{\boldsymbol{\beta}} - \mathbf{x}_i^\top\boldsymbol{\beta})^2$$

where the first term on the right hand side does not depend on $\boldsymbol{\beta}$.

If (3.26) has been minimized with respect to $\boldsymbol{\beta}$ subject to the constraint to obtain an estimate $\hat{\boldsymbol{\beta}}_{\text{pLasso}}$ (this in general depends on the constraint τ but we suppress this in the notation) then substituting in $\hat{\boldsymbol{\beta}}_{\text{pLasso}}$ and minimizing with respect to σ^2 gives

$$\begin{aligned}\hat{\sigma}^2_{\text{pLasso}} &= \frac{\sum_{i=1}^{N}\text{Var}(y_i^f|\mathbf{x}_i^f, D) + \sum_{i=1}^{N}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}} - (\mathbf{x}_i^f)^\top\hat{\boldsymbol{\beta}}_{\text{pLasso}}\right)^2}{N} \\ &= \frac{\sum_{i=1}^{N} s^2 w(\mathbf{x}_i^f) + \sum_{i=1}^{N}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}} - (\mathbf{x}_i^f)^\top\hat{\boldsymbol{\beta}}_{\text{pLasso}}\right)^2}{N}.\end{aligned} \tag{3.27}$$

The weighted version of pLasso. One extension we can consider is the following. Suppose that instead of considering sampling distributions in our predictive Lasso where the variance does not depend on $\mathbf{x}$, we predict y_i^f with

$$p(y_i^f|\boldsymbol{\beta}, \sigma^2 w(\mathbf{x}_i^f)) = N((x_i^f)^\top\boldsymbol{\beta}, \sigma^2 w(\mathbf{x}_i^f)).$$

That is, we allow our normal distributions to have variances vary in proportion to the true predictive variances in the full model $\text{Var}(y_i^f|\mathbf{x}_i^f,D)$. The standard deviation in the full model $(\text{Var}(y_i^f|\mathbf{x}_i^f,D))^{1/2}$ is often considered a more realistic estimate of the standard error, because it incorporates model uncertainty. We now consider minimization of

$$\sum_{i=1}^{N}\int\left[-\log p(y_i^f|\boldsymbol{\beta},\sigma^2 w(\mathbf{x}_i^f))\right]p(y_i^f|\mathbf{x}_i^f,D)dy_i^f$$

subject to the constraint, and following a similar argument to our previous one, we must minimize

$$\sum_{i=1}^{N}\frac{1}{w(\mathbf{x}_i^f)}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}}-(\mathbf{x}_i^f)^\top\boldsymbol{\beta}\right)^2$$

subject to the constraint in order to estimate $\boldsymbol{\beta}$. This is similar to before, but now with weights of $1/w(\mathbf{x}_i^f)$ for the different design points. We will refer to this producdure as the weighted pLasso (wpLasso). After $\boldsymbol{\beta}$ has been estimated as $\hat{\boldsymbol{\beta}}_{\text{wpLasso}}$ say, the minimization with respect to σ^2 gives

$$\begin{aligned}\hat{\sigma}^2_{\text{wpLasso}} &= \frac{\sum_{i=1}^N\frac{1}{w(x_i)}\text{Var}(y_i^f|\mathbf{x}_i^f,D)+\sum_{i=1}^N\frac{1}{w(x_i^f)}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}}-(\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}}_{\text{wpLasso}}\right)^2}{N}\\ &= \frac{\sum_{i=1}^N s^2+\sum_{i=1}^n\frac{1}{w(x_i^f)}\left((\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}}-(\mathbf{x}_i^f)^\top\tilde{\boldsymbol{\beta}}_{\text{wpLasso}}\right)^2}{N}.\end{aligned}$$

Elicitation of hyperparameters. We now discuss on the choice of the hyperparameters $\mathbf{m}$, V, a and d. There are many different ways proposed for choosing the matrix V in the literature. For example, Zellner [1986] proposed the so-called *g-prior* in which V is set equal to $c(X^\top X)^{-1}$ with some $c>0$ ($c=n$ is a common choice). Raftery et al. [1997] proposed an alternative where V is a block-diagonal matrix. For noncategorical covariates, V is a diagonal matrix $\text{diag}(s_y^2,\kappa^2 s_1^{-2},...,\kappa^2 s_p^{-2})$ where s_y^2 is the sample variance of $\mathbf{y}$, and s_i^2 are the variances of the columns of X. For a categorical covariate, the corresponding diagonal element will be a matrix induced from the corresponding dummy variables. Raftery et al.

[1997] proposed a value of 2.85 for κ together with $a=0.72$ and $d=2.58$. For the parameter $\mathbf{m}$, they proposed the default value of $\mathbf{m}=(\hat{\beta}_0^{\text{OLS}},0,...,0)^\top$ where $\hat{\beta}_0^{\text{OLS}}$ is the OLS estimate of β_0. An alternative is $\mathbf{m}=\mathbf{0}$. These two choices of $\mathbf{m}$ often lead to very similar inferences. We will use the setup of Raftery et al. [1997] in our following numerical examples.

Prior specifications for generalized linear models

There is an extensive literature on prior specifications for GLMs. We will briefly present here two of them: the first one is due to Chen and Ibrahim [2003] and the second is proposed recently by Gelman et al. [2008].

The Chen and Ibrahim prior. Recall that the sampling distribution of observables $\mathbf{y}=(y_1,...,y_n)^\top$ in the GLM case is

$$p(\mathbf{y}|X,\boldsymbol{\beta},\phi) \propto \exp\left(\sum_1^n \frac{1}{a(\phi)}\left[y_i\theta(\mathbf{x}_i^\top\boldsymbol{\beta}) - b(\theta(\mathbf{x}_i^\top\boldsymbol{\beta}))\right]\right) = \exp\left(\frac{1}{a(\phi)}\left[\mathbf{y}^\top\boldsymbol{\theta} - \mathbb{1}^\top\mathbf{b}(\boldsymbol{\theta})\right]\right)$$

where $\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\beta})=(\theta_1,...,\theta_n)^\top$, $\theta_i=\theta(\mathbf{x}_i^\top\boldsymbol{\beta})$, $\mathbf{b}(\boldsymbol{\theta})=(b(\theta_1),...,b(\theta_n))^\top$ and $\mathbb{1}$ is an $n-$vector of 1s. For ease of exposition, we assume that ϕ is known (and therefore suppressed in the notation), as, for example, in logistic and Poisson regression. Chen and Ibrahim [2003] proposed the following prior for $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}) \propto \exp\left(\gamma_0\frac{1}{a(\phi)}\left[\boldsymbol{\alpha}_0^\top\boldsymbol{\theta} - \mathbb{1}^\top\mathbf{b}(\boldsymbol{\theta})\right]\right) \tag{3.28}$$

where $\gamma_0\geq 0$ and $\boldsymbol{\alpha}_0\in\mathbb{R}^n$ are hyperparameters determined later on. Denote this distribution by $\boldsymbol{\beta}|\phi\sim D(\gamma_0,\boldsymbol{\alpha}_0)$. They proved that the prior (3.28) is proper and that this prior is conjugate with the posterior $\boldsymbol{\beta}|X,\mathbf{y}\sim D(1+\gamma_0,(\gamma_0\boldsymbol{\alpha}_0+\mathbf{y})/(1+\gamma_0))$.

As shown by Chen and Ibrahim [2003], $E(\mathbf{y})=\boldsymbol{\alpha}_0$, it is natural to choose $\boldsymbol{\alpha}_0$ as a prior guess for $E(\mathbf{y})$. Therefore, in practice, $\boldsymbol{\alpha}_0$ should be obtained from experts in the field although default empirical Bayes alternatives such as choosing $\boldsymbol{\alpha}_0$ as the fitted values

based on the MLE or other methods are also possible. The parameter γ_0 weighs the importance of the prior guess. In general, γ_0 should be taken such that $\gamma_0 = \gamma_0(n) \to 0$ as $n \to \infty$, i.e., the prior has less influence when more data is available. An advantage of this prior specification is that it is interpretable in terms of observables rather than parameters which are sometimes not easy to elicit.

The Gelman et al. prior. Gelman et al. [2008] proposed a weakly informative prior distribution for GLMs, constructed by first standardizing the covariates to have mean zero and standard deviation 0.5, and then putting independent $t-$distributions on the coefficients. As a default choice, they recommended a central Cauchy distribution with scale 10 for the intercept and central Cauchy distributions with scale 2.5 for other coefficients. As argued by Gelman et al. [2008], this prior specification has many advantages; besides, it works in an automatic fashion with no hyperparameter elicitation needed.

Recall that all what we need to implement the pLasso is to calculate the quantities $\bar{y}_i^f = E(y_i^f | \mathbf{x}_i, D)$. After the prior has been specified, $\bar{y}_i^f$ can be estimated by MCMC or some other method. It is well-known that

$$E(\mathbf{y}|X, \boldsymbol{\beta}) = \dot{\mathbf{b}}(\boldsymbol{\theta}) = (\dot{b}(\theta_1), ..., \dot{b}(\theta_n))^\top,$$

so that

$$\bar{\mathbf{y}}^f = E(\mathbf{y}^f | X, \mathbf{y}) = E_{\beta|X,y}\left[E(\mathbf{y}^f | X, \boldsymbol{\beta})\right] = E_{\beta|X,y}[\dot{\mathbf{b}}(\boldsymbol{\theta}(\boldsymbol{\beta}))] \tag{3.29}$$

which can be easily estimated by MCMC samples from the posterior distribution $\boldsymbol{\beta}|X,\mathbf{y}$.

A procedure for fitting GLMs with the Gelman et al. prior has been implemented in `R` by Gelman et al. (available online at `http://cran.r-project.org/web/packages/arm`). In the following numerical examples for logistic regression where no expert advice is available, we use the default prior of Gelman et al. For high-dimensional cases where using MCMC may be time consuming, we suggest using the plug-in predictive density (i.e., the

density with its parameters fixed at their estimates) to estimate the predictions $\bar{y}_i^f$. Our experiences show that this is very fast compared to MCMC.

3.2.3 Experiments

In this section, we study the pLasso through simulations and real-data examples. We use the convenient prior specifications as in Section 3.2.2. The tuning parameter λ is selected by 5-fold cross-validation. The examples are carried out using R with the help of the R packages glmnet and arm.

As before, we use the popular PPS to measure predictive ability. We also use another interesting predictive measure, called continuous ranked probability score (CRPS) [Gneiting and Raftery, 2007]. Let F be the cumulative distribution function (cdf) of the predictive distribution in use and x be an actual observation. The CRPS is defined as

$$\text{CRPS}(F, x) = -\int_R (F(y) - \mathbb{1}_{y \geq x})^2 dy$$

which corresponds to the integral of the Brier scores [Hersbach, 2000]. A problem with using CRPS is that the above integral is in general not available in closed form and needs to be estimated in some way. However, when F is the cdf of the normal distribution with mean μ and variance σ^2, the CRPS is given by [Gneiting and Raftery, 2007, p. 367]

$$\text{CRPS}(N(\mu, \sigma^2), x) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\varphi(\frac{x-\mu}{\sigma}) - \frac{x-\mu}{\sigma}\Big(2\phi(\frac{x-\mu}{\sigma}) - 1\Big) \right]$$

where φ and ϕ are pdf and cdf of the standard Gaussian variable; when F is the cdf of a Bernoulli variable X with probability of success $p = P(X = 1)$, the CRPS is given by

$$\text{CRPS}(F(p), x = 0) = -p^2 \text{ and } \text{CRPS}(F(p), x = 1) = -(1-p)^2.$$

The CRPS evaluated on a prediction set D^P of the predictive distributions induced by

model parameters θ^* is defined as

$$\text{CRPS} \equiv \text{CRPS}(\theta^*) = -\frac{1}{|D^P|} \sum_{\Delta \in D^P} \text{CRPS}(F(\theta^*), \Delta). \tag{3.30}$$

Under this formulation, it is (similar to PPS) understood that smaller CRPS means better predictive performance.

In the simulation studies below, we also use mean squared errors (MSE) in terms of coefficients and numbers of zero-estimated (NZE) coefficients to measure the performance.

A simulation study for linear regression. Consider the following linear model

$$y = 2 + \mathbf{x}'\boldsymbol{\beta} + \sigma\epsilon \tag{3.31}$$

where $\boldsymbol{\beta}$=(3, 1.5, 0, 0, 0.5, 0.5, 0, 0)$'$ (so that there are some main and also small effects), ϵ is iid $N(0,1)$, and $\sigma > 0$ is the noise level. We want to compare the performance of the pLasso and the wpLasso to that of the adaptive Lasso (aLasso). We also consider the orginal Lasso and the non-adaptive pLasso (i.e., the adaptive penalty term $\lambda\sum w_j|\beta_j|$ in (3.22) is replaced by $\lambda\sum|\beta_j|$) which will be abbreviated as npLasso.

In our first simulation study, design points $\mathbf{x}_j$ are simulated from a multivariate normal distribution $N_8(\mathbf{0},\Sigma)$ with $\sigma_{ij} = 0.5^{|i-j|}$. We first generate from model (3.31) a dataset which serves as the training set D^T. Another dataset D^P then is generated, which is used to test the predictive performance. Table 3.3 presents the PPS (after ignoring the constants independent of models), MSE, NZE and CRPS averaged over 500 replications with various factors n^T (size of training set), n^P (size of prediction set) and σ. The numbers in parentheses are standard deviations. The result suggests that the pLasso and wpLasso have better predictive ability (having smaller PPS and CRPS) than the others, and the non-adaptive predictive Lasso npLasso also works better than the Lasso and even the aLasso. Furthermore, in terms of MSE, both pLasso and wpLasso give better and

more stable coefficent estimation. As one may expect for predictively motivated methods, models selected by the pLasso and wpLasso are less sparse than selected by the Lasso and aLasso. In order to better compare the behaviour of the methods over the replications, we plot in Figure 3.1 boxplots for the case $n=200$ and $\sigma=1$.

In our second simulation study, design points $\mathbf{x}_j$ are simulated from a multivariate t-distribution with degrees of freedom being 1.5. By doing so, we intend to simulate situations in which some predictors have high leverage points, i.e., their distributions have long tails. The simulation result is presented in Table 3.4. As one may expect, the wpLasso works better and more stable than the others because the variance is modeled to vary in proportion to the true predictive variance. Boxplots of the measures over replications for the case $n=200$, $\sigma=1$ are given in Figure 3.2.

In our last simulation study, we try a high-dimensional example. We consider the linear model (3.31) with $p=100$ and most of the coefficients are zero except $\beta_j=5$, $j=10,20,...,100$. The result reported in Table 3.5 suggests that the pLasso and wpLasso compare favourably with the others in this example. Boxplots for the case $n=200$, $\sigma=1$ are given in Figure 3.3.

A simulation study for logistic regression. We simulate independent observations from Bernoulli distributions with probabilities of success

$$\mu_i = P(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\exp(2 + \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(2 + \mathbf{x}_i'\boldsymbol{\beta})}$$

where the design points $\mathbf{x}_i$ are generated from the normal distributions as in the previous example. We consider two cases: a small p case with $\boldsymbol{\beta} = (3,\ 1.5,\ 0.5,\ 0.5,\ 0,\ 0,\ 0,\ 0)^\top$ and a large p case with most of β_j are zero except the first four entries are 3, 1.5, 0.5 and 0.5. We use the Gelman et al. prior and the plug-in method discussed earlier for estimating the predictions $\bar{y}_i^f$ (using MCMC would give a more accurate estimation but

Table 3.3: Simulation result for linear regression: small-p and normal predictors. The numbers in parentheses are standard deviations.

$n_T = n_P$	σ	measure	Lasso	aLasso	npLasso	pLasso	wpLasso
50	1	PPS	0.77 (0.19)	0.72 (0.19)	0.59 (0.13)	0.57 (0.13)	0.57 (0.12)
		MSE	0.59 (0.36)	0.56 (0.45)	0.23 (0.18)	0.19 (0.16)	0.18 (0.16)
		NZE	4.10 (0.82)	5.02 (0.64)	2.47 (1.17)	3.42 (1.05)	3.57 (1.03)
		CRPS	0.73 (0.12)	0.69 (0.11)	0.61 (0.07)	0.60 (0.06)	0.60 (0.07)
	3	PPS	1.86 (0.16)	1.86 (0.17)	1.70 (0.14)	1.70 (0.14)	1.69 (0.13)
		MSE	5.09 (2.68)	5.72 (3.16)	2.16 (1.68)	2.17 (1.81)	2.12 (1.83)
		NZE	5.97 (0.90)	6.64 (0.58)	3.40 (1.56)	4.22 (1.38)	4.46 (1.37)
		CRPS	2.17 (0.32)	2.17 (0.35)	1.83 (0.20)	1.83 (0.21)	1.83 (0.22)
100	1	PPS	0.68 (0.11)	0.65 (0.11)	0.54 (0.08)	0.54 (0.08)	0.54 (0.08)
		MSE	0.35 (0.17)	0.36 (0.23)	0.10 (0.07)	0.08 (0.07)	0.08 (0.07)
		NZE	3.82 (0.64)	4.63 (0.57)	2.18 (1.14)	3.23 (0.94)	3.17 (1.10)
		CRPS	0.67 (0.07)	0.65 (0.07)	0.59 (0.05)	0.58 (0.05)	0.58 (0.05)
	3	PPS	1.77 (0.10)	1.76 (0.10)	1.64 (0.08)	1.64 (0.08)	1.64 (0.08)
		MSE	3.06 (1.41)	3.54 (1.78)	1.01 (0.76)	0.93 (0.79)	0.92 (0.83)
		NZE	5.60 (0.84)	6.39 (0.69)	2.94 (1.43)	3.84 (1.20)	3.90 (1.36)
		CRPS	2.00 (0.20)	1.98 (0.20)	1.76 (0.13)	1.75 (0.13)	1.75 (0.13)
200	1	PPS	0.62 (0.08)	0.61 (0.07)	0.53 (0.06)	0.52 (0.06)	0.52 (0.06)
		MSE	0.19 (0.09)	0.22 (0.10)	0.06 (0.04)	0.05 (0.04)	0.05 (0.04)
		NZE	3.84 (0.39)	4.28 (0.48)	2.24 (1.22)	3.28 (0.94)	3.25 (1.09)
		CRPS	0.64 (0.05)	0.63 (0.05)	0.58 (0.03)	0.58 (0.03)	0.58 (0.03)
	3	PPS	1.72 (0.07)	1.71 (0.07)	1.62 (0.05)	1.62 (0.05)	1.62 (0.05)
		MSE	1.90 (0.84)	2.30 (1.10)	0.47 (0.34)	0.40 (0.34)	0.41 (0.35)
		NZE	5.48 (0.74)	6.29 (0.61)	2.53 (1.32)	3.59 (1.15)	3.57 (1.31)
		CRPS	1.90 (0.14)	1.89 (0.13)	1.73 (0.09)	1.73 (0.09)	1.73 (0.09)

Table 3.4: Simulation result for linear regression: the small-p with long-tailed t-distribution predictors. The numbers in parentheses are standard deviations.

$n_T = n_P$	σ	measure	Lasso	aLasso	npLasso	pLasso	wpLasso
50	1	PPS	2.58 (9.04)	1.89 (2.67)	1.29 (3.52)	1.05 (1.81)	0.66 (0.14)
		MSE	0.21 (0.19)	0.21 (0.33)	0.11 (0.11)	0.10 (0.11)	0.09 (0.12)
		NZE	3.07 (0.84)	3.75 (0.89)	1.65 (1.17)	3.14 (1.05)	3.29 (1.12)
		CRPS	0.90 (0.38)	0.91 (0.54)	0.74 (0.25)	0.71 (0.22)	0.70 (0.22)
	3	PPS	4.15 (12.03)	3.97 (11.47)	3.00 (6.47)	2.81 (6.13)	1.74 (0.16)
		MSE	1.36 (1.11)	1.71 (1.56)	0.80 (0.71)	0.77 (0.69)	0.78 (0.72)
		NZE	4.06 (1.32)	5.12 (1.11)	2.22 (1.33)	3.45 (1.14)	3.71 (1.24)
		CRPS	2.65 (1.20)	2.69 (1.19)	2.17 (0.73)	2.14 (0.69)	2.15 (0.70)
100	1	PPS	1.53 (3.26)	1.74 (4.66)	0.82 (0.81)	0.76 (0.60)	0.61 (0.15)
		MSE	0.07 (0.05)	0.09 (0.12)	0.04 (0.05)	0.03 (0.06)	0.03 (0.06)
		NZE	3.25 (0.99)	3.86 (0.64)	1.53 (1.19)	3.10 (1.03)	3.16 (1.12)
		CRPS	0.74 (0.16)	0.76 (0.24)	0.66 (0.13)	0.64 (0.12)	0.64 (0.12)
	3	PPS	5.95 (26.29)	7.77 (41.60)	2.34 (2.45)	2.06 (1.58)	1.67 (0.09)
		MSE	0.51 (0.40)	0.58 (0.43)	0.26 (0.21)	0.25 (0.22)	0.25 (0.22)
		NZE	3.41 (1.09)	4.61 (0.92)	2.00 (1.17)	3.18 (0.98)	3.37 (1.01)
		CRPS	2.33 (1.05)	2.41 (1.23)	1.95 (0.36)	1.91 (0.30)	1.93 (0.38)
200	1	PPS	1.31 (3.91)	1.41 (3.62)	0.65 (0.32)	0.64 (0.33)	0.55 (0.08)
		MSE	0.02 (0.02)	0.03 (0.05)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
		NZE	3.22 (0.89)	3.78 (0.46)	1.39 (1.14)	2.90 (1.03)	3.13 (1.13)
		CRPS	0.65 (0.11)	0.66 (0.16)	0.60 (0.05)	0.60 (0.05)	0.60 (0.05)
	3	PPS	2.59 (3.52)	3.89 (12.03)	1.81 (0.87)	2.15 (4.38)	1.63 (0.05)
		MSE	0.26 (0.22)	0.30 (0.31)	0.12 (0.10)	0.12 (0.10)	0.11 (0.10)
		NZE	3.26 (0.93)	4.08 (0.72)	1.93 (1.03)	3.17 (0.83)	3.31 (0.93)
		CRPS	2.01 (0.35)	2.07 (0.61)	1.80 (0.15)	1.80 (0.24)	1.80 (0.23)

Table 3.5: Simulation result for linear regression: the large-p with normal predictors. The numbers in parentheses are standard deviations.

n_T	σ	measure	Lasso	aLasso	npLasso	pLasso	wpLasso
50	1	PPS	3.08 (3.32)	1.96 (3.49)	2.66 (3.16)	2.64 (8.81)	2.56 (0.24)
		MSE	14.35 (36.79)	10.45 (35.60)	6.01 (15.99)	5.34 (18.43)	5.67 (21.01)
		NZE	75.21 (5.98)	86.06 (4.59)	64.20 (9.07)	73.36 (9.98)	80.15 (7.58)
		CRPS	1.67 (1.55)	1.30 (1.49)	1.34 (0.79)	1.18 (0.95)	1.22 (1.15)
	3	PPS	3.60 (3.14)	3.54 (12.83)	8.47 (10.16)	6.72 (9.16)	3.29 (0.40)
		MSE	58.20 (63.86)	42.42 (67.93)	38.16 (41.15)	30.88 (34.58)	35.86 (50.35)
		NZE	78.66 (7.11)	87.30 (4.97)	67.45 (8.84)	75.12 (8.54)	81.29 (7.01)
		CRPS	4.39 (2.00)	3.63 (2.20)	3.99 (1.50)	3.63 (1.46)	3.97 (1.95)
100	1	PPS	2.45 (4.53)	1.43 (0.68)	1.04 (1.65)	1.04 (1.20)	1.01 (0.14)
		MSE	1.38 (1.47)	13.53 (19.42)	1.23 (3.36)	1.17 (2.60)	0.49 (1.94)
		NZE	54.74 (23.75)	85.69 (8.66)	61.06 (18.93)	71.14 (21.22)	83.50 (12.80)
		CRPS	0.86 (0.24)	1.75 (1.26)	0.76 (0.30)	0.75 (0.30)	0.67 (0.21)
	3	PPS	2.79 (2.99)	2.23 (0.37)	3.11 (4.37)	3.47 (4.52)	1.95 (0.39)
		MSE	8.79 (6.60)	27.02 (25.73)	9.88 (17.38)	10.86 (19.11)	4.01 (8.84)
		NZE	64.06 (20.88)	86.61 (7.98)	59.23 (19.65)	68.97 (21.16)	81.70 (12.46)
		CRPS	2.38 (0.47)	3.19 (1.13)	2.36 (0.78)	2.36 (0.95)	2.14 (0.49)
200	1	PPS	0.95 (0.12)	0.57 (0.08)	0.60 (0.06)	0.57 (0.06)	0.57 (0.04)
		MSE	1.50 (0.55)	0.17 (0.13)	0.27 (0.11)	0.18 (0.13)	0.09 (0.07)
		NZE	89.19 (1.09)	89.49 (0.66)	65.66 (10.23)	78.05 (9.08)	86.15 (4.58)
		CRPS	0.88 (0.10)	0.60 (0.05)	0.62 (0.04)	0.60 (0.04)	0.59 (0.03)
	3	PPS	1.97 (0.11)	1.70 (0.08)	1.72 (0.08)	1.69 (0.11)	1.66 (0.05)
		MSE	9.98 (3.18)	1.81 (1.32)	2.46 (0.96)	1.82 (1.42)	0.91 (0.59)
		NZE	88.88 (1.30)	89.64 (0.64)	66.25 (10.14)	76.61 (10.69)	85.17 (4.55)
		CRPS	2.46 (0.25)	1.86 (0.15)	1.88 (0.12)	1.83 (0.14)	1.80 (0.11)

Figure 3.1: Boxplots of the performance measures over replications in linear regression: the small p case with normal predictors, $n = 200$ and $\sigma = 1$.

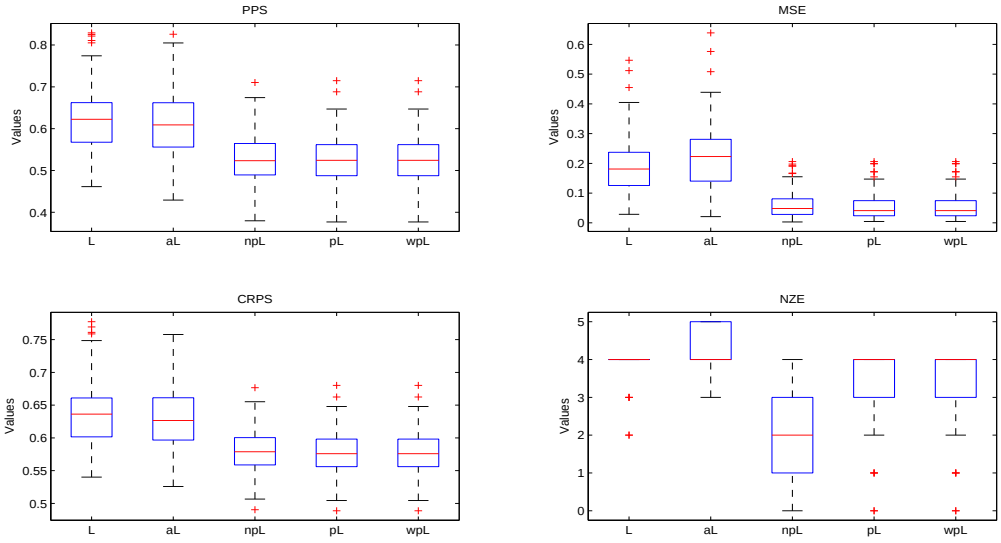


Figure 3.2: Boxplots of the performance measures over replications in linear regression: the small p case with long-tailed predictors, $n = 200$ and $\sigma = 1$.

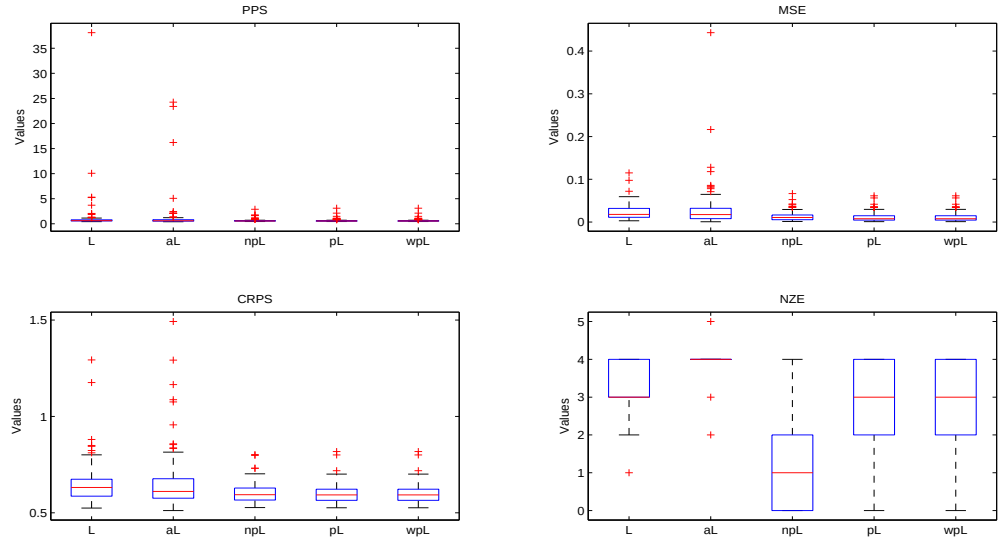
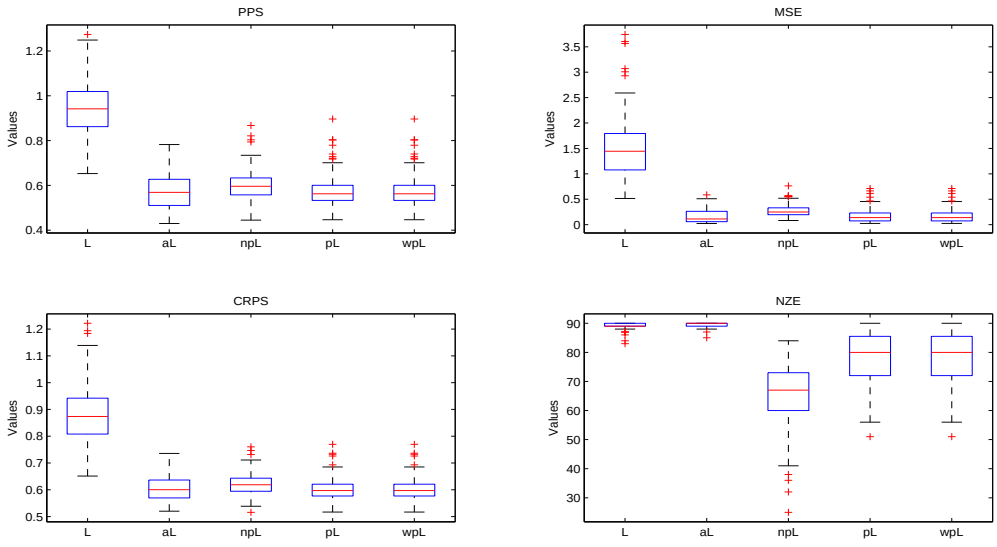


Figure 3.3: Boxplots of the performance measures over replications in linear regression: the large p case with normal predictors, $n = 200$ and $\sigma = 1$.

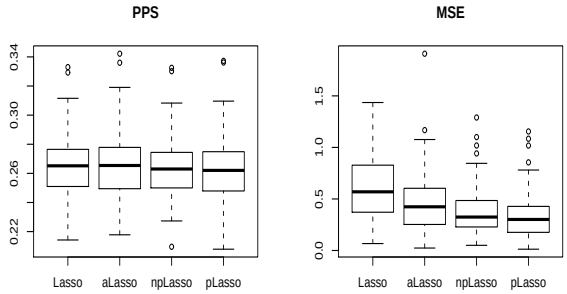


may be time consuming in simulation when performance measures are to be averaged over many replications). The simulation results are summarized in Tables 3.6 and 3.7 with various sample sizes. Boxplots for two cases are given in Figures 3.4 and 3.5. As shown, both pLasso methods outperform the Lasso methods in terms of both PPS, CRPS and MSE. Furthermore, the aLasso seems to be unstable and work poorly in the large p case with small number of observations. The pLasso with the regularization prior of Gelman et al. [2008] works surprisingly well in this example.

Application: Linear regression - predicting percent body fat. Percentage of body fat is one important measure of health, which can be accurately estimated by underwater weighing techniques [Bailey, 1994]. These techniques often require special equipment and are sometimes not convenient, thus fitting percent body fat to simple body measurements is a convenient way to predict body fat. Johnson [1996] introduced a dataset in which

Table 3.6: Simulation result for logistic regression: the small p case.

$n_T = n_P$	measure	Lasso	aLasso	npLasso	pLasso
100	PPS	0.28 (0.05)	0.28 (0.07)	0.27 (0.05)	0.27 (0.05)
	MSE	3.52 (2.19)	5.20 (29.81)	2.45 (1.12)	1.95 (1.26)
	NZE	1.95 (1.47)	3.11 (1.08)	0.65 (1.41)	2.41 (1.40)
	CRPS	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)
200	PPS	0.27 (0.03)	0.27 (0.03)	0.27 (0.03)	0.27 (0.03)
	MSE	1.40 (0.76)	1.09 (0.81)	0.95 (0.48)	0.90 (0.52)
	NZE	1.63 (1.42)	3.01 (1.32)	0.40 (1.21)	2.13 (1.42)
	CRPS	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)
500	PPS	0.26 (0.02)	0.26 (0.02)	0.26 (0.02)	0.26 (0.02)
	MSE	0.60 (0.33)	0.47 (0.29)	0.38 (0.23)	0.35 (0.22)
	NZE	2.07 (1.24)	3.82 (1.12)	1.09 (0.97)	2.66 (1.20)
	CRPS	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)



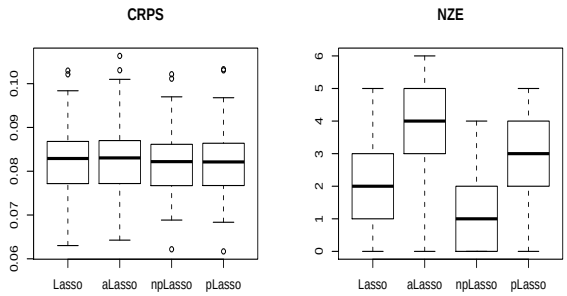
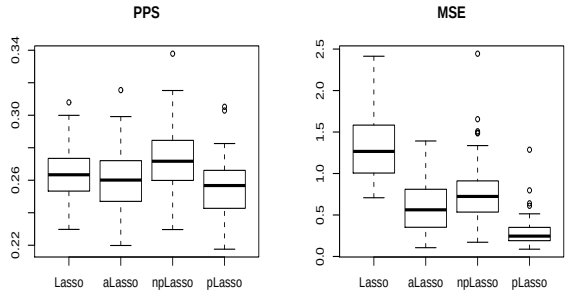


Figure 3.4: Boxplots of the performance measures over replications in logistic regression: the small p case with $n=500$



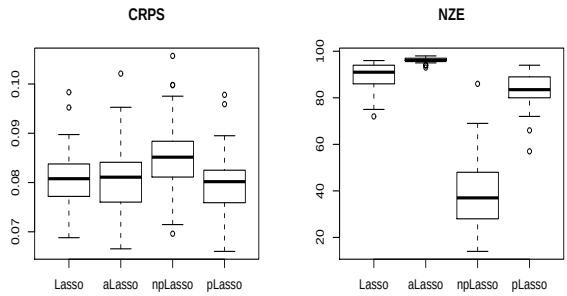


Figure 3.5: Boxplots of the performance measures over replications in logistic regression: the large p case with $n=1000$

percent body fat and 13 simple body measurements (such as weight, height and abdomen circumference) are recorded for 252 men. After omitting observations 39 (because a weight value of 363.15 pounds is unusually large), 42 (because a height value of 29.5 inches is unreasonable), and 182 (because the response value is 0), we obtain a dataset of size 249.

We are concerned with the problem of constructing a model that predicts the response from the covariates. Following Hoeting et al. [1999], we use a linear regression model. The primary goal is prediction accuracy for future observations; besides this, parsimony is another important objective, since a simple model is preferred for the sake of scientific insight into the $x-y$ relationship.

Table 3.7: Simulation result for logistic regression: the large p case.

$n_T = n_P$	measure	Lasso	aLasso	npLasso	pLasso
100	PPS	0.33 (0.04)	0.56 (0.21)	0.32 (0.04)	0.31 (0.05)
	MSE	4.56 (1.26)	17.5 (8.90)	4.37 (1.24)	2.63 (1.08)
	NZE	91.8 (5.27)	69.5 (6.22)	89.2 (7.67)	96.5 (2.15)
	CRPS	0.10 (0.02)	0.14 (0.04)	0.10 (0.02)	0.09 (0.02)
500	PPS	0.28 (0.02)	0.69 (0.43)	0.27 (0.03)	0.26 (0.03)
	MSE	2.10 (0.59)	15.9 (18.1)	1.13 (0.43)	0.69 (0.36)
	NZE	89.6 (6.92)	45.8 (30.5)	60.8 (17.7)	82.1 (9.22)
	CRPS	0.09 (0.01)	0.12 (0.03)	0.08 (0.01)	0.08 (0.01)
1000	PPS	0.26 (0.01)	0.26 (0.02)	0.27 (0.02)	0.25 (0.02)
	MSE	1.32 (0.40)	0.59 (0.28)	0.79 (0.41)	0.29 (0.20)
	NZE	89.5 (5.53)	96.1 (1.02)	38.9 (14.9)	83.5 (6.98)
	CRPS	0.08 (0.01)	0.08 (0.01)	0.08 (0.01)	0.07 (0.01)

Using the full dataset, the aLasso, pLasso and wpLasso estimates of $\boldsymbol{\beta}$ are given in Table 3.8. The abbreviations "al", "pl" and "wpl" stand for aLasso, pLasso and wpLasso, respectively. These methods simultaneously do parameter estimation and variable selection, because some of the estimated coefficients are exact zero. Recall that the goals at which the methods aim are somewhat different: pLasso and wpLasso have a more explicit predictive motivation; besides, the wpLasso in some cases is somewhat more realistic in the sense that it allows the variances to vary in proportion to the predictive variance of the full model.

We now examine the predictive performance of these three procedures. To this end, we split the dataset into two parts: the first 125 observations are used as the training set

Table 3.8: Predicting percent body fat.

	full data			case I			case II			case III		
	al	pl	wpl	al	pl	wpl	al	pl	wpl	al	pl	wpl
	-18.0	6.79	-0.18	-14.8	2.88	-0.28	-15.7	-2.95	-4.59	-23.3	-0.61	-3.87
1	0	0.06	0.04	0.02	0.09	0.08	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	-0.20	-0.29	-0.27	-0.26	-0.40	-0.39	0	-0.17	-0.14	0	-0.24	-0.22
4	0	-0.30	-0.11	0	-0.24	-0.17	0	0	0	0	-0.34	-0.25
5	0	-0.09	0	0	0	0	0	0	0	0	0	0
6	0.55	0.78	0.68	0.55	0.70	0.68	0.38	0.66	0.66	0.45	0.69	0.69
7	0	-0.09	0	0	-.09	0	0	0	0	0	0	0
8	0	0.09	0	0	0.16	0.08	0	0	0	0	0	0
9	0	0	0	0	0.09	0	0	0	0	0	0	0
10	0	0.09	0	0	0.22	0.17	0	-0.39	-0.43	0	-0.04	0
11	0	0.13	0.04	0	0	0	0	0.10	0.10	0	0.20	0.20
12	0	0.19	0	0	0	0	0	0	0	0	0.19	0.07
13	0	-1.62	-1.31	0	-1.34	-1.20	0	-1.16	-1.15	0	-1.44	-1.35
PPS				1.95	1.93	1.93	2.11	1.91	1.90	2.08	1.96	1.95
CRPS				2.44	2.36	2.37	3.00	2.35	2.35	2.94	2.34	2.26

D, the remaining observations are used as the prediction set D^P. The aLasso, pLasso and wpLasso estimates and their PPS are given in Table 3.8 (case I). As a second examination, the first 125 observations are used as the prediction set D^P, the remaining observations are used as the training set D. For a third examination, we randomly split the full dataset into two (roughly) equal parts which serve as the training and prediction sets. The coefficient estimates, PPS and CRPS are summarized in Table 3.8. As one may expect for predictively motivated methods, the variables selected by pLasso and wpLasso in general contain those selected by aLasso, i.e., the models selected by pLasso and wpLasso are bigger than the one selected by aLasso. In all cases, the pLasso and wpLasso show a better predictive performance over the aLasso. Indeed, the PPS of the aLasso, pLasso and wpLasso averaged over such 50 random partitions are 2.055, 1.998, 1.924, respectively and the averaged CRPS are 2.703, 2.385, 2.370, respectively. It seems that modelling the variances to vary in proportion to the predictive variance of the full model is appropriate in this example, because the wpLasso has a similar or better predictive performance compared with the pLasso.

Application: Logistic regression - the spambase data. We consider in this example an application of the predictive Lasso in the logistic regression framework with many predictors and instances. We consider the spam email data set created by Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt at the Hewlett-Packard Labs. The data set consists of 4061 messages, each has been already classified as email or spam together with 57 attributes (predictors) which are relative frequencies of commonly occurring words. The goal is to design a spam filter that could filter out spam before clogging the users' mailboxes. Our goal as usual is to construct a parsimonious model with a good prediction accuracy.

With a large number of predictors and observations, using MCMC may be time con-

suming so that we use the plug in method discussed earlier. To access the performance of the aLasso and pLasso methods, we randomly split the data set into two parts (training set and prediction set) and record performance measures PPS, CRPS and NZE across such 50 random partitions. The averaged PPS, CRPS and NZE for the aLasso are 0.261, 0.072, 27.2 and for the pLasso are 0.251, 0.067, 25.1, respectively. The pLasso gives a better predictive performance overall while selecting roughly 2 predictors more than the aLasso.

Chapter 4

Some results on variable selection

While the last two chapters discussed in turn two general procedures for model selection, this chapter focuses mainly on variable selection. We shall present two novel algorithms for variable selection in two broad frameworks. In Section 4.1, we look at the regularization approaches like the Lasso and its variants (adaptive Lasso, group Lasso, etc.) from a Bayesian point of view. We propose the Bayesian adaptive Lasso (BaLasso) for variable selection and group selection in a unified framework including GLMs, Cox's model and many others. The BaLasso is adaptive to the signal level in the sense that it adopts different shrinkage for different coefficients. Furthermore, our Bayesian formulation enables us to incorporate prior information on grouping and hierarchical structures present within the variables.

We then in Section 4.2 consider the problem of variable selection for heteroscedastic linear regression (i.e., the variance is allowed to vary with covariates) and propose a novel fast greedy search algorithm for variable seletion in both mean and variance model using a variational approximation method. Table 4.14 gives a brief summary of some of the commonly used variable selection methods as well as the methods proposed in this thesis.

This chapter is based on joint works with David Nott and Chenlei Leng [Leng et al., 2010, Nott et al., 2010, Tran et al., 2011].

4.1 Bayesian adaptive Lasso

Let us start the discussion with the usual linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

As is usual in regression analysis, our major interests are to estimate $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^\top$, to identify its important covariates and to make accurate predictions. Without loss of generality, we assume $\mathbf{y}$ and X are centered so that the intercept is zero and can be omitted from the model.

The Lasso of Tibshirani [1996], formulated in the penalized likelihood framework, minimizes the residual sum of squares with a constraint on the ℓ_1 norm of $\boldsymbol{\beta}$. Formally, the Lasso solves

$$\min_{\beta} (\mathbf{y} - X\boldsymbol{\beta})^\top(\mathbf{y} - X\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{4.1}$$

where $\lambda > 0$ is the tuning parameter controlling the amount of penalty. The least angle regression (LARS) algorithm provides fast implementation of the Lasso solution [Efron et al., 2004, Osborne et al., 2000]. Furthermore, the Lasso can be model selection consistent provided that the so-called irrepresentable condition on the design matrix is satisfied and that λ is chosen judiciously [Zhao and Yu, 2006].

However, if this condition does not hold, Zou [2006] and Zhao and Yu [2006] showed that the Lasso chooses a wrong model with non-vanishing probability, regardless of the sample size and how λ is chosen. The condition is almost necessary and sufficient for model selection consistency of the Lasso, which requires that the predictors not in the

model are not representable by predictors in the true model. This condition can be easily violated due to the collinearity between the predictors. To address this issue, Zou [2006] proposed to use the adaptive Lasso (aLasso) which gives consistent model selection. The final inference procedure, thereafter, is based on a single selected model. This may bring undesirable risk properties as discussed by Poetscher and Leeb [2009].

The Lasso estimator can be interpreted as the posterior mode using normal likelihood and iid Laplace prior for $\boldsymbol{\beta}$ [Tibshirani, 1996]. The first explicit treatment of the Bayesian Lasso (BLasso), which exploits model inference via posterior distributions, has been proposed by Park and Casella [2008]. Griffin and Brown [2010] proposed an extension of this approach but focused on finding posterior modes via an EM algorithm which does not provide exploration of the posterior distribution.

Although the Lasso was originally designed for variable selection, the BLasso loses this attractive property, not setting any of the coefficients to zero. A post hoc thresholding rule may overcome this difficulty but it brings the problem of threshold selection. Alternatively, Kyung et al. [2010] recommended to use the credible interval on the posterior mean. Although it gives variable selection, this suggestion fails to explore the uncertainty in the model space.

This work is motivated by the need to explore model uncertainty and to achieve parsimony. With these objectives, we consider the following adaptive Lasso estimator:

$$\min_{\beta} (\mathbf{y} - X\boldsymbol{\beta})^{\top}(\mathbf{y} - X\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{4.2}$$

where different penalty parameters are used for the regression coefficients. Naturally, for the unimportant covariates, we should put larger penalty parameters λ_j on their corresponding coefficients. This strategy was proposed by Zou [2006] by using some preliminary estimates of $\boldsymbol{\beta}$ such as the least squares estimate $\hat{\boldsymbol{\beta}}^{\text{ols}}$ and modifying λ_j as $\lambda/|\hat{\beta}_j^{\text{ols}}|$. Our

treatment is completely different and is motivated by the following arguments. Suppose tentatively that we have a posterior distribution on $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_p)^\top$. By drawing random samples from this distribution and plugging these into (4.2), we can solve for $\boldsymbol{\beta}$ using fast algorithms developed for the Lasso [Efron et al., 2004, Figueiredo et al., 2007] and subsequently obtain an array of (sparse) models. These models can be used not only for exploring model uncertainty, but also for prediction with a variety of methods akin to Bayesian model averaging. Since there are p tuning parameters, a hierarchical model is proposed to alleviate the problem of estimating many parameters. We develop an efficient Gibbs sampler for posterior inference.

We further propose a unified framework for variable/group selection using flexible penalties. This unified framework encompasses generalized linear models, Cox's model and other parametric models as special cases. We outline novel applications of the BaLasso when structured penalties are present, for example, grouped variable selection [Yuan and Lin, 2006] and variable/group selection with a prior hierarchical structure [Zhao et al., 2009].

A Matlab implementation of our method is available from the author's homepage. The software is general enough to deal with most of the models encountered in practice. Systematic simulation studies and real-data analysis strongly support the use of our method.

The rest of this section is organized as follows. The BaLasso in linear regression is presented in Section 4.1.1 and is extended in Section 4.1.4 to a unified framework with structured penalties. Section 4.1.2 discusses model selection and Bayesian model averaging. In Section 4.1.3, the finite sample performance of the BaLasso is illustrated via simulation studies, and analysis of real datasets.

4.1.1 Bayesian adaptive Lasso for linear regression

The ℓ_1 penalty corresponds to a conditional Laplace prior [Tibshirani, 1996] as

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}},$$

which can be represented as a scale mixture of normals with an exponential mixing density [Andrews and Mallows, 1974]

$$\frac{\lambda}{2} e^{-\lambda|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{\lambda^2}{2} e^{-\lambda^2 z/2} ds.$$

This motivates the following hierarchical BLasso model [Park and Casella, 2008]

$$\begin{aligned} \mathbf{y}|X, \boldsymbol{\beta}, \sigma^2 &\sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n) \\ \boldsymbol{\beta}|\sigma^2, \tau_1^2, ..., \tau_p^2 &\sim N_p(\mathbf{0}_p, \sigma^2 D_\tau) \\ D_\tau &= \text{diag}(\tau_1^2, ..., \tau_p^2) \end{aligned} \tag{4.3}$$

with the following priors on σ^2 and $\boldsymbol{\tau} = (\tau_1^2, ..., \tau_p^2)^\top$:

$$\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2 \tag{4.4}$$

for $\sigma^2 > 0$ and $\tau_1^2, ..., \tau_p^2 > 0$. Park and Casella [2008] suggested to use the improper prior $\pi(\sigma^2) \propto 1/\sigma^2$ to model the error variance.

As discussed in the introduction, the Lasso uses the same shrinkage for every coefficient and may not be consistent for certain design matrices in terms of model selection. This motivates us to replace (4.4) in the hierarchical structure by a more adaptive penalty

$$\sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda_j^2}{2} e^{-\lambda_j^2 \tau_j^2/2} d\tau_j^2. \tag{4.5}$$

The major difference of this formulation is to allow different λ_j^2, one for each coefficient. Intuitively, if a small penalty is applied to those covariates that are important and a large

penalty is applied to those which are unimportant, the Lasso estimate, as the posterior mode, can be model selection consistent [Zou, 2006]. Indeed, as we will see below and in later numerical experiments, in the posterior distribution, the λ_j's for zero β_j's will be much larger than those λ_j's for nonzero β_j's.

The Gibbs sampling scheme follows Park and Casella [2008]. For Bayesian inference, the full conditional distribution of $\boldsymbol{\beta}$ is multivariate normal with mean $A^{-1}X^{\top}\mathbf{y}$ and variance $\sigma^2 A^{-1}$, where $A = X^{\top}X + D_{\tau}^{-1}$. The full conditional for σ^2 is inverse-gamma with shape parameter $(n-1)/2+p/2$ and scale parameter $(\mathbf{y}-X\boldsymbol{\beta})^{\top}(\mathbf{y}-X\boldsymbol{\beta})/2+\boldsymbol{\beta}^{\top}D_{\tau}^{-1}\boldsymbol{\beta}/2$ and τ_1^2,...,τ_p^2 are conditionally independent, with $1/\tau_j^2$ conditionally inverse-Gaussian with parameters

$$\tilde{\mu}_j = \frac{\lambda_j \sigma}{|\beta_j|} \text{ and } \tilde{\lambda}_j = \lambda_j^2$$

where the inverse-Gaussian density is given by

$$f(x) = \sqrt{\tilde{\lambda}} 2\pi x^{-3/2} \exp\{-\frac{\tilde{\lambda}(x-\tilde{\mu}_j)^2}{2(\tilde{\mu})^2 x}\},\ x > 0.$$

As observed in Park and Casella [2008], the Gibbs sampler with block updating of $\boldsymbol{\beta}$ and $(\tau_1^2,...,\tau_p^2)$ is very fast.

Choosing the Bayesian adaptive Lasso parameters

We discuss here two approaches for choosing the BaLasso parameters λ_j in the Bayesian framework.

The first one is the empirical Bayes (EB) method which aims at estimating the λ_j via the marginal maximum likelihood. A natural choice is to estimate the hyper-parameters λ_j by marginal maximum likelihood. However, in our framework, the marginal likelihood for the λ_j is not available in closed form. To deal with this problem, Casella [2001] proposed a multi-step approach based on an EM algorithm with the expectation in the E-step being

approximated by the average from the Gibbs sampler. The updating rule then for λ_j is easily seen to be

$$\lambda_j^{(k)} = \sqrt{\frac{2}{E_{\lambda_j^{(k-1)}}(\tau_j^2|\mathbf{y})}} \tag{4.6}$$

where $\lambda_j^{(k)}$ is the estimate of λ_j at the kth stage and the expectation $E_{\lambda_j^{(k-1)}}(.)$ is approximated by the average from the Gibbs sampler with the hyper-parameters set to $\lambda_j^{(k-1)}$.

Casella's method may be computationally expensive because many Gibbs sampler runs are needed. Atchade [2009] proposed a single-step approach based on stochastic approximation which can obtain the MLE of the hyper-parameters using a single Gibbs sampler run. In our framework, making the transformation $\lambda_j = e^{s_j}$, the updating rule for the hyper-parameters s_j can be seen as [Atchade, 2009, Algorithm 3.1]

$$s_j^{(n+1)} = s_j^{(n)} + a_n(2 - e^{2s_j^{(n)}}\tau_{n+1,j}^2)$$

where $s_j^{(n)}$ is the value of s_j at the nth iteration, $\tau_{n,j}^2$ is the nth Gibbs sample of τ_j^2, and $\{a_n\}$ is a sequence of step-sizes such that

$$a_n \searrow 0,\ \sum a_n = \infty,\ \sum a_n^2 < \infty.$$

In the following simulation, a_n is set to $1/n$. Strictly speaking, choosing a proper a_n is an important problem of stochastic approximation which is beyond the scope of our discussion in this thesis. In practice, a_n is often set after a few trials by justifying the convergence of iterations graphically.

The second method for estimating the BaLasso parameters uses hyper priors on λ_j which enable posterior inference on these shrinkage parameters. The λ_j themselves can be treated as random variables and join the Gibbs updating by using an appropriate prior on λ_j^2. Here for simplicity and numerical tractability, we take the following gamma prior

[Park and Casella, 2008]

$$\pi(\lambda_j^2) = \frac{\delta^r}{\Gamma(r)}(\lambda_j^2)^{r-1}e^{-\delta\lambda_j^2}. \tag{4.7}$$

The advantage of using such a prior is that the Gibbs sampling algorithm can be easily implemented. More specifically, when this prior is used, the full conditional of λ_j^2 is gamma with shape parameter $1+r$ and rate parameter $\tau_j^2+\delta$. This specification allows λ_j^2 to join the other parameters in the Gibbs sampler. Although the number of the penalty parameters λ_j has increased to p in the BaLasso from a single parameter in the Lasso, the fact that the same prior is used on these parameters greatly reduces the degrees of freedom in specifying the prior.

As a first choice, we can fix hyper-parameters r and δ to some small values in order to get a flat prior. Alternatively, we can fix r and use an empirical Bayes approach where δ is estimated. The updating rule for δ [Casella, 2001] can be seen as

$$\delta^{(k)} = \frac{pr}{\sum_{j=1}^{p} E_{\delta^{(k-1)}}(\lambda_j^2|\mathbf{y})}.$$

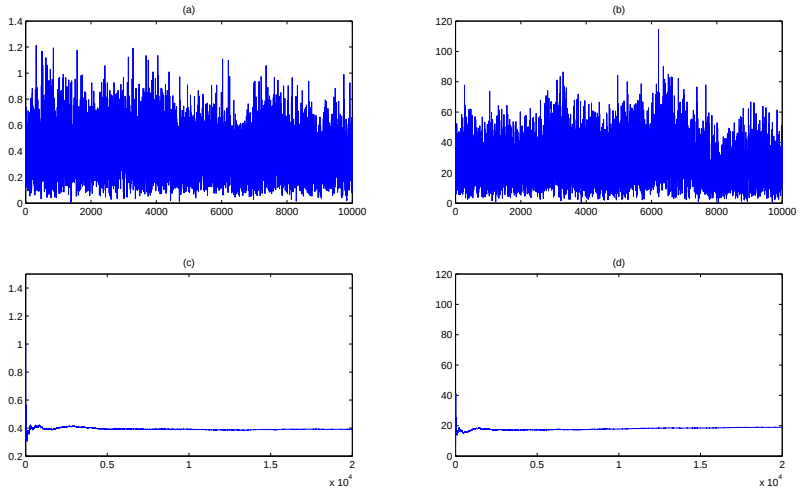
Theoretically, we need not worry so much about how to select r because parameters that are deeper in the hierarchy have less effect on inference [Lehmann and Casella, 1998, p.260]. In our simulation study and data analysis, we use $r=.1$ which gives a fairly flat prior and stable results. In our experience, both the EB and full treatment methods for estimating λ_j often give very similar results. In what follows we focus on the latter only.

Adaptive shrinkage

By allowing different λ_j^2, adaptive shrinkage on the coefficients is possible. We demonstrate the adaptivity by a simple simulation in which a data set of size 50 is generated from the model

$$y=\beta_1 x_1+\beta_2 x_2+\sigma\epsilon$$

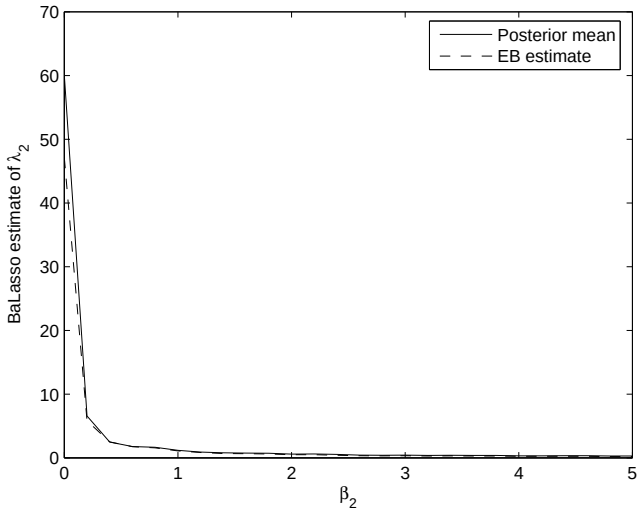
Figure 4.1: (a)-(b): Gibbs samples for λ_1 and λ_2, respectively. (c)-(d): Trace plots for $\lambda_1^{(n)}$ and $\lambda_2^{(n)}$ by Atchade's method.



with $\beta = (3,\ 0)'$, $\sigma = 1$, $\epsilon \sim N(0,1)$.

Because $\beta_1 \neq 0$, $\beta_2 = 0$ we expect that the EB and posterior estimate of λ_2 would be much larger than that of λ_1. As a result, a heavier penalty is put on β_2 such that β_2 is more likely to be shrunk to zero. This phenomenon is demonstrated graphically in Figure 4.1. Figure 4.1 (a)-(b) plot 10,000 Gibbs samples (after discarding 10,000 *burn-in* samples) for λ_1 and λ_2 (not λ_1^2, λ_2^2), respectively. The posterior distribution of λ_2 is central around a value of 22 which is much larger than .39, the posterior median of λ_1. Figure 4.1 (c)-(d) shows the trace plots of iterations $\lambda_1^{(n)}$, $\lambda_2^{(n)}$ from Atchade's method. Marginal maximum likelihood estimates of λ_1 and λ_2 are 0.39 and 19, respectively. In Figure 4.2 we plot EB and posterior mean estimates of λ_2 versus β_2 when β_2 varies from 0 to 5. Clearly, both the EB and the posterior estimates of λ_2 decrease as β_2 increases, which demonstrates that lighter penalty is applied for stronger signals.

Figure 4.2: Plots of the EB and posterior estimates of λ_2 versus β_2



4.1.2 Inference

Estimation and model selection

For the adaptive Lasso, the usual methods to choose the λ_j would be computationally demanding. From the Bayesian perspective, one can draw MCMC samples based on the BaLasso and get an estimated posterior quantity for $\boldsymbol{\beta}$. Like the original Bayesian Lasso, however, a full posterior exploration gives no sparse models and would fail as a model selection method. Here we take a hybrid Bayesian-frequentist point of view in which coefficient estimation and variable selection are simultaneously conducted by plugging in an estimate of $\boldsymbol{\lambda}$ into (4.2), where $\boldsymbol{\lambda}$ might be the marginal maximum likelihood estimator, posterior median or posterior mean. Hereafter these suggested strategies are abbreviated as BaLasso-EB, BaLasso-Median, and BaLasso-Mean, respectively.

With the presence of a posterior sample, we also propose another strategy for exploring model uncertainty. Let $\{\boldsymbol{\lambda}^{(s)}\}_{s=1}^{N}$ be Gibbs samples drawn from the hierarchical model (4.3), (4.5) and (4.7). For the sth Gibbs sample $\boldsymbol{\lambda}^{(s)} = (\lambda_1^{(s)},...,\lambda_p^{(s)})^\top$, we plug $\boldsymbol{\lambda}^{(s)}$ into

(4.2) and then record the frequencies of each variable being chosen out of N samples. The final chosen model consists of those variables whose frequencies are not less than 0.5. This strategy will be abbreviated as BaLasso-Freq. The chosen model is somewhat similar in spirit to the so-called median probability (MP) model proposed by Barbieri and Berger [2004] (see Section 3.1)

As we will see in Section 4.1.3, all of our proposed strategies have surprising improvement in terms of variable selection over the original Lasso and the adaptive Lasso.

Bayesian model averaging

When model uncertainty is present, making inferences based on a single model may be dangerous. Using a set of models helps to account for this uncertainty and can provide improved inference. As discussed in Section 3.1, Bayesian model averaging is widely used for prediction and generally provides better predictive performance than a chosen single model. For making inference via multiple models, we use the hierarchical model approach for estimating $\boldsymbol{\lambda}$ and refer to the following strategy as BaLasso-BMA.

Let $\Delta=(\mathbf{x}_\Delta,y_\Delta)$ be a future observation and $D=(X,\mathbf{y})$ be the past data. The posterior predictive distribution of Δ is given by

$$p(\Delta|D) = \int p(\Delta|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\lambda},D)d\boldsymbol{\beta}p(\boldsymbol{\lambda}|D)d\boldsymbol{\lambda}. \tag{4.8}$$

Suppose that we measure predictive performance via a logarithmic scoring rule [Good, 1952], i.e., if $g(\Delta|D)$ is some distribution we use for prediction then our predictive performance is measured by $\log g(\Delta|D)$ (where larger is better). Then for any fixed smoothing parameter vector $\boldsymbol{\lambda}_0$

$$E(\log p(\Delta|D) - \log p(\Delta|\boldsymbol{\lambda}_0,D)) = \int \log \frac{p(\Delta|D)}{p(\Delta|\boldsymbol{\lambda}_0,D)} p(\Delta|D)d\Delta$$

is nonnegative because the right hand side is the Kullback-Leibler divergence between $p(\Delta|D)$ and $p(\Delta|\boldsymbol{\lambda}_0,D)$. Hence prediction with $p(\Delta|D)$ is superior in this sense to prediction with $p(\Delta|\boldsymbol{\lambda}_0,D)$ with any choice of $\boldsymbol{\lambda}_0$.

Our hierarchical model (4.3), (4.5) and (4.7) offers a natural way to estimate the predictive distribution (4.8), in which the integral is approximated by the average from Gibbs samples of $\boldsymbol{\lambda}$. For example, in the case of point prediction for y_Δ with squared error loss, the ideal prediction is

$$E(y_\Delta|D) = \int \mathbf{x}_\Delta^\top E(\boldsymbol{\beta}|\boldsymbol{\lambda}, D)p(\boldsymbol{\lambda}|D)d\boldsymbol{\lambda} = \mathbf{x}_\Delta^\top E(\boldsymbol{\beta}|D),$$

where $E(\boldsymbol{\beta}|D)$ can be estimated by the mean of Gibbs samples for $\boldsymbol{\beta}$. Write $\hat{\boldsymbol{\beta}}_\lambda$ as the conditional posterior mode for $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$. One could approximate $\mathbf{x}_\Delta^\top E(\boldsymbol{\beta}|D)$ by replacing $E(\boldsymbol{\beta}|D)$ with the conditional posterior mode $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ for some fixed value $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$. However, this ignores uncertainty in estimating the penalty parameters. An alternative strategy is to replace $E(\boldsymbol{\beta}|D,\boldsymbol{\lambda})$ in the integral above with $\hat{\boldsymbol{\beta}}_\lambda$ and to integrate it out accordingly. This should provide a better approximation to the full Bayes solution than the approach which uses a fixed $\hat{\boldsymbol{\lambda}}$. In fact, we predict $E(y_\Delta|D)$ by $s^{-1}\sum_{i=1}^s \mathbf{x}_\Delta^\top \hat{\boldsymbol{\beta}}_{\lambda^{(i)}}$ where $\boldsymbol{\lambda}^{(i)}$, i=1,...,s, denote MCMC samples drawn from the posterior distribution of $\boldsymbol{\lambda}$. Note that this approach has advantages in interpretation over the fully Bayes' solution. By considering the models selected by the conditional posterior mode for different draws of $\boldsymbol{\lambda}$ from $p(\boldsymbol{\lambda}|\mathbf{y})$ we gain an ensemble of sparse models that can be used for interpretation. As will be seen in Section 4.1.3, when there is model uncertainty, BaLasso-BMA provides an ensemble of sparse models and may have better predictive performance than conditioning on a single fixed smoothing parameter vector $\boldsymbol{\lambda}$.

Table 4.1: Frequency of correctly-fitted models over 100 replications for Example 1.

n	σ	Lasso	aLasso	BaLasso-Freq	BaLasso-Median	BaLasso-Mean	BaLasso-EB
30	1	50	71	86	86	97	78
	3	17	8	35	34	18	39
60	1	66	76	81	79	100	83
	3	44	38	54	53	55	46
120	1	73	76	87	87	100	87
	3	58	55	81	81	97	86

4.1.3 Examples

In this section we study the proposed methods through numerical examples. These methods are also compared to the Lasso, aLasso and BLasso in terms of variable selection and prediction. We use the LARS algorithm of Efron et al. [2004] for the Lasso and the aLasso in which fivefold cross-validation is used to choose shrinkage parameters. In the adaptive Lasso, we either use the least squares estimate (Examples 1 and 2) or the Lasso estimate (Example 3) as the preliminary estimate. For the optimization problem (4.2), we use the gradient projection algorithm developed by Figueiredo et al. [2007].

Example 1 (simple example). We consider again model (2.40) of Tibshirani [1996]. We compare the performance of the proposed methods for model selection described above to that of the original Lasso and adaptive Lasso. The performance is measured by the frequency of correctly-fitted models over 100 replications. The simulation results summarized in Table 4.1 suggest that the proposed methods perform better than the Lasso and aLasso in model selection.

Example 2 (difficult example). For the second example, we use Example 1 in Zou [2006], for which the Lasso does not give consistent model selection, regardless of the

Table 4.2: Frequency of correctly-fitted models over 100 replications for Example 2.

n	σ	Lasso	aLasso	BaLasso-Freq	BaLasso-Median	BaLasso-Mean	BaLasso-EB
60	9	0	5	8	8	9	12
120	5	10	45	66	65	66	51
300	3	12	65	83	83	85	83
300	1	12	100	100	100	100	100

sample size and how the tuning parameter $\boldsymbol{\lambda}$ is chosen. Here $\boldsymbol{\beta}=(5.6,\ 5.6,\ 5.6,\ 0)^\top$ and the correlation matrix of covariates is such that $\text{cor}(x_j,x_k)=-.39,\ j<k<4$ and $\text{cor}(x_j,x_4)=.23,\ j<4$.

The experimental results are summarized in Table 4.2 in which the frequencies of correct selection are shown. We see that the original Lasso does not seem to give consistent model selection. For all the other methods, the frequencies of correct selection go to 1 as n increases and σ decreases. In general, our proposed method for model selection performs better than the aLasso.

Example 3 (large p example). We consider a large-p example in which $p=100$ with various sample sizes $n=50,\ 100,\ 200$. We set up a *sparse recovery problem* in which most of coefficients are zero except $\beta_j=5,\ j=10,20,...,100$. From the previous examples, the performances of the four methods BaLasso-Freq, BaLasso-Median, BaLaso-Mean and BaLasso-EB are similar. We therefore just consider the BaLasso-Mean as a representative and compare it to the adaptive Lasso which is generally superior to the Lasso.

Table 4.3 summarizes our simulation results, in which the design matrix is simulated as in Example 1. The BaLasso-Mean performs satisfactorily in this example and outperforms the aLasso in variable selection.

Example 4 (prediction). In this example, we examine the predictive ability of the

Table 4.3: Frequency of correctly-fitted models over 100 replications for Example 3.

n	σ	aLasso	BaLasso-Mean
50	1	24	39
	3	24	35
	5	8	29
100	1	40	100
	3	39	99
	5	20	86
200	1	100	100
	3	88	100
	5	78	97

BaLasso-BMA experimentally. As discussed in Section 4.1.2, when there is model uncertainty, making predictions conditioning on a single fixed parameter vector is not optimal predictively. Suppose that the dataset D is split into two sets: a *training set* D^T and *prediction set* D^P. Let $\Delta=(\mathbf{x}_\Delta,y_\Delta)\in D^P$ be a future observation and $\hat{y}_\Delta$ be a prediction of y_Δ based on D^T. We measure the predictive performance by the prediction squared error (PSE)

$$\text{PSE} = \frac{1}{|D^P|}\sum_{\Delta\in D^P}|y_\Delta - \hat{y}_\Delta|^2. \tag{4.9}$$

We compare PSE of the BaLasso-BMA to that of the BaLasso-Mean in which $\hat{y}_\Delta=\mathbf{x}_\Delta^\top\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the solution to (4.2) with smoothing parameter vector fixed at the posterior mean of $\boldsymbol{\lambda}$. We also compare the predictive performance of the BaLasso-BMA to that of the Lasso, aLasso, and the original Bayesian Lasso (BLasso). The implementation of the BLasso is similar to the BaLasso except that the BLasso has a single smoothing parameter.

We first consider a small-p case in which data sets are generated from Tibshirani's

Table 4.4: Prediction squared errors averaged over 100 replications for the small-p case.

$n_T = n_P$	σ	Lasso	aLasso	BLasso	BaLasso-Mean	BaLasso-BMA
30	1	2.02	1.97	1.27	1.17	1.16
	3	17.43	17.37	10.88	15.51	11.06
	5	42.74	42.13	29.43	41.32	29.56
	10	126.6	126.2	109.6	123.9	109.9
100	1	1.44	1.43	1.04	1.07	1.03
	3	12.69	12.58	9.66	9.62	9.48
	5	34.89	34.79	25.79	27.55	25.83
	10	117.6	117.5	105.7	118.2	106.5
200	1	1.27	1.27	1.01	1.03	1.01
	3	11.44	11.40	9.42	9.32	9.32
	5	31.30	31.18	25.32	25.36	25.19
	10	120.7	120.7	103.9	108.8	104.3

model (2.40) but now with $\boldsymbol{\beta} = (3,\ 1.5,\ 0.1,\ 0.1,\ 2,\ 0,\ 0,\ 0)^\top$. By adding two small effects we expect there to be model uncertainty. Table 4.4 presents the prediction squared errors averaged over 100 replications with various factors n_T (size of training set), n_P (size of prediction set) and σ. The experiment shows that the BaLasso-BMA performs slightly better than the BLasso and BaLasso-Mean, and much better than the Lasso and aLasso.

Similarly, we consider a large-p case as in Example 3 but now with $\beta_{10} = \beta_{20} = \beta_{30} = \beta_{40} = \beta_{50} = .5$ in order to get model uncertainty. The results are summarized in Table 4.5. Unlike for the small-p case, the BLasso now performs surprisingly badly. This may be due to the fact that the BLasso uses the same shrinkage for every coefficient. As shown, the BaLasso-BMA outperforms the others.

Table 4.5: Prediction squared errors averaged over 100 replications for the large-p case.

$n_T = n_P$	σ	Lasso	aLasso	BLasso	BaLasso-Mean	BaLasso-BMA
100	1	3.50	4.17	9.57	1.67	1.23
	3	15.49	17.70	27.42	10.88	10.42
	5	34.45	39.81	42.43	28.66	28.19
	10	149.3	178.1	161.0	124.5	117.6
200	1	2.46	2.41	5.23	1.11	1.07
	3	17.11	17.09	15.12	10.42	10.22
	5	44.49	44.39	33.92	27.18	27.06
	10	148.1	147.5	136.1	112.0	108.9

Example 5: Prostate cancer data. We now apply our methodologies to the prostate cancer data set which was considered in Section 2.6.2. We first consider the variable selection problem. The data set of size 97 is standardized so that the intercept β_0 is excluded. Table 4.6 summarizes the selected smoothing parameters and estimated coefficients by various methods. Note that, for the Lasso and aLasso there is just one smoothing parameter and putting the values on the first row as presented in the table does not mean these parameters are only associated with the first predictor.

The EB estimation here is implemented using the stabilized Algorithm 2.2 of Atchade [2009], in which the compact sets are selected to be $\otimes[-n-1,n+1]$, and the step-size $a_n = 2/n$ is obtained after a few trials by justifying the convergence of iterations $\boldsymbol{\lambda}^{(n)}$ graphically. As shown in Table 4.6, the BaLasso-EB, BaLasso-Mean and BaLasso-Median give very similar estimates for λ_j corresponding to nonzero-estimated coefficients, but fairly different estimates for λ_j corresponding to zero-estimated coefficients. The effects of increased penalty parameters on the zero coefficients are obvious: smaller shrinkage is

Table 4.6: Prostate cancer example: selected smoothing parameters and coefficient estimates

Selected $\boldsymbol{\lambda}$					Coefficient estimate $\hat{\boldsymbol{\beta}}$				
	BaLasso		Lasso	aLasso		BaLasso		Lasso	aLasso
-EB	-Median	-Mean			-EB	-Median	-Mean		
1.2	1.2	1.4	2.4	1.9	0.56	0.56	.56	.56	.57
1.6	1.5	1.8			0.44	0.44	.44	.36	.44
332.8	841.1	1066			0	0	0	-.02	0
55.8	16.7	20.4			0	0	0	.1	0
1.2	1.1	1.3			0.59	0.59	.58	.43	.51
97.6	86.6	113.2			0	0	0	0	0
89.8	78.7	105.1			0	0	0	0	0
754.4	1242	1824			0	0	0	.01	0

applied to the nonzero coefficients and larger shrinkage is applied to those which should be removed.

The adaptive Lasso and all of the proposed strategies (including the BaLasso-Freq also) for variable selection produce the same model whose BIC is -25.19, while BIC of the model selected by the Lasso is -21.38. Therefore the model chosen by our methods is favorable according to this criterion at least. Note that the model selected here by the BaLasso methods is the same as the one selected by the loss rank criterion in Section 2.6.2.

We now proceed to explore model uncertainty inherent in this dataset. Let $M(\boldsymbol{\lambda})$ be the model selected w.r.t. shrinkage parameter vector $\boldsymbol{\lambda}$. The posterior model probability (PMP) of a model M will be

$$p(M|D) = \int_{\lambda:M(\lambda)=M} p(\boldsymbol{\lambda}|D)d\boldsymbol{\lambda}.$$

Table 4.7: Prostate cancer example: 10 models with highest posterior model probability

Models							PMP (%)
1	2			5			27.9
1	2			5		8	16.1
1			4	5			6.3
1	2		4	5		8	5.9
1	2					8	5.7
1	2		4	5			5.1
1	2	3		5		8	4.9
1	2	3	4	5		8	4.9
1			4	5		8	3.2
1	2						3.1

From the Gibbs samples of $\boldsymbol{\lambda}$, it is straightforward to estimate these PMPs. Table 4.7 presents 10 models with highest PMP. The most frequently selected model is the same as the one selected by the aLasso and our methods. In comparison to the examples in Section 3.1.5, the model uncertainty indicator defined in (3.12) MUI=.58 suggests that the presence of model uncertainty is not very clear in this case. The model with highest posterior probability accounts for 27.9% of the total. Moreover, this probability is also considerably different from that of the model with second highest posterior probability.

To examine the predictive performance, we split the data set (without standardizing) into two sets: the first 50 observations form the training set D^T, the rest form the prediction set D^P. The PSEs of the aLasso, BLasso, BaLasso-Median, BaLasso-BMA are 1.89, 1.91, 1.91, 1.86, respectively. Therefore, although the presence of model uncertainty is not very clear, the BaLasso-BMA still provides comparable and slightly better estimates in terms

of prediction.

4.1.4 A unified framework

So far, we have focused on the BaLasso for linear regression. This section extends the BaLasso to general linear models, such as generalized linear models, Cox's model and so on, with other penalties, such as the group penalty [Yuan and Lin, 2006] and the composite absolute penalty [Zhao et al., 2009]. This unified framework enables us to study variable selection in a much broader context.

Denote by $L(\boldsymbol{\beta})$ the minus log-likelihood. In order to use the BaLasso developed for linear regression, we approximate $L(\boldsymbol{\beta})$ by the least squares approximation (LSA)

$$\begin{aligned} L(\boldsymbol{\beta}) &\approx L(\tilde{\boldsymbol{\beta}}) + \frac{\partial L(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\top}\frac{\partial^2 L(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \\ &= \text{constant} + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\hat{\Sigma}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\hat{\Sigma}^{-1} := \partial^2 L(\tilde{\boldsymbol{\beta}})/\partial \boldsymbol{\beta}^2$. The LSA was proposed by Wang and Leng [2007] for a unified treatment of variable selection using the Lasso. To use the BaLasso for a general model, the sampling distribution of $\mathbf{y}$, conditional on $\boldsymbol{\beta}$, can be approximately written as

$$\mathbf{y}|\boldsymbol{\beta} \sim \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\top}\hat{\Sigma}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right).$$

Using this approximation, we only need to update the hierarchical model for $\mathbf{y}$ in the linear model as

$$\mathbf{y}|\boldsymbol{\beta} \sim \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^{\top}\hat{\Sigma}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right).$$

Now we discuss in detail three novel applications of the BaLasso for general linear models with flexible penalties.

BaLasso for general linear models. The frequentist adaptive Lasso for general models estimates $\boldsymbol{\beta}$ by minimizing

$$L(\boldsymbol{\beta}) + \sum \lambda_j |\beta_j|. \tag{4.10}$$

Its Bayesian version is the following

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta} &\sim \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top \hat{\Sigma}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right), \\
\boldsymbol{\beta}|\boldsymbol{\tau}^2 &\sim N_p(\mathbf{0}, D_\tau),\ D_\tau = \mathrm{diag}(\boldsymbol{\tau}^2), \\
\boldsymbol{\tau}^2|\boldsymbol{\lambda}^2 &\sim \prod_{j=1}^{p} \frac{\lambda_j^2}{2} e^{-\lambda_j^2 \tau_j^2/2}, \\
\boldsymbol{\lambda}^2 &\sim \prod_{j=1}^{p} (\lambda_j^2)^{r-1} e^{-\delta\lambda_j^2}
\end{aligned}$$

where $\boldsymbol{\tau}^2 := (\tau_1^2,...,\tau_p^2)^\top$, $\boldsymbol{\lambda}^2 := (\lambda_1^2,...,\lambda_p^2)^\top$. Note that we no longer have σ^2 in the hierarchy. The full conditionals are specified by

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}^2, \boldsymbol{\lambda}^2 &\sim N_p\left((\hat{\Sigma}^{-1} + D_\tau^{-1})^{-1}\hat{\Sigma}^{-1}\tilde{\boldsymbol{\beta}}, (\hat{\Sigma}^{-1} + D_\tau^{-1})^{-1}\right), \\
\frac{1}{\tau_j^2} = \gamma_j|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\lambda}^2 &\sim \text{inverse-Gaussian}\left(\frac{\lambda_j}{|\beta_j|}, \lambda_j^2\right),\ j = 1, ..., p, \\
\lambda_j^2|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau}^2 &\sim \text{gamma}(r+1, \delta + \frac{\tau_j^2}{2}),\ j = 1, ..., p.
\end{aligned}$$

BaLasso for group Lasso. The adaptive group Lasso [Yuan and Lin, 2006] for general models minimizes

$$L(\boldsymbol{\beta}) + \sum_{j=1}^{J} \lambda_j \|\boldsymbol{\beta}_j\|_{l_2} \tag{4.11}$$

where $\boldsymbol{\beta}_j$ is the coefficient vector of the jth group, $j = 1,...,J$. The corresponding Bayesian

hierarchy is as follows:

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta} &\sim \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^\top\hat{\Sigma}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\right), \\
\boldsymbol{\beta}_j|\boldsymbol{\tau}^2 &\sim N_{m_j}(\mathbf{0}, \tau_j^2 I_{m_j}),\ j=1,...,J \\
\tau_j^2|\boldsymbol{\lambda}^2 &\sim \text{gamma}\left(\frac{m_j+1}{2}, \frac{\lambda_j^2}{2}\right),\ j=1,...,J \\
\lambda_j^2 &\sim \text{gamma}(r,\delta),\ j=1,...,J
\end{aligned}$$

where m_j is the size of group j, I_{m_j} is the identity matrix of order m_j. This prior was also used by Kyung et al. [2010] for grouped variable selection in linear regression.

The full conditionals can be obtained as follows. Let $\tilde{X}$ be the square root matrix of $\hat{\Sigma}^{-1}$ and $\tilde{\mathbf{y}} := \tilde{X}\tilde{\boldsymbol{\beta}}$. Write $\tilde{X} = [\tilde{X}_1,...,\tilde{X}_J]$ with block matrices $\tilde{X}_j$ of size $p \times m_j$. We have

$$\begin{aligned}
\boldsymbol{\beta}_j|\mathbf{y},\boldsymbol{\beta}_{-j},\boldsymbol{\tau}^2,\boldsymbol{\lambda}^2 &\sim N_{m_j}\left(A_j^{-1}\tilde{X}_j^\top(\tilde{\mathbf{y}}-\sum_{j'\neq j}\tilde{X}_{j'}\boldsymbol{\beta}_{j'}), A_j^{-1}\right), \\
\frac{1}{\tau_j^2}=\gamma_j|\mathbf{y},\boldsymbol{\beta},\boldsymbol{\lambda}^2 &\sim \text{inverse Gaussian}\left(\frac{\lambda_j}{\|\boldsymbol{\beta}_j\|}, \lambda_j^2\right), \\
\lambda_j^2|\mathbf{y},\boldsymbol{\beta},\boldsymbol{\tau}^2 &\sim \text{gamma}\left(r+\frac{m_j+1}{2}, \delta+\frac{\tau_j^2}{2}\right),\ j=1,...,J,
\end{aligned}$$

where $\boldsymbol{\beta}_{-j} = (\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)$ and $A_j = \tilde{X}_j^\top\tilde{X}_j + (1/\tau_j^2)I_{m_j}$.

BaLasso for composite absolute penalty. We now consider the group selection problem in which a natural ordering among the groups is present. By $j \to j'$, we mean that group j should be added into the model before another group j', i.e., if group j' is selected then group j must be included in the model as well. We extend the composite absolute penalty [Zhao et al., 2009] by allowing different tuning parameters for different groups

$$\sum_{\text{group } j}\lambda_j\|(\boldsymbol{\beta}_j,\boldsymbol{\beta}_{\text{all } j':j\to j'})\|_{l_2},$$

where $\boldsymbol{\beta}_j$ is a coefficient vector and this penalty represents some hierarchical structure in

the model. From this, the desired prior for $\boldsymbol{\beta}$ is the multi-Laplace

$$\pi(\boldsymbol{\beta}) \propto \exp\left(\sum_j \lambda_j \|(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j':j\to j'})\|_{l_2}\right)$$

which can be expressed as the following normal-gamma mixture

$$\int \left(\frac{1}{2\pi\tau_j^2}\right)^{\frac{k_j}{2}} \exp\left(-\frac{\|(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j':j\to j'})\|^2}{2\tau_j^2}\right) \frac{(\frac{\lambda_j^2}{2})^{\frac{k_j+1}{2}} (\tau_j^2)^{\frac{k_j+1}{2}-1}}{\Gamma(\frac{k_j+1}{2})} \exp(-\frac{\lambda_j^2\tau_j^2}{2}) d\tau_j^2 = \exp\big(\lambda_j \|(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j':j\to j'})\|\big) \tag{4.12}$$

where $k_j := m_j + \sum_{j':j\to j'} m_{j'}$. Similar to the Bayesian formulations before, this identity leads to the idea of using a hierarchical Bayesian formulation with a normal prior for $\boldsymbol{\beta}|\boldsymbol{\tau}^2$ and a gamma prior for τ_j^2. More specifically, the prior for $\boldsymbol{\beta}|\boldsymbol{\tau}^2$ will be

$$\boldsymbol{\beta}|\boldsymbol{\tau}^2 \propto \exp\left(-\sum_j \frac{\|(\boldsymbol{\beta}_j, \boldsymbol{\beta}_{j':j\to j'})\|^2}{2\tau_j^2}\right) = \prod_j \exp\left(-\frac{1}{2}\Big(\frac{1}{\tau_j^2} + \sum_{j':j'\to j} \frac{1}{\tau_{j'}^2}\Big)\|\boldsymbol{\beta}_j\|^2\right).$$

This suggests that the hierarchical prior for $\boldsymbol{\beta}_j|\boldsymbol{\tau}^2$ is independently multinormal with mean $\mathbf{0}$ and covariance matrix $(1/\tau_j^2 + \sum_{j':j'\to j} 1/\tau_{j'}^2)^{-1} I_{m_j}$, $j = 1,...,J$. We therefore have the following hierarchy

$$\begin{aligned}
\mathbf{y}|\boldsymbol{\beta} &\sim \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \hat{\Sigma}^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right), \\
\boldsymbol{\beta}_j|\boldsymbol{\tau}^2 &\sim N_{m_j}\left(\mathbf{0}, \sigma_j^2 I_{m_j}\right), \text{ where } \sigma_j^2 := (\frac{1}{\tau_j^2} + \sum_{j':j'\to j} \frac{1}{\tau_{j'}^2})^{-1} \\
\tau_j^2|\boldsymbol{\lambda}^2 &\sim \text{gamma}\left(\frac{k_j+1}{2}, \frac{\lambda_j^2}{2}\right) \\
\lambda_j^2 &\sim \text{gamma}(r, \delta) \text{ for } j = 1, ..., J.
\end{aligned}$$

Full conditionals. It is now straightforward to derive the full conditionals as follows

$$
\begin{aligned}
\boldsymbol{\beta}_j|\mathbf{y},\boldsymbol{\beta}_{-j},\boldsymbol{\tau}^2,\boldsymbol{\lambda}^2 &\sim N_{m_j}\left(A_j^{-1}\tilde{X}_j^{\top}(\tilde{\mathbf{y}}-\sum_{j'\neq j}\tilde{X}_{j'}\boldsymbol{\beta}_{j'}),A_j^{-1}\right),\\
\frac{1}{\tau_j^2}=\gamma_j|\mathbf{y},\boldsymbol{\beta},\boldsymbol{\lambda}^2 &\sim \text{inverse Gaussian}\left(\frac{\lambda_j}{\|(\boldsymbol{\beta}_j,\boldsymbol{\beta}_{j':j\to j'})\|},\lambda_j^2\right),\\
\lambda_j^2|\mathbf{y},\boldsymbol{\beta},\boldsymbol{\tau}^2 &\sim \text{gamma}\left(r+\frac{k_j+1}{2},\delta+\frac{\tau_j^2}{2}\right),\ j=1,...,J
\end{aligned}
$$

where $\boldsymbol{\beta}_{-j}=(\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{j-1},\boldsymbol{\beta}_{j+1},...,\boldsymbol{\beta}_J)$ and $A_j=\tilde{X}_j'\tilde{X}_j+(1/\sigma_j^2)I_{m_j}$.

We now assess the usefulness of this unified framework by three examples. For brevity, we only report the performance of various methods in terms of model selection.

Example 6: BaLasso in logistic regression. We simulate independent observations from Bernoulli distributions with probabilities of success

$$\mu_i=P(y_i=1|x_i,\boldsymbol{\beta})=\frac{\exp(5+x_i^{\top}\boldsymbol{\beta})}{1+\exp(5+x_i^{\top}\boldsymbol{\beta})}$$

where $\boldsymbol{\beta}=(3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0)^{\top}$, and $\mathbf{x}_i=(x_{i1},...,x_{ip})^{\top}\sim N_p(0,\Sigma)$ with $\sigma_{ij}=0.5^{|i-j|}$. We compare the performance of the BaLasso to that of the Lasso and the aLasso. The performance is measured by the frequency of correct fitting and average number of zero coefficients over 100 replications. The weight vector in the aLasso is as usual assigned as $\hat{\mathbf{w}}=1/|\hat{\boldsymbol{\beta}}^{\text{MLE}}|$, where $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ is the MLE of $\boldsymbol{\beta}$. The shrinkage parameters in the Lasso and aLasso are tuned by 5-fold cross-validation. Table 4.8 presents the simulation result for various sample size n. The aLasso in this example works better than the Lasso. The suggested BaLasso works very well, especially when the sample size n is large. In addition, the BaLasso often produces sparser models than the others do.

Example 7: BaLasso for group selection. We consider in this example the group selection problem in a linear regression framework. We follow the simulation setup of Yuan and Lin [2006]. A vector of 15 latent variables $\mathbf{Z}\sim N_{15}(\mathbf{0},\Sigma)$ with $\sigma_{ij}=0.5^{|i-j|}$ are

Table 4.8: Example 6: Frequency of correctly-fitted models over 100 replications. The numbers in parentheses are average numbers of zero-estimated coefficients. The oracle average number is 5.

n	Lasso	aLasso	BaLasso
200	3(2.15)	35(3.97)	36(6.19)
300	5(2.42)	42(4.07)	90(5.10)
500	4(2.66)	41(4.00)	100(5.00)

first simulated. For each latent variable Z_i, a 3-level factor F_i is determined according to whether Z_i is smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$ or in between. The factor F_i then is coded by two dummy variables. There are totally 30 dummy variables $X_1,...,X_{30}$ and 15 groups with $\boldsymbol{\beta}_j=(\beta_{2j-1},\beta_{2j})^\top$, $j=1,...,J=15$. After having the design matrix X, a vector of responses is generated from the following linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, I) \tag{4.13}$$

where most of $\boldsymbol{\beta}_j=\mathbf{0}$ except $\boldsymbol{\beta}_1=(-1.2,\, 1.8)^\top$, $\boldsymbol{\beta}_3=(1,\, 0.5)^\top$, $\boldsymbol{\beta}_5=(1,\, 1)^\top$. We compare the performance of the BaLasso to that of the gLasso in Yuan and Lin [2006] and the adaptive group Lasso (agLasso) [Wang and Leng, 2008] in terms of frequencies of correct fitting and average numbers of not-selected factors over 100 replications. We follow Wang and Leng [2008] to take the weights $\hat{w}_j=1/\|\hat{\boldsymbol{\beta}}_j^{\text{MLE}}\|$ with $\hat{\boldsymbol{\beta}}_j^{\text{MLE}}$ are the MLE of $\boldsymbol{\beta}_j$. The tuning parameters in gLasso and agLasso are tuned using AIC with the degrees of freedom as in Yuan and Lin [2006]. We use 1000 values of λ equally spaced from 0 to $\lambda_{\max}$ to search for the optimal value. Table 4.9 reports the simulation result. Both gLasso and agLasso seem to select unnecessarily large models and have low rate of correct fitting. In contrast, the BaLasso seems to produce more parsimonious models when n is small. In general, the BaLasso works much better than the others in terms of model selection consistency.

Table 4.9: Example 7: Frequency of correctly-fitted models and average numbers (in parentheses) of not-selected factors over 100 replications. The oracle average number is 12.

n	gLasso	agLasso	BaLasso
100	5(6.64)	22(9.60)	15(14.86)
200	8(6.92)	48(10.72)	90(12.04)
500	7(7.24)	70(11.34)	100(12.00)

Example 8: BaLasso for main and interaction effect selection. In this example we demonstrate the BaLasso with composite absolute penalty for selecting main and interaction effects in a linear framework. We consider the model II of Yuan and Lin [2006]. First, 4 factors are created as in the previous example, each factor is then coded by two dummy variables. The true model is generated from (4.13) with main effects $\beta_1 = (3,\ 2)^\top$, $\beta_2 = (3,\ 2)^\top$ and interaction $\beta_{1\cdot 2} = (1,\ 1.5,\ 2,\ 2.5)^\top$. There are totally 10 groups (4 main effects and 6 second-order interaction effects) with the natural ordering in which main effects should be selected before their corresponding interaction effects. We use the BaLasso formulation with composite absolute penalty to account for this ordering. Table 4.10 reports the simulation result. We observe that both gLasso and agLasso sometimes select effects in a "wrong" order (interactions are seclected while the corresponding main effects are not). As a result, they have low rates of correct fitting. The BaLasso always produce the models with effects in the "right" order. This fact has been theoretically proven in Zhao et al. [2009]. In general, the BaLasso outperforms its competitors.

Table 4.10: Example 8: Frequency of correctly-fitted models and average numbers (in parentheses) of not-selected effects over 100 replications. The oracle average number is 7.

n	gLasso	agLasso	BaLasso
100	18(4.25)	45(5.45)	72(7.28)
200	36(5.16)	88(6.78)	100(7.00)
500	34(5.24)	96(6.92)	100(7.00)

4.2 Variable selection for heteroscedastic linear regression

Consider the heteroscedastic linear regression model

$$y_i = x_i^\top \beta + \sigma_i \epsilon_i, \ i = 1, \ldots, n \tag{4.14}$$

where y_i is a response, $x_i = (x_{i1},...,x_{ip})^\top$ is a corresponding p-vector of predictors, $\beta = (\beta_1,...,\beta_p)^\top$ is a vector of unknown mean parameters, $\epsilon_i \sim N(0,1)$ are independent errors and

$$\log \sigma_i^2 = z_i^\top \alpha,$$

where $z_i = (z_{i1},...,z_{iq})^\top$ is a q-vector of predictors and $\alpha = (\alpha_1,...,\alpha_q)^\top$ is a vector of unknown variance parameters. In this model the standard deviation σ_i of y_i is being modelled in terms of the predictors z_i; this heteroscedastic model is contrasted with the usual homoscedastic model which assumes σ_i is constant. We take a Bayesian approach to inference in this model and consider a prior distribution $p(\theta)$ on $\theta = (\beta^\top, \alpha^\top)^\top$ of the form $p(\theta) = p(\beta)p(\alpha)$ with $p(\beta)$ and $p(\alpha)$ both normal, $N(\mu_\beta^0, \Sigma_\beta^0)$ and $N(\mu_\alpha^0, \Sigma_\alpha^0)$, respectively. It is possible to consider hierarchical extensions for the priors on $p(\beta)$ and $p(\alpha)$, but we do not consider this here.

We will consider a variational Bayes approach to inference (which will be discussed in detail in Section 4.2.1). The term variational approximation refers to a wide range of different methods where the common idea is to convert a problem of integration into an optimization problem. For Bayesian inference, variational approximation provides a fast alternative to Monte Carlo methods for approximating posterior distributions in complex models, especially in high-dimensional problems. In the heteroscedastic linear regression model, we will consider a variational approximation to the joint posterior distribution of β and α as $q(\beta,\alpha)=q(\beta)q(\alpha)$, where $q(\beta)$ and $q(\alpha)$ are both normal densities, $N(\mu_\beta^q,\Sigma_\beta^q)$ and $N(\mu_\alpha^q,\Sigma_\alpha^q)$, respectively. It is also possible to give a variational treatment in which independence is not assumed between β and α but this complicates the variational optimization somewhat. We attempt to choose the parameters in the variational posterior μ_β^q, μ_α^q, Σ_β^q and Σ_α^q to minimize the Kullback-Leibler divergence between the true posterior distribution $p(\beta,\alpha|y)$ and $q(\beta,\alpha)$. This results in a lower bound on the log marginal likelihood $\log p(y)$ - a key quantity in Bayesian model selection. Our first contribution is the derivation of a closed form for the lower bound and the proposal of an iterative scheme for maximizing it. This lower bound maximization plays a crucial role in the variable selection problem discussed in Section 4.2.2.

Variable selection is a fundamental problem in statistics and machine learning, which has attracted many researchers recently. A large number of methods have been proposed for variable selection in homoscedastic regression. The traditional approach in the Bayesian framework is Bayesian variable selection which consists in building a hierarchical Bayes model and using MCMC algorithms to estimate posterior model probabilities [George and McCulloch, 1993, Smith and Kohn, 1996]. This methodology is computationally demanding in high-dimensional problems and there is a need for fast alternatives in some applications. The reader is referred to Nott et al. [2011] and Tran et al. [2011]

for some detailed reports on real computational time savings of variational approximation methods compared to MCMC in the regression context. In high-dimensional settings, commonly-used alternatives include the family of greedy algorithms [Tropp, 2004, Zhang, 2009]. Greedy algorithms, also known as *matching pursuit* [Mallat and Zhang, 1993] in signal processing, are closely related to the Lasso [Tibshirani, 1996] and the LARS algorithm [Efron et al., 2004]. See Zhao and Yu [2007], Efron et al. [2004] and Zhang [2009] for excellent comparisons of these families of algorithms. In the statistical context, greedy algorithms have been proven to be very efficient for variable selection in linear regression under the assumption of homoscedasticity, i.e., where the variance is assumed to be constant [Zhang, 2009].

In many applications the assumption of constant variance may be unrealistic. Ignoring heteroscedasticity may lead to serious problems in inference, such as misleading assessments of significance, poor predictive performance and inefficient estimation of mean parameters. In some cases, learning the structure in the variance may be the primary goal. See Chan et al. [2006] and Carroll and Ruppert [1988] for a more detailed discussion on heteroscedastic modelling. Despite a large number of works on heteroscedastic regression and modelling covariate-dependent overdispersion in overdispersed generalized linear models [Efron, 1986, Smyth, 1989, Yee and Wild, 1996, Rigby and Stasinopoulos, 2005], methods for model selection seem to be somewhat overlooked. Yau and Kohn [2003] and Chan et al. [2006] consider Bayesian variable selection and MCMC approaches to computation in heteroscedastic Gaussian models and extensions involving flexible modelling of the mean and variance functions. Cottet et al. [2008] consider extensions to overdispersed generalized linear and generalized additive models. These approaches are computationally demanding in high dimensional settings. A general and flexible framework for modelling overdispersed data is considered by Yee and Wild [1996] and Rigby and Stasinopoulos

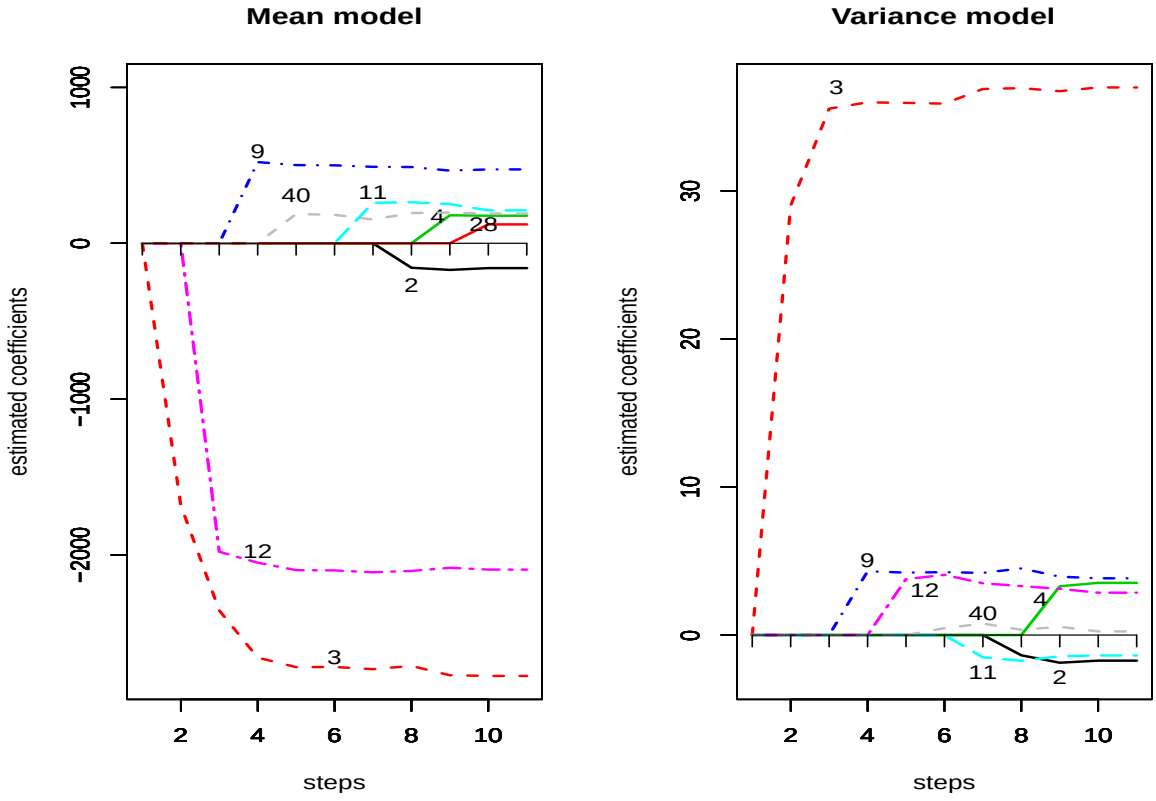
[2005]. Methods for model selection are less well developed in these general models. A common approach is to use information criteria such as generalized AIC and BIC together with forward stepwise methods (see, e.g., [Rigby and Stasinopoulos, 2005, Section 6]). We compare our own approaches to such methods later. Our main contribution in this section is to propose a novel fast greedy algorithm for variable selection in heteroscedastic linear regression. We show that the proposed algorithm is in homoscedastic cases similar to currently used methods while having many attractive properties and working efficiently in high-dimensional problems. An efficient R program is available on the author's website.

Our methodology has potential for extension to more complicated frameworks such as variable selection in regression density estimation [Tran et al., 2011] in which the density of the response variable is smoothy estimated at all points in the covariate space with a mixtute of experts. However, we do not discuss this extension here.

In Section 4.2.3 we apply our algorithm to the analysis of the diabetes data [Efron et al., 2004] using heteroscedastic linear regression. This data set consists of 64 predictors (constructed from 10 input variables for a "quadratic model") and 442 observations. We show in Figure 4.3 the estimated coefficients corresponding to selected predictors as functions of iteration steps in our algorithm, for both mean and variance models. The algorithm stops after 11 forward selection steps with 8 and 7 predictors selected for the mean and variance models, respectively.

The rest of this section is organized as follows. The closed form of the lower bound and the iterative scheme for maximizing it are presented in Section 4.2.1. We present in Section 4.2.2 our novel fast greedy algorithm, and compare it to existing greedy algorithms in the literature for homoscedastic regression. Numerical examples are presented in Section 4.2.3. Technical derivation is relegated to the Appendix.

Figure 4.3: Solution paths as functions of iteration steps for analyzing the diabetes data using heteroscedastic linear regression. The algorithm stops after 11 iterations with 8 and 7 predictors selected for the mean and variance models, respectively. The selected predictors enter the mean (variance) model in the order 3, 12, ..., 28 (3, 9, ..., 4).



4.2.1 Variational Bayes

We now give a brief introduction to the variational approximation method. For a more detailed exposition see, for example, Jordan et al. [1999], [Bishop, 2006, Chapter 10], or see Ormerod and Wand [2010] for a statistically oriented introduction. The term variational approximation refers to a wide range of different methods where the common idea is to convert a problem of integration into an optimization problem. Here we will only be concerned with applications of variational methods in Bayesian inference and only with a particular approach sometimes referred to as parametric variational approximation. Write θ for all our unknown parameters, $p(\theta)$ for the prior distribution and $p(y|\theta)$ for the likelihood. In Bayesian inference, decisions are based on the posterior distribution $p(\theta|y) \propto p(\theta)p(y|\theta)$, and a common difficulty in applications is how to compute quantities of interest with respect to the posterior. These computations often involve the evaluation of high-dimensional integrals. Variational approximation proceeds by approximating the posterior distribution directly. Formally, we consider a family of distributions $q(\theta|\lambda)$ where λ denotes some unknown parameters and attempt to choose λ so that $q(\theta|\lambda)$ is closest to $p(\theta|y)$ in some sense. In particular, we attempt to minimize the Kullback-Leibler divergence

$$\int \log \frac{q(\theta|\lambda)}{p(\theta|y)} q(\theta|\lambda) d\theta$$

with respect to λ. Using the identity

$$\log p(y) \quad = \quad \int \log \frac{p(\theta)p(y|\theta)}{q(\theta|\lambda)} q(\theta|\lambda) d\theta + \int \log \frac{q(\theta|\lambda)}{p(\theta|y)} q(\theta|\lambda) d\theta, \tag{4.15}$$

where $p(y) = \int p(\theta)p(y|\theta)d\theta$, we see that minimizing the Kullback-Leibler divergence is equivalent to maximization of

$$\int \log \frac{p(\theta)p(y|\theta)}{q(\theta|\lambda)} q(\theta|\lambda) d\theta. \tag{4.16}$$

Here (4.16) is a lower bound on the log marginal likelihood log$p(y)$ due to the non-negativity of the Kullback-Leibler divergence term in (4.15). The lower bound (4.16), when maximized with respect to λ, is often used as an approximation to the log marginal likelihood log$p(y)$ and clearly (again from (4.15)) the error in the approximation is the Kullback-Leibler divergence between the approximation $q(\theta|\lambda)$ and the true posterior. The approximation is useful, since log$p(y)$ is a key quantity in Bayesian model selection.

For our heteroscedastic linear model the lower bound (4.16) can be expressed as

$$L = T_1 + T_2 + T_3,$$

where

$$\begin{aligned} T_1 &= \int \log p(\beta, \alpha) q(\beta) q(\alpha) d\beta d\alpha, \\ T_2 &= \int \log p(y|\beta, \alpha) q(\beta) q(\alpha) d\beta d\alpha, \\ T_3 &= -\int \log \left(q(\beta)q(\alpha)\right) q(\beta) q(\alpha) d\beta d\alpha. \end{aligned}$$

We show (see the Appendix) that these three terms, which are all expectations with respect to the (assumed normal) variational posterior, can be evaluated analytically. Putting the terms together we obtain that the lower bound (4.16) on the log marginal likelihood is

$$\begin{aligned} L = & \frac{p+q}{2} - \frac{n}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_\beta^q \Sigma_\beta^{0\,-1}| + \frac{1}{2}\log|\Sigma_\alpha^q \Sigma_\alpha^{0\,-1}| - \frac{1}{2}\mathrm{tr}(\Sigma_\beta^{0\,-1}\Sigma_\beta^q) \\ & -\frac{1}{2}\mathrm{tr}(\Sigma_\alpha^{0\,-1}\Sigma_\alpha^q) - \frac{1}{2}(\mu_\beta^q - \mu_\beta^0)^\top \Sigma_\beta^{0\,-1}(\mu_\beta^q - \mu_\beta^0) - \frac{1}{2}(\mu_\alpha^q - \mu_\alpha^0)^\top \Sigma_\alpha^{0\,-1}(\mu_\alpha^q - \mu_\alpha^0) \\ & -\frac{1}{2}\sum_{i=1}^n z_i^\top \mu_\alpha^q - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - x_i^\top \mu_\beta^q)^2 + x_i^\top \Sigma_\beta^q x_i}{\exp\left(z_i^\top \mu_\alpha^q - \frac{1}{2} z_i^\top \Sigma_\alpha^q z_i\right)}. \end{aligned} \tag{4.17}$$

This needs to be maximized with respect to μ_β^q, μ_α^q, Σ_β^q, Σ_α^q. We consider an iterative scheme in which we maximize with respect to each of the blocks of parameters μ_β^q, μ_α^q, Σ_β^q, Σ_α^q with the other blocks held fixed.

Write X for the design matrix with ith row $x_i^\top$ and D for the diagonal matrix with ith diagonal element $1/\exp(z_i^\top \mu_\alpha^q - 1/2 z_i^\top \Sigma_\alpha^q z_i)$. Maximization with respect to μ_β^q with other terms held fixed leads to

$$\mu_\beta^q = \left(X^\top DX + {\Sigma_\beta^0}^{-1}\right)^{-1} \left({\Sigma_\beta^0}^{-1}\mu_\beta^0 + X^\top Dy\right).$$

Maximization with respect to Σ_β^q with other terms held fixed leads to

$$\Sigma_\beta^q = \left({\Sigma_\beta^0}^{-1} + X^\top DX\right)^{-1}.$$

Handling the parameters μ_α^q and Σ_α^q in the variational posterior for α is more complex. We proceed in the following way. If no parametric form for the variational posterior $q(\alpha)$ is assumed (that is, if we do not assume that $q(\alpha)$ is normal), but only assume the factorization $q(\theta) = q(\beta)q(\alpha)$, then the optimal choice for $q(\alpha)$ for a given $q(\beta) = N(\mu_\beta^q, \Sigma_\beta^q)$ is (see Ormerod and Wand [2010], for example)

$$q(\alpha) \propto \exp\left[E(\log p(\theta)p(y|\theta))\right], \tag{4.18}$$

where the expectation is with respect to $q(\beta)$. Similar to the derivation of the lower bound (4.17), it is easy to see that

$$q(\alpha) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n z_i^T\alpha - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - x_i^T\mu_\beta^q)^2 + x_i^T\Sigma_\beta^q x_i}{\exp(z_i^T\alpha)} - \frac{1}{2}(\alpha - \mu_\alpha^0)^T {\Sigma_\alpha^0}^{-1}(\alpha - \mu_\alpha^0)\right),$$

which takes the form of the posterior (apart from a normalization constant) for a Bayesian generalized linear model with gamma response and log link, coefficient of variation $\sqrt{2}$, and responses $w_i = (y_i - x_i^T\mu_\beta^q)^2 + x_i^T\Sigma_\beta^q x_i$ with the log of the mean response being $z_i^T\alpha$. The prior in this gamma generalized linear model is $N(\mu_\alpha^0, \Sigma_\alpha^0)$. If we use a quadratic approximation to $\log q(\alpha)$ then this results in a normal approximation to $q(\alpha)$. We choose the mean and variance of the normal approximation simply by the posterior mode and

the negative inverse Hessian of the log posterior at the mode for the gamma generalized linear model described above. The computations required are standard ones involving iteratively weighted least squares in a Bayesian generalized linear model. With μ_α^q the posterior mode, we obtain for Σ_α^q the expression

$$\Sigma_\alpha^q = \left(Z^\top W Z + {\Sigma_\alpha^0}^{-1}\right)^{-1},$$

where W is diagonal with ith diagonal element $w_i \exp(-z_i^\top \mu_\alpha^q)/2$. Our optimization over μ_α^q and Σ_α^q is only approximate, so that we only retain the new values in the optimization if they result in an improvement in the lower bound (4.17). The advantage of our approximate approach is the closed form expression for the update of Σ_α^q once μ_α^q is found, so that explicit numerical optimization for a possibly high-dimensional covariance matrix is avoided.

The explicit algorithm for our method is the following.

Algorithm 1: Maximization of the variational lower bound.

1. Initialize parameters μ_α^q, Σ_α^q.
2. $\mu_\beta^q \leftarrow \left(X^\top D X + {\Sigma_\beta^0}^{-1}\right)^{-1}\left({\Sigma_\beta^0}^{-1}\mu_\beta^0 + X^\top D y\right)$ where D is the diagonal matrix with ith diagonal entry $1/\exp\left(z_i^\top \mu_\alpha^q - 1/2 z_i^\top \Sigma_\alpha^q z_i\right)$.
3. $\Sigma_\beta^q \leftarrow \left(X^\top D X + {\Sigma_\beta^0}^{-1}\right)^{-1}$.
4. Obtain μ_α^q as the posterior mode for a gamma generalized linear model with normal prior $N(\mu_\alpha^0, \Sigma_\alpha^0)$, gamma responses $w_i = (y_i - x_i^\top \mu_\beta^q)^2 + x_i^\top \Sigma_\beta^q x_i$, coefficient of variation $\sqrt{2}$ and where the log of the mean is $z_i^\top \alpha$.
5. $\Sigma_\alpha^q \leftarrow \left(Z^\top W Z + {\Sigma_\alpha^0}^{-1}\right)^{-1}$ where W is diagonal with ith diagonal element $w_i \exp(-z_i^\top \mu_{\alpha j}^q)/2$.

6. If the updates done in steps 3 and 4 do not improve the lower bound (4.17) then their old values are retained.

7. Repeat steps 2-6 until the increase in the variational lower bound (4.17) is less than some user specified tolerance.

For initialization, we first perform an OLS fit for the mean model to get an estimate $\hat{\beta}$ of β. Then we take the residuals from this fit, $r_i = (y_i - x_i^\top \hat{\beta})^2$ say, and do an OLS fit of $\log r_i$ to the predictors z_i to obtain our initial estimate of μ_α^q. The initial value of Σ_α^q is then set to the covariance matrix of the least squares estimator. When the OLS fits are not valid, some other method such as the Lasso can be used instead. The application of this algorithm to the problem of variable selection in Section 4.2.2 always involves only situations in which the above OLS fits are available.

We mention one further extension of our method. We have assumed above that the prior covariance matrices Σ_β^0 and Σ_α^0 are known. Later we will assume $\Sigma_\beta^0 = \sigma_\beta^2 I$ and $\Sigma_\alpha^0 = \sigma_\alpha^2 I$ where I denotes the identity matrix and σ_β^2 and σ_α^2 are scalar variance parameters. We further assume that $\mu_\beta^0 = 0$ and $\mu_\alpha^0 = 0$. It may be helpful to perform some data driven shrinkage so that σ_β^2 and σ_α^2 are considered unknown and to be estimated from the data. Our lower bound (4.17) can be considered as an approximation to $\log p(y|\sigma_\beta^2, \sigma_\alpha^2)$, and the log posterior for $\sigma_\beta^2, \sigma_\alpha^2$ is apart from an additive constant

$$\log p(\sigma_\beta^2, \sigma_\alpha^2) + \log p(y|\sigma_\beta^2, \sigma_\alpha^2).$$

If we assume independent inverse gamma priors, $IG(a,b)$, for σ_β^2 and σ_α^2 and if we replace the log marginal likelihood by the lower bound and maximize, we get

$$\sigma_\beta^2 = \frac{b + \frac{1}{2}\mu_\beta^{q\top}\mu_\beta^q + \frac{1}{2}\mathrm{tr}(\Sigma_\beta^q)}{a + 1 + p/2}$$

and

$$\sigma_\alpha^2 = \frac{b + \frac{1}{2}\mu_\alpha^{q\top}\mu_\alpha^q + \frac{1}{2}\mathrm{tr}(\Sigma_\alpha^q)}{a + 1 + q/2}.$$

These updating steps can be added to the Algorithm 1 given above.

4.2.2 Variable selection

In the discussion of the previous section the choice of predictors in the mean and variance models was fixed. We now wish to consider the problem of variable selection in the heteroscedastic linear model, and the question of computationally efficient model search when the number of candidate predictors is very large, perhaps much larger than the sample size. In Section 4.2.1 we denoted the marginal likelihood by $p(y)$ without making explicit conditioning on the model but now we write $p(y|m)$ for the marginal likelihood in a model m. If we have a prior distribution $p(m)$ on the set of all models under consideration, then Bayes' rule leads to the posterior distribution on the model given by $p(m|y) \propto p(m)p(y|m)$. We can use the variational lower bound for $\log p(y|m)$ as a replacement for $\log p(y|m)$ in this formula as one strategy for Bayesian variable selection when $p(y|m)$ is difficult to compute, and we follow that strategy here. For a more thorough review of the Bayesian approach to model selection see, for example, O'Hagan and Forster [2004].

Before presenting our strategy for ranking variational lower bounds, we discuss here the model prior. Suppose we have a current model with predictors x_i, $i \in C \subset D = \{1,...,p\}$, in the mean model and z_i, $i \in V \subset E = \{1,...,q\}$, in the variance model. The subsets C and V give indices for the currently active predictors in the mean and variance models. Let π_i^μ (π_j^σ) be the prior probability for inclusion of x_i (z_j) in the mean (variance) model, and write $\pi^\mu = (\pi_1^\mu,...,\pi_p^\mu)^\top$, $\pi^\sigma = (\pi_1^\sigma,...,\pi_q^\sigma)^\top$. We assume that the inclusions of predictors are

independent a priori with

$$p(C|\pi^\mu) = \prod_{i\in C} \pi_i^\mu \prod_{i\notin C}(1-\pi_i^\mu), \;\; p(V|\pi^\sigma) = \prod_{j\in V} \pi_j^\sigma \prod_{j\notin V}(1-\pi_j^\sigma),$$

and the prior probability of a model m with index sets C and V in its mean and variance models is assumed to be

$$p(m) = p(C, V|\pi^\mu, \pi^\sigma) = p(C|\pi^\mu)p(V|\pi^\sigma). \tag{4.19}$$

If no such detailed prior information is available for each individual predictor (which is the situation we consider here), one may assume that $\pi_1^\mu = ... = \pi_p^\mu = \pi_\mu$ and $\pi_1^\sigma = ... = \pi_q^\sigma = \pi_\sigma$ (we note a slight abuse of notation here). Then

$$p(C|\pi_\mu) = \pi_\mu^{|C|}(1-\pi_\mu)^{p-|C|}, \;\; p(V|\pi_\sigma) = \pi_\sigma^{|V|}(1-\pi_\sigma)^{q-|V|}, \tag{4.20}$$

where hyperparameters π_μ, $\pi_\sigma \in [0,1]$ are user-specified. One can encourage parsimonious models by setting small ($<1/2$) π_μ and π_σ. The smaller the π_μ and π_σ, the smaller prior probabilities are put on complex models. By setting $\pi_\mu = \pi_\sigma = 1/2$, one can set the uniform prior on the models. Another option is to put uniform distributions on π_μ and π_σ, then

$$p(C) = \int_0^1 p(C|\pi_\mu)d\pi_\mu \propto \binom{p}{|C|}^{-1}, \;\; p(V) = \int_0^1 p(V|\pi_\sigma)d\pi_\sigma \propto \binom{q}{|V|}^{-1}. \tag{4.21}$$

This prior agrees with the one used in the extended BIC proposed by Chen and Chen [2008]. It has the advantage of requiring no hyperparameter while still encouraging parsimony. We recommend using this as the default prior.

We now consider adding a single variable in either the mean or the variance model, and then a one-step update to the current variational lower bound in the proposed model as a computationally thrifty way of ranking the predictors for their possible inclusion. In our one-step update, we consider a variational approximation in which the variational

posterior distribution factorizes into independent parts for the added parameter and the parameters in the current model. We stress that this factorization is only assumed for the purpose of ranking predictors for inclusion - once a variable has been selected for inclusion the posterior distribution is approximated using the method outlined in Section 4.2.1. Write β_C for the parameters in the current mean model and X_C for the corresponding design matrix, and α_V for the parameters in the current variance model with Z_V the corresponding design matrix. Write x_{Ci} for the ith row of X_C and z_{Vi} for the ith row of Z_V.

Ranking predictors in the mean model

Let us consider first the effect of adding the predictor x_j, $j \in D \setminus C$, to the mean model. We write β_j for the coefficient of x_j and we consider a variational approximation to the posterior of the form

$$q(\theta) = q(\beta_C)q(\beta_j)q(\alpha_V), \tag{4.22}$$

with $q(\beta_C) = N(\mu^q_{\beta C}, \Sigma^q_{\beta C})$, $q(\alpha_V) = N(\mu^q_{\alpha V}, \Sigma^q_{\alpha V})$ and $q(\beta_j) = N(\mu^q_{\beta j}, (\sigma^q_{\beta j})^2)$. Suppose that we have fitted a variational approximation for the current model (i.e., the model without x_j) using the procedure of Section 4.2.1. We now consider fitting the extended model with $\mu^q_{\beta C}, \Sigma^q_{\beta C}, \mu^q_{\alpha V}$ and $\Sigma^q_{\alpha V}$ fixed at the optimized values obtained for the current model, and consider just one step of a variational algorithm for maximizing the variational lower bound in the new model with respect to the parameters $\mu^q_{\beta j}, (\sigma^q_{\beta j})^2$. In effect for our variational lower bound (4.17), we are assuming that the variational posterior distribution for $({\beta_C}^\top, \beta_j)^\top$ is normal with mean $({\mu^q_{\beta C}}^\top, \mu^q_{\beta j})^\top$ and covariance matrix

$$\begin{bmatrix} \Sigma^q_{\beta C} & 0 \\ 0 & (\sigma^q_{\beta j})^2 \end{bmatrix}.$$

Substituting these forms into (4.17) and further assuming $\mu_\beta^0 = 0$, $\mu_\alpha^0 = 0$, $\Sigma_\beta^0 = \sigma_\beta^2 I$ and $\Sigma_\alpha^0 = \sigma_\alpha^2 I$ (see the remarks at the end of Section 4.2.1) we obtain the lower bound

$$L = L_{\text{old}} + \frac{1}{2} + \frac{1}{2}\log\frac{(\sigma_{\beta j}^q)^2}{\sigma_\beta^2} - \frac{(\sigma_{\beta j}^q)^2}{2\sigma_\beta^2} - \frac{(\mu_{\beta j}^q)^2}{2\sigma_\beta^2} - \frac{1}{2}\sum_{i=1}^n \frac{x_{ij}^2(\sigma_{\beta j}^q)^2 + x_{ij}^2(\mu_{\beta j}^q)^2 - 2x_{ij}\mu_{\beta j}^q(y_i - x_{Ci}^\top \mu_{\beta C}^q)}{\exp\left(z_{iV}^\top \mu_{\alpha V}^q - \frac{1}{2} z_{iV}^\top \Sigma_{\alpha V}^q z_{iV}\right)} \tag{4.23}$$

where L_{old} is the previous lower bound for the current model without predictor j. Here we are writing x_{ij} for the value of predictor j for observation i. Optimizing the above bound with respect to $\mu_{\beta j}^q$ and $(\sigma_{\beta j}^q)^2$ and writing $\hat{\mu}_{\beta j}^q$ and $(\hat{\sigma}_{\beta j}^q)^2$ for the optimizers gives

$$\hat{\mu}_{\beta j}^q = \left(\sum_{i=1}^n \frac{x_{ij}(y_i - x_{Ci}^\top \mu_{\beta C}^q)}{\exp\left(z_{Vi}^\top \mu_{\alpha V}^q - \frac{1}{2} z_{Vi}^\top \Sigma_{\alpha V}^q z_{Vi}\right)}\right) \Big/ \left(\frac{1}{\sigma_\beta^2} + \sum_{i=1}^n \frac{x_{ij}^2}{\exp\left(z_{Vi}^\top \mu_{\alpha V}^q - \frac{1}{2} z_{Vi}^\top \Sigma_{\alpha V}^q z_{Vi}\right)}\right), \tag{4.24}$$

and

$$(\hat{\sigma}_{\beta j}^q)^2 = \left(\frac{1}{\sigma_\beta^2} + \sum_{i=1}^n \frac{x_{ij}^2}{\exp\left(z_{Vi}^\top \mu_{\alpha V}^q - \frac{1}{2} z_{Vi}^\top \Sigma_{\alpha V}^q z_{Vi}\right)}\right)^{-1}. \tag{4.25}$$

Substituting these back into the lower bound (4.23) gives

$$L_{\text{old}} + \frac{1}{2}\log\frac{(\hat{\sigma}_{\beta j}^q)^2}{\sigma_\beta^2} + \frac{1}{2}\frac{(\hat{\mu}_{\beta j}^q)^2}{(\hat{\sigma}_{\beta j}^q)^2}. \tag{4.26}$$

If the variance model contains only an intercept, this result agrees with greedy selection algorithms where predictors are ranked according to the correlation between a predictor and the residuals from the current model (see, e.g., Zhang [2009]). We will discuss this point in detail in the case of homoscedasticity below. Later we write the optimized value of (4.23) as $L_j^M(C,V)$, the superscript M means the lower bound associated with the model for *m*ean.

Ranking predictors in the variance model

So far we have considered only the addition of a predictor in the mean model. We now attempt a similar analysis of the effect of inclusion of a predictor in the variance

model. With the mean model fixed, suppose we are considering adding a predictor z_j, $j \in E \setminus V$, to the variance model. We consider a normal approximation to the posterior $q(\theta)=q(\beta_C)q(\alpha_V)q(\alpha_j)$ with $q(\beta_C)=N(\mu^q_{\beta C},\Sigma^q_{\beta C})$, $q(\alpha_V)=N(\mu^q_{\alpha V},\Sigma^q_{\alpha V})$ and $q(\alpha_j)=N(\mu^q_{\alpha j},(\sigma^q_{\alpha j})^2)$. The variational lower bound is

$$L_{\text{old}}+\frac{1}{2}+\frac{1}{2}\log\frac{(\sigma^q_{\alpha j})^2}{\sigma^2_\alpha}-\frac{(\sigma^q_{\alpha j})^2}{2\sigma^2_\alpha}-\frac{(\mu^q_{\alpha j})^2}{2\sigma^2_\alpha}-\frac{1}{2}\sum_i z_{ij}\mu^q_{\alpha j}$$
$$-\frac{1}{2}\sum_{i=1}^n\left\{\frac{1}{\exp(z^\top_{Vi}\mu^q_{\alpha V}-\frac{1}{2}z^\top_{Vi}\Sigma^q_{\alpha V}z_{Vi}+z_{ij}\mu^q_{\alpha j}-\frac{1}{2}z^2_{ij}(\sigma^q_{\alpha j})^2)}-\frac{1}{\exp(z^\top_{Vi}\mu^q_{\alpha V}-\frac{1}{2}z^\top_{Vi}\Sigma^q_{\alpha V}z_{Vi})}\right\}\times$$
$$\big((y_i-x^\top_{Ci}\mu^q_{\beta C})^2+x^\top_{Ci}\Sigma^q_{\beta C}x_{Ci}\big), \tag{4.27}$$

where L_{old} is the lower bound for the current model without predictor z_j. To obtain good values for $\mu^q_{\alpha j}$ and $(\sigma^q_{\alpha j})^2$ we use an approximation similar to the one used for the variance parameters in Section 4.2.1. If we do not assume a normal form for $q(\alpha_j)$ but just the factorization $q(\theta)=q(\beta_C)q(\alpha_V)q(\alpha_j)$ and with the current $q(\beta_C)$ and $q(\alpha_V)$ fixed, then the optimal $q(\alpha_j)$ is

$$q(\alpha_j) \propto \exp[E(\log p(\alpha_j) + \log p(y|\theta))],$$

where the expectation is with respect to $q(\beta_C)q(\alpha_V)$. We have that

$$\begin{aligned} E(\log p(\alpha_j) + \log p(y|\theta)) &= E\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2_\alpha - \frac{\alpha_j^2}{2\sigma^2_\alpha} - \frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^n z^\top_{Vi}\alpha_V\right. \\ &\quad \left. -\frac{1}{2}\sum_{i=1}^n z_{ij}\alpha_j - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - x^\top_{Ci}\beta_C)^2}{\exp\left(z^\top_{Vi}\alpha_V + z_{ij}\alpha_j\right)}\right) \\ &= -\frac{n+1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2_\alpha - \frac{\alpha_j^2}{2\sigma^2_\alpha} - \frac{1}{2}\sum_{i=1}^n z^\top_{Vi}\mu^q_{\alpha V} - \frac{1}{2}\sum_{i=1}^n z_{ij}\alpha_j \\ &\quad -\frac{1}{2}\sum_{i=1}^n \frac{(y_i - x^\top_{Ci}\mu^q_{\beta C})^2 + x^\top_{Ci}\Sigma^q_{\beta C}x_{Ci}}{\exp\left(z^\top_{Vi}\mu^q_{\alpha V} + z_{ij}\alpha_j - \frac{1}{2}z^\top_{Vi}\Sigma^q_{\alpha V}z_{Vi}\right)}. \end{aligned} \tag{4.28}$$

We will make a normal approximation $N(\hat\mu^q_{\alpha j},(\hat\sigma^q_{\alpha j})^2)$ to the optimal $q(\alpha_j)$ via the mode and negative inverse second derivative of (4.28). Differentiating with respect to α_j, we

obtain

$$-\frac{\alpha_j}{\sigma_\alpha^2}-\frac{1}{2}\sum_{i=1}^{n} z_{ij}+\frac{1}{2}\sum_{i=1}^{n}\frac{z_{ij}v_i}{\exp(z_{ij}\alpha_j)} \quad \text{where} \quad v_i=\frac{(y_i-x_{Ci}^\top\mu_{\beta C}^q)^2+x_{Ci}^\top\Sigma_{\beta C}^q x_{Ci}}{\exp\left(z_{Vi}^\top\mu_{\alpha V}^q-\frac{1}{2}z_{Vi}^\top\Sigma_{\alpha V}^q z_{Vi}\right)}.$$

Approximating $\exp(-z_{ij}\alpha_j)\approx 1-z_{ij}\alpha_j$ (i.e., using a Taylor series expansion about zero), setting the derivative to zero and solving gives

$$\hat{\mu}_{\alpha j}^q=\left(\frac{1}{2}\sum_{i=1}^{n} z_{ij}(v_i-1)\right)\Big/\left(\frac{1}{\sigma_\alpha^2}+\frac{1}{2}\sum_{i=1}^{n} z_{ij}^2 v_i\right). \tag{4.29}$$

To get more accurate estimation of the mode, some optimization procedure may be used here with (4.29) used as an initial point. In our R implementation, the Newton method was used because (4.28) has its second derivative available in a closed form (see (4.30) below). We found that (4.29) is a very good approximation as the Newton iteration very often stops after a small number of iterations (with a stopping tolerance as small as 10^{-10}).

Differentiating (4.28) once more, and finding the negative inverse of the second derivative at $\hat{\mu}_{\alpha j}^q$ gives

$$(\hat{\sigma}_{\alpha j}^q)^2=\left(\frac{1}{\sigma_\alpha^2}+\frac{1}{2}\sum_{i=1}^{n}\frac{z_{ij}^2 v_i}{\exp(z_{ij}\hat{\mu}_{\alpha j}^q)}\right)^{-1}. \tag{4.30}$$

We can plug these values back into the lower bound in order to rank different predictors for inclusion in the variance model. We write the optimized value of (4.27) as $L_j^D(C,V)$, the superscript D means the lower bound associated with the model for standard *d*eviance.

Summary of the algorithm

We summarize our variable selection algorithm below. We write $L(C,V)$ for the optimized value of the lower bound (4.17) with the predictor set C in the mean model and the predictor set V in the variance model. Write C_{+j} for the set $C\cup\{j\}$ and V_{+j} for the set $V\cup\{j\}$.

Algorithm 2: Variational approximation ranking (VAR) algorithm.

1. Initialize C and V and set $L_{\text{opt}} := L(C,V)$.

2. Repeat the following steps until stop

   (a) Store $C_{\text{old}} := C$, $V_{\text{old}} := V$.

   (b) Let $j^* = \text{argmax}_j\{L_j^M(C,V) + \log p(C_{+j},V)\}$. If $L(C_{+j^*},V) + \log p(C_{+j^*},V) > L_{\text{opt}} + \log p(C,V)\}$ then set $C := C_{+j^*}$, $L_{\text{opt}} = L(C_{+j^*},V)$.

   (c) Let $j^* = \text{argmax}_j\{L_j^D(C,V) + \log p(C,V_{+j})\}$. If $L(C,V_{+j^*}) + \log p(C,V_{+j^*}) > L_{\text{opt}} + \log p(C,V)$ then set $V := V_{+j^*}$, $L_{\text{opt}} = L(C,V_{+j^*})$.

   (d) If $C = C_{\text{old}}$ and $V = V_{\text{old}}$ then stop, else return to (a).

Forward-backward ranking algorithm

The ranking algorithm described above can be regarded as a forward greedy algorithm because it considers adding at each step another predictor to the current model. Hereafter we refer to this algorithm as forward variational ranking algorithm or fVAR in short. Like the other forward greedy algorithms that have been widely used in many scientific fields, the fVAR works well in most of the examples that we have encountered. However, a major drawback with the forward selection algorithms is that if a predictor has been wrongly selected then it can not be removed anymore. A natural remedy for this is to add a backward elimination process in order to correct mistakes made in the earlier forward selection. We present here a recipe for ranking predictors for exclusion in mean and variance models.

Let C, V be the current sets of predictors in the mean and variance models, respectively. With $j \in C$, we write C_{-j} for the set $C \setminus \{j\}$ and consider now the effect of removing the predictor x_j to the lower bound. In order to reduce computational burden, we need some way to avoid the need to do lower bound maximization for each model C_{-j} when

ranking x_j for exclusion. Similar as before, we consider a variational approximation using the factorization (4.22) for the variational posterior distribution. Following steps (4.23)-(4.26), we can approximately write the lower bound for the current model (i.e.,the model contains x_j) as the sum of the lower bound for the model without x_j and a x_j-based term

$$L(C,V) \approx L(C_{-j},V) + \Gamma^M_{C_{-j},V}(j), \tag{4.31}$$

with

$$\Gamma^M_{C_{-j},V}(j) := \frac{1}{2}\log\frac{(\hat{\sigma}^q_{\beta j})^2}{\sigma^2_\beta} + \frac{1}{2}\frac{(\hat{\mu}^q_{\beta j})^2}{(\hat{\sigma}^q_{\beta j})^2}, \tag{4.32}$$

where $\hat{\mu}^q_{\beta j}$, $\hat{\sigma}^q_{\beta j}$ are as in (4.24) and (4.25) with C replaced by C_{-j}. All the relevant quantities needed in the calculation of $\Gamma^M_{C_{-j},V}(j)$ are fixed at optimized values maximizing the lower bound for the current model. The subscripts C_{-j},V is to emphasize that the quantities needed are adjusted correspondingly when the predictor j is removed from the mean model. The most plausible candidate for exclusion from the current mean model then is

$$j^* = \text{argmax}_{j\in C}\{L(C_{-j},V) + \log p(C_{-j},V)\} = \text{argmin}_{j\in C}\{\Gamma^M_{C_{-j},V}(j) - \log p(C_{-j},V)\}. \tag{4.33}$$

We now rank the predictors for exclusion in the variance model. Following the arguments above, we can write

$$L(C,V) \approx L(C,V_{-j}) + \Gamma^D_{C,V_{-j}}(j) \tag{4.34}$$

with

$$\begin{aligned}\Gamma^D_{C,V}(j) = &\frac{1}{2} + \frac{1}{2}\log\frac{(\hat{\sigma}^q_{\alpha j})^2}{\sigma^2_\alpha} - \frac{(\hat{\sigma}^q_{\alpha j})^2}{2\sigma^2_\alpha} - \frac{(\hat{\mu}^q_{\alpha j})^2}{2\sigma^2_\alpha} - \frac{1}{2}\sum_i z_{ij}\hat{\mu}^q_{\alpha j} \\ &-\frac{1}{2}\sum_{i=1}^n\left\{\frac{1}{\exp(z^\top_{Vi}\mu^q_{\alpha V} - \frac{1}{2}z^\top_{Vi}\Sigma^q_{\alpha V}z_{Vi} + z_{ij}\hat{\mu}^q_{\alpha j} - \frac{1}{2}z^2_{ij}(\hat{\sigma}^q_{\alpha j})^2} - \frac{1}{\exp(z^\top_{Vi}\mu^q_{\alpha V} - \frac{1}{2}z^\top_{Vi}\Sigma^q_{\alpha V}z_{Vi})}\right\}\times \\ &\left((y_i - x^\top_{Ci}\mu^q_{\beta C})^2 + x^\top_{Ci}\Sigma^q_{\beta C}x_{Ci}\right),\end{aligned} \tag{4.35}$$

where $\hat{\mu}^q_{\alpha j}$, $\hat{\sigma}^q_{\alpha j}$ are as in (4.29)-(4.30) with V replaced by V_{-j}. The most plausible candidate for exclusion from the current variance model then is

$$j^* = \text{argmax}_{j\in V}\{L(C,V_{-j}) + \log p(C,V_{-j})\} = \text{argmin}_{j\in V}\{\Gamma^D_{C,V_{-j}}(j) - \log p(C,V_{-j})\}. \quad (4.36)$$

Algorithm 3: Forward-backward variational approximation ranking algorithm.

1. Initialize C and V, and set $L_{\text{opt}} = L(C,V)$.

2. Forward selection: as in Step 2 in Algorithm 2.

3. Backward elimination: Repeat the following steps until stop

   (a) Store $C_{\text{old}} := C$, $V_{\text{old}} := V$.

   (b) Find j^* as in (4.33). If $L(C_{-j^*},V) + \log p(C_{-j^*},V) > L_{\text{opt}} + \log p(C,V)$ then set $C = C_{-j^*}$, $L_{\text{opt}} = L(C_{-j^*},V)$.

   (c) Find j^* as in (4.36). If $L(C,V_{-j^*}) + \log p(C,V_{-j^*}) > L_{\text{opt}} + \log p(C,V)$ then set $V = V_{-j^*}$, $L_{\text{opt}} = L(C,V_{-j^*})$.

   (d) If $C = C_{\text{old}}$ and $V = V_{\text{old}}$ then stop, else return to (a).

Hereafter we refer to this algorithm as fbVAR.

In some applications where $X \equiv Z$, it might be meaningful to restrict the search for inclusion in the variance model to those predictors that have been included in the mean model. To this end, in the forward selection we just need to restrict the search for the most plausible candidate j^* in Step 2(c) of Algorithm 2 to set C, i.e., $j^* = \text{argmax}_{j\in C}\{L^D_j(C,V) + \log p(C,V_j)\}$. Also, when consider the removal of a candidate j from the mean model in the backward elimination, we need to remove j from the variance model as well if $j \in V$, i.e., Step 3(b) of Algorithm 3 must be modified to

3(b') Let $j^* = \text{argmin}_{j \in C} \{\Gamma^M_{C_{-j}, V_{-j}}(j) - \log p(C_{-j}, V_{-j})\}$. If $L(C_{-j^*}, V_{-j^*}) + \log p(C_{-j^*}, V_{-j^*}) > L_{\text{opt}} + \log p(C, V)$ then set $C = C_{-j^*}$, $V = V_{-j^*}$, $L_{\text{opt}} = L(C_{-j^*}, V_{-j^*})$.

Later we compare with the variable selection approaches for heteroscedastic regression implemented in the GAMLSS (generalized additive model for location, scale and shape) package [Rigby and Stasinopoulos, 2005]. The GAMLSS framework allows modeling of the mean and other parameters (like the standard deviation, skewness and kurtosis) of the response distribution as flexible functions of predictors. Variable selection is done with stepwise selection using a generalized AIC or BIC as the stopping rule. The GAMLSS uses a Fisher scoring algorithm to maximize the likelihood for ranking every predictor for inclusion/exclusion rather than only the most plausible one as in the VAR algorithm, which leads to a heavy computational burden for large-p problems.

The ranking algorithm for homoscedastic regression

In order to get more insight into our VAR algorithm, we discuss now the algorithm for the homoscedastic linear regression model. In the case of constant variance, the variance parameter α now becomes scalar, we rename the quantities Σ^0_α, Σ^q_α as $(\sigma^0_\alpha)^2$, $(\sigma^q_\alpha)^2$, respectively. The optimal choice (4.18) for $q(\alpha)$ becomes

$$q(\alpha) \propto \exp\left(-\frac{n}{2}\alpha - \frac{1}{2} v e^{-\alpha} - \frac{1}{2}\frac{\alpha^2}{(\sigma^0_\alpha)^2}\right) \quad \text{where} \quad v := \sum_{i=1}^{n} \left((y_i - x_i^\top \mu^q_\beta)^2 + x_i^\top \Sigma^q_\beta x_i\right).$$

Using the approximation $\exp(-\alpha) \approx 1 - \alpha$, it is easy to see that the mean and variance of the normal approximation are

$$\mu^q_\alpha = \frac{v - n}{v + 2/(\sigma^0_\alpha)^2} \quad \text{and} \quad (\sigma^q_\alpha)^2 = \left(\frac{v}{2} e^{-\mu^q_\alpha} + \frac{1}{(\sigma^0_\alpha)^2}\right)^{-1},$$

respectively. We now can replace steps 4 and 5 in Algorithm 1 by these two closed forms so that the computations can be reduced greatly. Similar to the above discussion, the

Newton method may be used here in order to get a more accurate estimate of the mode. In our experience, however, this is not necessary here.

For the variable selection problem we now just need to rank the predictors for inclusion/exclusion in the mean model. Assume that we are using the uniform model prior, i.e., $p(C,V) \equiv$constant, or a model prior as in (4.20), the ranking of predictors then follows the ranking of the lower bounds. We further assume that the design matrix X has been standardized such that $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = n$, the optimizer $(\hat{\sigma}_{\beta j}^q)^2$ in (4.25) then does not depend on j, and the ranking of the lower bound (4.26) follows the ranking of $\left|\sum_{i=1}^n x_{ij}(y_i - x_{Ci}^\top \mu_{\beta C}^q)\right|$ (i.e., it follows the ranking of the absolute correlation of the predictors with the standardized residuals from the current model). This result agrees with frequentist matching pursuit and greedy algorithms where predictors are ranked according to the correlation between a predictor and the residuals from the current model [Mallat and Zhang, 1993, Zhang, 2009, Efron et al., 2004]. This is also similar to computationally thrifty path following algorithms (the LARS of Efron et al. [2004], the BLasso of Zhao and Yu [2007]).

For all the existing frequentist algorithms for variable selection in the literature, extra tuning parameters are involved (shrinkage parameters in penalization procedures like the Lasso, number of iterations in matching pursuit, stopping parameter ϵ in greedy algorithms) and their performance depends essentially on the method used to choose these tuning parameters. An advantage of our method is that no extra tuning parameters is required, the final model is chosen when the lower bound (which is a good approximation of the logarithm of the evidence $\log p(y)$) is maximized - a natural stopping rule in Bayesian model selection with uniform model prior. Unlike many commonly used greedy algorithms, our Bayesian framework is able to incorporate prior information (if available) on models and/or to encourage parsimonious models if desired. Besides involving ex-

tra tuning parameters, penalized estimates are often biased (see, for example, Friedman [2008], Efron et al. [2004]). While our method can penalize non-zero coefficients through the prior if desired, it does not rely on shrinkage of coefficients to do variable selection, so that in principle it might produce better estimation of non-zero coefficients. Simulation studies in Section 4.2.3 confirm this point. Note that we do not consider models of all sizes, the algorithm stops when important predictors have been included in the model so that computations of Algorithm 1 just involve matrices with low-dimension. This is another advantage which makes our method potentially valuable for variable selection in high-dimensional problems. Our experience shows that the VAR algorithm is as fast as the LARS algorithm in problems with thousands of predictors.

4.2.3 Numerical examples

Heteroscedastic case. We present here a simulation study for our VAR method for simultaneous variable selection and parameter estimation in the heteroscedastic linear regression model, and compare its performance to that of the GAMLSS and aLasso methods. Data sets were generated from the following model

$$y = 2 + x^\top \tilde{\beta} + \sigma e^{\frac{1}{2}x^\top \tilde{\alpha}} \epsilon, \tag{4.37}$$

with $\tilde{\beta}=(3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0)^\top$, $\epsilon \sim N(0,\ 1)$. Predictors x were first generated from normal distributions $N(0,\Sigma)$ with $\Sigma_{ij}=0.5^{|i-j|}$ and then transformed into the unit interval by the cumulative distribution function $\Phi(.)$ of the standard normal. The reason for making the transformation was to control the magnitude of noise level, i.e., the quantity $\sigma e^{\frac{1}{2}x^\top \tilde{\alpha}}$. Let $\beta=(2,\ \tilde{\beta}^\top)^\top$ and $\alpha=(\log\sigma^2,\tilde{\alpha}^\top)^\top$ be the mean and variance parameters, respectively, where $\tilde{\alpha}=(0,\ 3,\ 0,\ 0,\ -3,\ 0,\ 0,\ 0)^\top$. Note that the true predictors in the variance model were

among those in the mean model. This *prior* information was employed in the GAMLSS and VAR.

The performance was measured by correctly-fitted rates (CFR), numbers of zero-estimated coefficients (NZC) (for both the mean and variance models), mean squared error (MSE) of predictions and partial prediction score (PPS) averaged over 100 replications. MSE and PPS were evaluated based on independent prediction sets generated in the same manner as the training set. We compared the performance of the VAR and GAMLSS methods (when heteroscedasticity was assumed) to that of the aLasso (when homoscedasticity was assumed). The simulation results are summarized in Table 4.11 for various factors sample size n, n_P (size of prediction sets D^P) and σ. As shown, the VAR method did a good job and outperformed the others.

We also considered a "large p, small n" case in which $\tilde{\beta}$ and $\tilde{\alpha}$ in model (4.37) were vectors of dimension 500 with most of the components zero except $\tilde{\beta}_{50}=\tilde{\beta}_{100}=...=\tilde{\beta}_{250}=5$, $\tilde{\beta}_{300}=\tilde{\beta}_{350}=...=\tilde{\beta}_{500}=-5$ and $\tilde{\alpha}_{100}=\tilde{\alpha}_{200}=5$, $\tilde{\alpha}_{300}=\tilde{\alpha}_{400}=-5$. The simulation results are summarized in Table 4.12. Note that the GAMLSS is not applicable when $n<p$, and moreover that in the case with $n\geq p$ and with large p the current implementation version of the GAMLSS is much more time consuming compared to the VAR and even not working with p as large as 500. We are not aware of any existing methods in the literature for variable selection in heteroscedastic linear models for "large p, small n" case.

Homoscedastic case. We also considered a simulation study when the data come from homoscedastic models. Data sets were generated from the linear model (4.37) with $\tilde{\alpha}\equiv 0$, i.e.

$$y = 2 + x^{\top}\tilde{\beta} + \sigma\epsilon$$

with predictors x generated from normal distributions $N(0,\Sigma)$ with $\Sigma_{ij}=0.5^{|i-j|}$. We were concerned with simulating a sparse, high-dimensional case. To this end, $\tilde{\beta}$ was set to be a

Table 4.11: Small-p case: CFR, NZC, MSE and PPS averaged over 100 replications. The numbers in parentheses are NZC.

$n=n_P$	σ	measures	aLasso	GAMLSS	VAR
50	0.5	CFR in mean	64 (4.56)	36 (4.06)	80 (4.88)
		CFR in var.	nil	70 (5.74)	80 (5.96)
		MSE	0.56	0.49	0.48
		PPS	1.17	0.89	0.87
	1	CFR in mean	22 (4.72)	38 (4.60)	56 (5.00)
		CFR in var.	nil	50 (5.88)	60 (6.22)
		MSE	2.45	2.29	2.24
		PPS	2.02	1.78	1.69
100	0.5	CFR in mean	74 (4.50)	30 (3.98)	88 (4.84)
		CFR in var.	nil	64 (5.62)	90 (5.90)
		MSE	0.52	0.48	0.48
		PPS	1.12	0.87	0.77
	1	CFR in mean	36 (4.68)	42 (4.30)	66 (4.76)
		CFR in var.	nil	58 (5.72)	76 (5.84)
		MSE	2.20	2.08	2.03
		PPS	1.83	1.62	1.51
200	0.5	CFR in mean	94 (4.90)	48 (4.14)	100 (5.00)
		CFR in var.	nil	70 (5.70)	94 (5.94)
		MSE	0.48	0.46	0.46
		PPS	1.06	0.87	0.74
	1	CFR in mean	56 (4.36)	36 (4.06)	88 (4.88)
		CFR in var.	nil	82 (5.80)	100 (6.00)
		MSE	2.01	1.93	1.92
		PPS	1.77	1.52	1.43

Table 4.12: Large-p case: CFR, NZC, MSE and PPS averaged over 100 replications. The numbers in parentheses are NZC.

		VAR				aLasso		
$n=n_P$	σ	CFR in mean	CFR in var.	MSE	PPS	CFR in mean	MSE	PPS
100	0.5	80 (489.75)	90 (495.90)	5.40	1.91	20 (491.80)	11.65	2.65
	1	70 (489.05)	65 (495.80)	20.29	2.30	0 (495.75)	35.11	3.27
150	0.5	100 (490.00)	95 (495.90)	13.76	0.84	40 (491.95)	20.02	3.40
	1	95 (489.95)	85 (495.85)	28.97	1.52	5 (495.05)	43.18	3.68

vector of 1000 dimensions with the first 5 entries were 5, -4, 3, -2, 2 and the rest were zeros. We used the modified ranking algorithm with both forward and backward moves and the default prior (4.21). The performance was measured as before by CFR, NZC and MSE but MSE was defined as the squared error between the true vector β and its estimate. The simulation results are summarized in Table 4.13. The big improvement of the VAR over the aLasso in this example is surprising and probably due to the reasons discussed at the end of Section 4.2.2.

Application to the diabetes data. As an application, we applied the VAR method to analyzing a benchmark data set in the literature on progression of diabetes [Efron et al., 2004]. Ten baseline variables, age, sex, body mass index, average blood pressure and six blood serum measurements, were obtained for each of $n = 442$ diabetes patients, as well as the response of interest y, a quantitative measure of disease progression one year after baseline. We constructed a (heteroscedastic, if necessary) linear regression model to predict y from these ten input variables. In the hope of improving prediction accuracy, we considered a "quadratic model" with 64 predictors. We distinguish between input variables and predictors, for example, in a quadratic regression model on two input variables age and

Table 4.13: Homoscedastic case: CFR, MSE and NZC averaged over 100 replications for the aLasso and VAR.

		CFR (NZC)		MSE	
$n=n_P$	σ	aLasso	VAR	aLasso	VAR
50	1	0 (994.42)	38 (994.34)	31.21	17.72
	2	0 (994.54)	2 (992.36)	38.20	33.16
100	1	46 (995.62)	96 (994.96)	8.40	0.09
	2	16 (996.14)	32 (993.56)	11.86	2.08
200	1	90 (995.10)	98 (994.98)	6.34	0.04
	2	44 (995.56)	32 (993.40)	7.78	0.62

income, there are five predictors (age, income, age×age, income×income and age×income).

The analysis of the full data set showed clear evidence of heteroscedasticity. See again Figure 4.3 for the solution paths resulting from our VAR algorithm (only forward selection was implemented and the search for inclusion in the variance model was restricted). The VAR and GAMLSS both selected some predictors to include in the variance model. Furthermore, there was quite a clear pattern in the plot of the OLS studentized residuals indicating heteroscedasticity (results not shown). Interestingly, when fitting y with only ten input variables as the predictors, diagnostics and the selected model by VAR showed no evidence of heteroscedasticity. This result agreed with the homoscedasticity assumption often used in the literature for this diabetes data set.

To assess predictive performance, we randomly selected 300 instances to form the training set, with the remainder serving as the validation set. Of 64 predictors, the VAR selected 13 to include in the mean model and 12 to include in the variance model, while the GAMLSS selected 23 and 7, respectively. Under the assumption of constant variance,

the aLasso selected 43 predictors. On the validation set, the models estimated by the aLasso, GAMLSS and VAR had PPS of 5.50, 15.93, 5.41 and MSE of 3264.95, 3506.32, 2993.16, respectively. In order to reduce the uncertainty in training-validation separation, we averaged the MSE and PPS over 50 replications, and obtained the MSE for the aLasso, GAMLSS and VAR of 3560.50, 4843.40, 2970.67, and the PPS of 5.63, 59.76, 5.38, respectively. The GAMLSS method performed poorly in this example but it should be stressed that we have only used the default implementation (i.e., stepwise selection with both forward and backward moves and the generalized AIC used as the stopping rule) in the GAMLSS R package. Further experimentation with tuning parameters in the information criterion might produce better results.

Remarks on calculations. The VAR algorithm was implemented using R and the code is freely available on the author's website. The weights used in the aLasso were assigned as usual as $1/|\hat{\beta}_j|$ with $\hat{\beta}_j$ being the MLE (when $p<n$) or the Lasso estimate (when $p\geq n$) of β_j. The tuning parameter λ was selected by 5-fold cross-validation. The implementation of the aLasso and GAMLSS was carried out with the help of the R packages glmnet and gamlss.

4.2.4 Appendix

Below we write $E_q(\cdot)$ for an expectation with respect to the variational posterior. In the notation of Section 4.2.1 we have

$$\begin{aligned}
T_1 &= -\frac{p+q}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_\beta^0| - \frac{1}{2}\log|\Sigma_\alpha^0| \\
&\quad -\frac{1}{2}E_q((\beta-\mu_\beta^0)^T{\Sigma_\beta^0}^{-1}(\beta-\mu_\beta^0)) - \frac{1}{2}E_q((\alpha-\mu_\alpha^0)^T{\Sigma_\alpha^0}^{-1}(\alpha-\mu_\alpha^0)) \\
&= -\frac{(p+q)}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_\beta^0| - \frac{1}{2}\log|\Sigma_\alpha^0| \\
&\quad -\frac{1}{2}\mathrm{tr}({\Sigma_\beta^0}^{-1}\Sigma_\beta^q) - \frac{1}{2}\mathrm{tr}({\Sigma_\alpha^0}^{-1}\Sigma_\alpha^q) - \frac{1}{2}(\mu_\beta^q-\mu_\beta^0)^T{\Sigma_\beta^0}^{-1}(\mu_\beta^q-\mu_\beta^0) \\
&\quad -\frac{1}{2}(\mu_\alpha^q-\mu_\alpha^0)^T{\Sigma_\alpha^0}^{-1}(\mu_\alpha^q-\mu_\alpha^0), \\
T_2 &= -\frac{n}{2}\log 2\pi - \frac{1}{2}E_q(\sum_{i=1}^n z_i^T\alpha) - \frac{1}{2}E_q\left(\sum_{i=1}^n \frac{(y_i-x_i^T\beta)^2}{\exp(z_i^T\alpha)}\right) \\
&= -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^n z_i^T\mu_\alpha^q - \frac{1}{2}\sum_{i=1}^n \frac{x_i^T\Sigma_\beta^q x_i + (y_i-x_i^T\mu_\beta^q)^2}{\exp\left(z_i^T\mu_\alpha^q - \frac{1}{2}z_i^T\Sigma_\alpha^q z_i\right)}
\end{aligned}$$

and

$$\begin{aligned}
T_3 &= \frac{p+q}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_\beta^q| + \frac{1}{2}\log|\Sigma_\alpha^q| \\
&\quad +\frac{1}{2}E_q((\beta-\mu_\beta^q)^T{\Sigma_\beta^q}^{-1}(\beta-\mu_\beta^q)) + \frac{1}{2}E_q((\alpha-\mu_\alpha^q)^T{\Sigma_\alpha^q}^{-1}(\alpha-\mu_\alpha^q)) \\
&= \frac{p+q}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_\beta^q| + \frac{1}{2}\log|\Sigma_\alpha^q| + \frac{p+q}{2}.
\end{aligned}$$

In evaluating T_2 above we made use of the independence of β and α in the variational posterior and of the moment generating function for the multivariate normal variational posterior distribution for α. Putting the terms together, the variational lower bound simplifies to (4.17).

Table 4.14: A brief summary of some variable selection methods

Method	Description	Comment	References
♠ Subset selection, forward/backward selection	Search over all possible subsets with some criterion such as AIC, BIC, C_p as stopping rule.	Traditionally widely used, a sub-optimal model may be selected.	Miller [2002]
♠ Stochastic search variable selection	Based on a Bayesian hierarchy and Gibbs sampling.	Efficient, flexible to design. May be time demanding in high-dimensional cases.	George and McCulloch [1993], Smith and Kohn [1996]
♠ Lasso-type	Minimize an empirical loss with constraints, such as l_1, on the coefficients.	Efficient, a modern method, widely adopted. May cause bias on non-zero coefficients.	Tibshirani [1996], Fan and Li [2001]
♠ pLasso	A version of Lasso using the KL divergence to the full/BMA model instead of empirical loss.	Achieve a good predictive performance.	Proposed in this thesis
♠ BaLasso	An extension of Lasso using adaptive constraints on coefficients.	Efficient practically in achieving model selection consistency.	Proposed in this thesis
♠ VAR	Rank variables for inclusion via maximizing the VB lower bound.	Designed for heteroscedastic regression.	Proposed in this thesis

Chapter 5

Conclusions and future work

The thesis has approached the model selection problem from different angles and made some contributions to the model selection literature. This chapter gives some concluding remarks and discusses some open research questions raised from our works.

Calculation of the loss rank in the general case. As presented in Chapter 2, the LoRP is a general procedure for model selection whose main goal is to learn the underlying structure in the data. The LoRP can be regarded as a guiding principle for deriving model selection criteria that can avoid overfitting. This thesis has only scratched at the surface of this new methodology and discusses here several interesting questions that are worth investigating in future.

For non-linear regression we did not present an efficient algorithm for calculating the loss rank/volume $\mathrm{LR}_M(\mathbf{y}|\mathbf{x})$. This high-dimensional volume may be computed by Monte Carlo algorithms. Resampling techniques may be applied too.

A potential solution is as follows. Recall the definition of the loss rank of a model M with output data $\mathbf{y}$ and fixed input data $\mathbf{x}$

$$\mathrm{Rank}_M(\mathbf{y}|\mathbf{x}) = \mathrm{Volume}\{\mathbf{y}' \in \mathcal{Y}^n : \mathrm{Loss}_M(\mathbf{y}'|\mathbf{x}) \leq \mathrm{Loss}_M(\mathbf{y}|\mathbf{x})\}$$

where $\text{Loss}_M(\mathbf{y}|\mathbf{x})$ is the empirical loss associated with some loss function $l(.)$. Assume that the loss $\text{Loss}_M(\mathbf{y}|\mathbf{x})$ as a function of $\mathbf{y}$ is twice differentiable and that the Hessian $H = \partial^2 \text{Loss}_M(\mathbf{y}|\mathbf{x})/\partial\mathbf{y}\partial\mathbf{y}^\top$ is positive definite. Let $\mathbf{b} = \partial \text{Loss}_M(\mathbf{y}|\mathbf{x})/\partial\mathbf{y}$. Using Taylor's expansion

$$\text{Loss}_M(\mathbf{y}'|\mathbf{x}) = \text{Loss}_M(\mathbf{y}|\mathbf{x}) + \mathbf{b}^\top(\mathbf{y}' - \mathbf{y}) + \frac{1}{2}(\mathbf{y}' - \mathbf{y})^\top H(\mathbf{y}' - \mathbf{y}) + O(\|\mathbf{y}' - \mathbf{y}\|^3),$$

and ignoring the last term, the logarithm of the loss rank now can be approximately written as

$$\text{LR}_M(\mathbf{y}|\mathbf{x}) \approx \frac{n}{2} \log \mathbf{b}^\top H^{-1}\mathbf{b} + \frac{1}{2} \log(\det H^{-1}). \tag{5.1}$$

Note that, in the case of $\mathbf{y}$-linear regression as considered in Section 2.2, this approximation is exact. Investigation of (5.1) is currently in progress.

What is the "right" definition of model complexity?. Model selection can typically be regarded as the question of choosing the "right" model complexity. Many popular methods such as AIC and BIC define the complexity of a model as (to be proportional to) its number of free parameters df. This has also been generalized in some cases to the trace formula $\text{df} = \text{tr}(M)$ where M is a regression matrix [Hastie et al., 2001, Section 7.6]. This definition is nicely motivated and widely used but is not without problem, because it is not associated with the loss function as it should be. This definition results from using the minus log likelihood as the loss, what if a different loss function such as l_p-loss is used?

General speaking, a model is said to be complex if it can fit many data well, i.e., having small empirical fit. The fitness here must be measured by some loss function. Therefore, model complexity must be defined in association with a loss function, or in other words, model complexity should be loss-dependent. Besides loss-dependency, data-adaptivity is another desirable property for model complexity.

The LoRP offers a neat way to define model complexity which can be both loss-dependent and data-adaptive. By virtue of (5.1) and the results elsewhere in Chapter 2, it seems to be reasonable to define the complexity of a model M by

$$\text{Com}(M) \equiv \text{Com}(M|\mathbf{y}, l(.)) := \log\det(H^{-1}). \tag{5.2}$$

Intuitively, for a flexible M, the loss $\text{Loss}_M(\mathbf{y}|\mathbf{x})$ is small and stays fairly constant with changes in $\mathbf{y}$. As the result, the Hessian H will be "small", thus leading to a large $\text{logdet}(H^{-1})$. In some cases such as ridge regression, $\text{logdet}(H^{-1})$ has a closed form and a meaningful interpretation [Tran, 2009, Section 3.1]. Because model complexity plays an essential role in model selection, a careful investigation of $\text{logdet}(H^{-1})$ is necessary.

The POPMOS and the predictive Lasso. The procedure for model selection POPMOS with an explicit predictive motivation was described in Chapter 3. A variant of the POPMOS, the pLasso, has been shown to be convenient for variable selection and efficient in terms of prediction accuracy. A notable feature of the pLasso is that we put no restriction on the reference predictive distribution $p(\Delta|D)$. Although we have considered $p(\Delta|D)$ as arising from a full model including all potential covariates, it can in fact arise from any model where a GLM approximation with variable selection is desired. The approximation can also be an appropriately local one in the covariate space through a judicious choice of the design points in the pLasso criterion, which need not correspond to the observed design points. We have motivated and developed the idea of the pLasso only for GLMs. It is clear that this idea can be extended to other models rather than GLMs, and this is a topic for future research.

Variable selection in complicated frameworks. The variational approximation ranking algorithm VAR described in Chapter 4 is efficient for variable selection in high-dimensional heteroscedastic regression. The idea of ranking covariates for inclusion has

potential for extensions to much more complicated frameworks like Bayesian (grouped) variable selection in GLMs. Another potential research direction is to extend the method to simultaneous variable selection and number of experts selection in flexible regression density estimation with mixtures of experts. This research direction is currently in progress [Tran et al., 2011].

Bibliography

J. Aitchison. Goodness of prediction fit. *Biometrika*, 62:547–554, 1975.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory*, pages 267–281, Budapest, Hungary, 1973. Akademiai Kaidó.

D. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36:99–102, 1974.

S. Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624, 2009.

Y. F. Atchade. A computational framework for empirical Bayes inference. *Statistics and computing*, 2009. URL `www.stat.lsa.umich.edu/~yvesa/EB.pdf`. to appear.

C. Bailey. *Smart Exercise: Burning Fat, Getting Fit*. Boston: Houghton-Mifflin, 1994.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics, identifying influential data and sources of collinearity*. New York, John Wiley, 1980.

C. M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

K. P. Burnham and D. Anderson. *Model selection and multimodel inference : a practical information-theoretic approach*. New York, Springer, 2002.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, 35:2313–2351, 2007.

R. J. Carroll and D. Ruppert. *Transformation and Weighting in Regression*. Chapman and Hall, London, 1988. Monographs on Statistics and Applied Probability.

G. Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2:485–500, 2001.

A. Chambaz. Testing the order of a model. *The Annals of Statistics*, 34(3):1166–1203, 2006.

D. Chan, R. Kohn, D. J. Nott, and C. Kirby. Adaptive nonparametric estimation of mean and variance functions. *Journal of Computational and Graphical Statistics*, 15:915–936, 2006.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008.

M. H. Chen and J. G. Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, 13:461–476, 2003.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

R. Cottet, R. Kohn, and D. J. Nott. Variable selection and model averaging in overdispersed generalized linear models. *Journal of the American Statistical Association*, 103:661–671, 2008.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the methods of generalized cross- validation. *Numerische Mathematik*, 31:377–403, 1979.

N. J. Delaney and S. Chatterjee. Use of the bootstrap and cross-validation in ridge regression. *Journal of Business and Economics Statsitics*, 4(2):225–262, 1986.

D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, 57(1):45–97, 1995.

N. R. Draper and H. Smith. *Applied Regression Analysis*. New York, John Wiley, 1981.

R. M. Dudley and W. Philipp. Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. Verw. Gebiete*, 62:509–552, 1983.

B. Efron. Double exponential families and their use in generalised linear regression. *Journal of the American Statistical Association*, 81:709–721, 1986.

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

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

I. Ehrlich. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of Political Economy*, 81:521–565, 1973.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing*, 1: 586–598, 2007.

J. H. Friedman. Fast sparse regression and classification. *Technical report*, 2008. URL `http://www-stat.stanford.edu/ jhf/ftp/GPSpaper.pdf`.

M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66: 165–207, 2007.

P. H. Garthwaite, Y. Fan, and S. A. Sisson. Adaptive optimal scaling of Metropolis-Hastings algorithms using the Robbins-Monro process. *arXiv:1006.3690v1*, 2010.

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:329–328, 1975.

S. Geisser. Discussion of "Sampling and Bayes' inference in scientific modelling and robustness" by g.e.p. box. *Journal of the Royal Statistical Society A*, 143:416–417, 1980.

S. Geisser. *Predictive Inference: An Introduction*. New York: Chapman & Hall, 1993.

A. Gelman, A. Jakulin, P. Grazia, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383, 2008.

A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of American Statistical Association*, 88:881–889, 1993.

W. R. Gilks, D. J. Spiegelhalter, and S. Richardson. *Markov Chain Monte Carlo in practice*. London, Chapman & Hall, 1996.

E. Gine and J. Zinn. Bootstrapping general empirical functions. *The Annals of Probability*, 18: 851–869, 1990.

T. Gneiting and A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

G.H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society B*, 14:107–114, 1952.

J. E. Griffin and P. J. Brown. Bayesian adaptive lassos with non-convex penalization. *Technical report*, 2010. URL `http://www.kent.ac.uk/ims/personal/jeg28/NEG.pdf`.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570, 2000.

A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

A. E. Hoerl, R.W. Kennard, and K. F. Baldwin. Ridge regression: Some simulations. *Communications in statistics*, 4:105–123, 1975.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

C. C. Holmes and N. M. Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal of the Royal Statistical Society B*, 64(2):295–306, 2002.

C. M. Hurvich and C. L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

M. Hutter. The loss rank principle for model selection. In N. Bshouty and C. Gentile, editors, *Proc. 20th Annual Conf. on Learning Theory (COLT'07)*, volume 4539 of *LNAI*, pages 589–603, San Diego, 2007. Springer, Berlin.

M. Hutter and M.-N. Tran. Model selection with the loss rank principle. *Computational Statistics and Data Analysis*, 54(5):1288–1306, 2010.

R. W. Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4, 1996.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in Graphical Models. M. I. Jordan (Ed.)*. MIT Press, Cambridge, 1999.

R. E. Kass and S. Vaidyanathan. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B*, 54:129–144, 1992.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47:1902–1914, 2001.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors and Bayesian lassos. *Beyesian Statistics*, 5:369–412, 2010.

E. E. Leamer. *Specification searches*. New York, Wiley, 1978.

E. L. Lehmann and G. Casella. *Theory of point estimation (2nd ed.)*. New York: Springer, 1998.

C. Leng, M.-N. Tran, and D. J. Nott. Bayesian adaptive lasso. *Submitted*, 2010. arXiv:1009.2300v1.

C. Leng, Y. Lin Y, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16:1273–1284, 2006.

D. V Lindley. The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, Series B*, 30:31–66, 1968.

F. Lozano. Model selection using Rademacher penalization. In *Proc. 2nd ICSC Symp. Neural Computation NC2000*. Berlin, Germany: ICSC Academic, 2000.

D. Madigan and A. E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1994.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41:3397–3415, 1993.

P. Massart. *Concentration inequalities and model selection. Ecole d'Et de Probabilits de Saint-Flour*, volume 1896. Springer, 2007. Lecture Notes in Mathematics.

N. Meinshausen and P. Buhlmann. Consistent neighbourhood selection for high-dimensional graphs the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

X.-L. Meng. Posterior predictive p-values. *The Annals of Statistics*, 22:1142–1160, 1994.

A. Miller. *Subset Selection in Regression*. Chapman & Hall/CRC, 2002.

D. J. Nott, S. L. Tan, M. Villani, and R. Kohn. Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 2011. To appear.

D. J. Nott, M.-N. Tran, and C. Leng. Variational approximation for heteroscedastic linear models and matching pursuit algorithms. *Statistics and Computing*, 2010. To appear [Preprint: arXiv:1011.4832v3].

A. O'Hagan and J. J. Forster. *Bayesian Inference*. Arnold, London, 2004.

J. T. Ormerod and M. P. Wand. Explaining variational approximation. *The American Statistician*, 64(2):140–153, 2010.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20:389–404, 2000.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

B. M. Poetscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100:2065–2082, 2009.

A. E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statatiscal Association*, 92(437), 1997.

A. Reusken. Approximation of the determinant of large sparse symmetric positive definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 23(3):799–818, 2002.

J. A. Rice. *Mathematical Statistics and Data Analysis*. California: Duxbury Press, 1995.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, 54:507–554, 2005.

J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *Annals of the Institute of Statistical Mathematics*, 35:415–423, 1983.

M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.

G. Smyth. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society B*, 51:47–60, 1989.

T. Stamey, J. Kabalin, J. McNeal, I. Johnstone, F. Freiha, E. Redwine, , and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii. radical prostatectomy treated patients. *Journal of Urology*, 16:1076–1083, 1989.

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

R. H. Taplin. Robust likelihood calculation for time series. *Journal of the Royal Statistical Society B*, 55:829–836, 1993.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

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

M.-N. Tran. Penalized maximum likelihood principle for choosing ridge parameter. *Communications in Statistics - Simulation and Computation*, 38:1610–1624, 2009.

M.-N. Tran. A criterion for optimal predictive model selection. *Communications in Statistics - Theory and Methods*, 40:893–906, 2011a.

M.-N. Tran. The loss rank criterion for variable selection in linear regression analysis. *Scandinavian Journal of Statistics*, 38(3):466–479, 2011b.

M.-N. Tran and M. Hutter. Model selection by loss rank for classification and unsupervised learning. *arXiv:1011.1379v1*, 2010.

M.-N. Tran, D. J. Nott, and R. Kohn. Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. *Working paper*, 2011.

M.-N. Tran, D. J. Nott, and C. Leng. The predictive lasso. *Statistics and computing*, 2010. To appear.

J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

V. N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probability and Its Application*, 16:264–280, 1971.

H. Wang and C. Leng. Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

H. Wang and C. Leng. A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52:5277–5286, 2008.

H. Wang, R. Li, and C. L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 3(94):553–568, 2007.

Y. Yang. Can the strengths of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

P. Yau and R. Kohn. Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13:191–208, 2003.

T. Yee and C. Wild. Vector generalized additive models. *Journal of the Royal Statistical Society, Series B*, 58:481–493, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68:49–67, 2006.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques: Essays in honour of Bruno De Finetti*, pages 233–243. North-Holland, Amsterdam, 1986.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497, 2009.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao and B. Yu. Stagewise lasso. *Journal of Machine Learning Research*, 8:2701–2726, 2007.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou, T. Hastie, and R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statitics*, 35:2173–2192, 2007.