DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1093

July 6, 2004

Automatic Smoothing and Variable Selection via Regularization[1]

Ming Yuan [2]

Department of Statistics, University of Wisconsin, Madison WI

[1]**Research supported in part by NSF Grant DMS-0772292 and NASA Grant NAG5-1073**

[2]Email: yuanm@stat.wisc.edu

AUTOMATIC SMOOTHING AND VARIABLE SELECTION VIA REGULARIZATION

By

Ming Yuan

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2004

Abstract

This thesis focuses on developing computational methods and the general theory of automatic smoothing and variable selection via regularization.

Methods of regularization are a commonly used technique to get stable solution to ill-posed problems such as nonparametric regression and classification. In recent years, methods of regularization have also been successfully introduced to address a classical problem in statistics, variable selection.

Smoothing parameters are introduced in regularization to balance the trade-off between the goodness-of-fit to the data such as the log likelihood, and the prior knowledge on the unknown such as the degree of smoothness. A successful regularization method requires an objective way to tune the smoothing parameter involved.

In this thesis, we demonstrate the use of methods of regularizations and different approaches for tuning parameter selection in three different settings, namely nonparametric Gaussian heteroscedastic regression, nonparametric Poisson regression, and variable selection for normal linear models. We start with the nonparametric heteroscedastic regression where both the conditional mean and variance are assumed unknown. A penalized log likelihood method, doubly penalized log likelihood estimate, is proposed to estimate both the conditional

mean and variance nonparametrically. Cross-validation based methods for tuning the smoothing parameters are also suggested. The technique is illustrated through a set of Monte Carlo simulations.

In the second part of the thesis, we consider the nonparametric Poisson regression. We recommend the use of an exact unbiased risk estimate for tuning the smoothing parameter. To reduce the computational cost, we also suggest an accurate approximation to the exact unbiased risk estimator which can be used in practice.

In the last part of the thesis, we propose an empirical Bayes method for variable selection and coefficient estimation in linear regression models. The method is based on a particular hierarchical Bayes formulation, and the empirical Bayes estimator is shown to be closely related to a penalized least squares estimate, the LASSO estimator. Such a intimate connection allows us to take advantage of the recently developed quick LASSO algorithm to compute the empirical Bayes estimate, and provides a new way to select the tuning parameter in the LASSO method. Unlike previous empirical Bayes variable selection methods, which in most practical situations can only be implemented through a greedy stepwise algorithm, our method gives a global solution efficiently. We also show how to assess model uncertainty by estimating the posterior probabilities of the models and the variables. Simulations show that the proposed method is very competitive in terms of variable selection, estimation accuracy, and computation speed when compared with other variable selection and estimation methods.

Acknowledgements

It has been an honor and a privilege to work with my advisor, Grace Wahba, and it is an opportunity for which I am grateful. Her support and guidence have sparked my interest in reproducing kernel methods and machine learning. I also express my sincerest gratitude to my coadvisor, Yi Lin. Stimulating discussions with Yi have always been a source of inspiration, which has greatly influenced the direction and interests of my research. This thesis would not be possible without discussions and insights from Grace and Yi.

I am thankful to Christina Kendziorski for introducing me into microarray data analysis which is such an exciting area to work on. Collaborations with Zhengyu Liu and Dan Vimont have also been a great motivation for the research work presented in this thesis and beyond.

I am grateful to the members of my thesis committee, Kjell Doksum, Christina Kendziorski, Yi Lin, Zhengyu Liu, and Grace Wahba. I would also like to thank my references, Christina Kendziorski, Bret Larget, Yi Lin, and Grace Wahba, for writing countless letters in behalf of me. I also thank the staff at Statistics Department, especially Denise Roder and Jude Grudzina, for their help in various ways during these years.

I would like to acknowledge my friends in Madison, and fellow graduate students in the Statistics Department, too many to name, for making my graduate

experience much more colorful and enjoyable. Especially, I am thankful to the former and current graduate students in the "Thursday group" for making up such a dynamic research environment.

I owe much more than I can express here to my parents and family-in-law for their unconditional love and support. I dedicate this thesis to Jiaqin, whose companionship has made my life so much better in countless ways.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The main focus of this dissertation research is to develop statistical methods for automatic smoothing and variable selection.

Nonparametric regression or classification allows the unknown parameter to reside in an infinitely dimensional space such as a reproducing kernel Hilbert space. In doing so, we face an ill-posed problem. Methods of regularization are a commonly used technique to get a stable solution to ill-posed problems. Popular examples include smoothing spline analysis of variance models. In recent years, methods of regularization have also been successfully introduced to address a classical problem in statistics, variable selection. The idea of regularization is to weigh the fidelity of the data measured by certain loss functions and a penalty on the unknown quantity. Smoothing parameters are often introduced to balance this tradeoff and they considerably affect the performance of an estimate from regularization. Ideally, we want to choose a smoothing parameter which minimizes the loss of the associated estimate, for example the Kullback Leibler distance to the truth. Since the risk is related to the truth which is not directly obtainable, it can only be controlled indirectly. There usually are three options to achieve this goal. We can minimize an unbiased estimate of the risk, or an

approximately unbiased estimate of the risk such as cross-validation. We can also give the regularization a Bayesian interpretation and choose a smoothing parameter in a Bayesian fashion, for example by maximizing the marginal likelihood. Each option should be tailored depending on the problem. In this thesis, we demonstrate the usage of regularization and the aforementioned approaches for smoothing parameter selection in three different settings: nonparametric Gaussian heteroscedastic regression, smoothing spline based Poisson regression and penalized least squares in normal linear regression model.

Chapter 2 concerns the nonparametric heteroscedastic regression problem. A penalized likelihood estimation procedure is developed for heteroscedastic regression. A distinguishing feature of the new methodology is that it estimates both the mean and variance functions simultaneously without parametric assumption for either. An efficient implementation of the estimating procedure and cross-validation based smoothing parameter tuning methods are also provided. The procedure is illustrated by a Monte Carlo example. A potential generalization, and application to the covariance modeling problem in numerical weather prediction is noted.

It is attempting to think cross-validation as a panacea to all smoothing parameter tuning problems in the light of its successes on various grounds. However, in Chapter 3, we find that the unbiased risk estimator based tuning method might be preferable over the cross-validation based methods for nonparametric Poisson regression. We recommended the use of an exact unbiased risk estimate

for smoothing spline based Poisson regression. A computable approximation of the unbiased risk estimate (AUBR) for Poisson regression is introduced. This approximation can be used to automatically tune the smoothing parameter for the penalized likelihood estimator. An alternative choice is the generalized approximate cross validation (GACV) proposed by Xiang and Wahba (1996). Although GACV enjoys a great success in practice when applying for nonparametric logistic regression, its performance for Poisson regression is not clear. Numerical simulations have been conducted to evaluate the GACV and AUBR based tuning methods.We found that GACV has a tendency to oversmooth the data when the intensity function is small. As a consequence, we suggest tuning the smoothing parameter using AUBR in practice.

In Chapter 4, we propose an empirical Bayes method for variable selection and coefficient estimation in linear regression models. The method is based on a particular hierarchical Bayes formulation, and the empirical Bayes estimator is shown to be closely related to the LASSO estimator. Such a connection allows us to take advantage of the recently developed quick LASSO algorithm to compute the empirical Bayes estimate, and provides a new way to select the tuning parameter in the LASSO method. Unlike previous empirical Bayes variable selection methods, which in most practical situations can only be implemented through a greedy stepwise algorithm, our method gives a global solution efficiently. We also show how to assess model uncertainty by estimating the posterior probabilities of the models and the variables. Simulations show that the

proposed method is very competitive in terms of variable selection, estimation accuracy, and computation speed when compared with other variable selection and estimation methods.

The dissertation is concluded by a few remarks on future works in Chapter 5.

Chapter 2

Doubly Penalized Likelihood Estimator in Heteroscedastic Regression

2.1 Heteroscedastic Regression Model

In this chapter, we consider the following heteroscedastic regression model

$$y_i = \mu(x_i) + \sigma(x_i)\varepsilon_i, \qquad i = 1, 2, ..., n, \tag{2.1.1}$$

where μ and σ are unknown functions to be estimated, $x_i, i = 1, \ldots, n$ are univariate or multivariate covariates and $\varepsilon's$ are i.i.d. noise with mean 0 and variance 1. The focus of most nonparametric regression methodologies centers around the conditional mean. However, the understanding of the local variability of the data, measured by the conditional variance is very important in many scientific studies (e.g, Andersen and Lund, 1997; Gallant and Tauchen, 1997). These applications necessitate the development of new nonparametric techniques that allow the modeling of varying variance.

Two step procedures are commonly used. First, the conditional mean is estimated. Then, an estimate of the conditional variance is constructed based on the regression residuals. Earlier proposals include Carroll (1982), Muller and Stadtmuller (1987), Hall and Carroll (1989), Ruppert *et. al.* (1997) and Fan and Yao (1998) among many others. More recently, a Bayesian treatment to (2.1.1) has also been introduced by Yau and Kohn (2003).

In this chapter, an alternative approach is proposed. We introduce a penalized likelihood based procedure which can take the heteroscedasticity into account and estimate both mean function and variance function simultaneously.

The rest of the chapter is organized as follows. The DPLE is introduced in the next section. Then it is expressed as a nonlinear optimization problem and a practical estimating procedure is given in Section 2.3. In Section 2.4, we construct the Bayesian confidence interval for the DPLE. The chapter is concluded by a Monte Carlo example.

2.2 Doubly Penalized Likelihood Estimator

By assuming that $\varepsilon_i \sim_{iid} N(0,1)$, we can write down the average negative log likelihood of (2.1.1), which, up to a constant not depending on μ or σ^2, is

$$L(\mu,\sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(y_i-\mu(x_i))^2}{2\sigma^2(x_i)} + \frac{1}{2}\log\sigma^2(x_i)\right)$$

Like other nonparametric settings, instead of assuming a parametric form of μ and σ^2, we allow them to reside in some Hilbert spaces $\mathcal{H}_\mu, \mathcal{H}_{\sigma^2}$ of smooth

functions. Using the penalized likelihood strategy, we may define our estimators $(\widehat{\mu}, \widehat{\sigma}^2)$ to be the minimizer of the following penalized negative log likelihood function over the corresponding functional spaces:

$$L(\mu, \sigma) + \frac{\lambda_\mu}{2} J_\mu(\mu) + \frac{\lambda_{\sigma^2}}{2} J_{\sigma^2}\left(\sigma^2\right), \tag{2.2.1}$$

where both penalty functions J_μ and J_{σ^2} are quadratic forms defined on $\mathcal{H}_\mu, \mathcal{H}_{\sigma^2}$ respectively and smoothing parameters λ_μ and λ_{σ^2} are nonnegative constants which control the tradeoff between L, the goodness-of-fit on the data, and the smoothness penalties J_μ and J_{σ^2}.

To work around the positivity constraint of σ^2, write

$$\sigma^2(x) = e^{g(x)}. \tag{2.2.2}$$

Then we assume that g lies in a Hilbert space $\mathcal{H}_g$ and re-express the penalized negative likelihood as

$$T_n(\mu, g; y, \lambda_\mu, \lambda_g) \equiv \frac{1}{n} \sum_{i=1}^{n} \left((y_i - \mu(x_i))^2 e^{-g(x_i)} + g(x_i) \right) + \lambda_\mu J_\mu(\mu) + \lambda_g J_g(g). \tag{2.2.3}$$

If both smoothing parameters are big enough, then it is not hard to see that (2.2.3) is convex in (μ, g) jointly. But if the λ's are small enough it can be seen, by looking at the Hessian, that bivariate convexity cannot be guaranteed. A practically useful statement concerning the exact nature of "big enough" is a subject of ongoing research. However, in a number of realistic simulations, for example, if the actual variance varies slowly compared to the data density, then

our simulations suggest that λ's of interest are "big enough", and possible lack of bivariate convexity is not a problem.

It is interesting to consider some special cases of the proposed DPLE. If we choose $\lambda_\mu = \infty$, then we end up with a model having a parametric conditional mean and a nonparametric conditional variance. A special case of this where $J_\mu(\mu) = \int (\mu'')^2$ has been treated by Carroll (1982) using the kernel smoother.

If we choose $\lambda_g = \infty$, then the conditional variance of model (2.1.1) is forced to take a parametric form. This includes the usual nonparametric regression model with constant variances in which our DPLE becomes the well-known smoothing spline estimator.

If we choose $\lambda_\mu = \lambda_g = \infty$, then our DPLE boils down to a maximum likelihood estimate for a parametric model where both μ and g lie in the null spaces in the sense that $J_\mu(\mu) = J_g(g) = 0$, for some penalty functions J_μ and J_g. A case of common interest is $J_\mu(\mu) = \int (\mu'')^2$. The null space corresponding to this penalty function is the usual linear model for μ. Jobson and Fuller (1980) argued that the maximum likelihood estimator dominates the simple least squares based estimators in this setting.

2.3 Algorithm

A multiple functional extension of the representer theorem of Kimeldorf and Wahba (1971) ensures that the minimizer of (2.2.3) lies in a finite dimensional space, even when the minimization is carried out in infinite dimensional Hilbert

spaces.

For brevity, let $\mathcal{H} \equiv \mathcal{H}_\mu = \mathcal{H}_g$ and $J \equiv J_\mu = J_g$. However, it is worth noting that in some applications, penalty functions J_μ and J_g could be different. Actually, μ and g could even be defined on entirely different sets of covariates.

Penalty J induces a decomposition of $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is the null space of J. Assume that $\{\phi_v : 1 \le v \le M\}$ are the basis functions of $\mathcal{H}_0$ and $K(\cdot,\cdot)$ is the reproducing kernel of $\mathcal{H}_1$. Then the representer theorem guarantees the minimizer of (2.2.3) to be

$$\mu(x) = \sum_{i=1}^{M} d_i^\mu \phi_i(x) + \sum_{i=1}^{n} c_i^\mu K(x_i, x), \tag{2.3.1}$$

$$g(x) = \sum_{i=1}^{M} d_i^g \phi_i(x) + \sum_{i=1}^{n} c_i^g K(x_i, x), \tag{2.3.2}$$

for some vectors c_μ, d_μ, c_g, d_g, where $d_\mu = (d_1^\mu, \cdots, d_M^\mu)'$, $c_\mu = (c_1^\mu, \cdots, c_n^\mu)'$ and c_g, d_g are defined in the same fashion.

For the ease of exposition, we shall use the matrix notation: Let T be a $n \times M$ matrix with the (i,ν)th entry $\phi_\nu(x_i)$ and $y = (y_1, \cdots, y_n)'$. With some abuse of notation, let $\mu = (\mu_1, \cdots, \mu_n)' \equiv (\mu(x_i), \cdots, \mu(x_n))'$, $g = (g_1, \cdots, g_n)' \equiv (g(x_i), \cdots, g(x_n))'$, and K be the $n \times n$ matrix with (i,j)th entry $K(x_i, x_j)$. Rewrite (2.3.1) and (2.3.2) as

$$\mu = Kc_\mu + Td_\mu, \qquad g = Kc_g + Td_g. \tag{2.3.3}$$

The variational problem (2.2.3) now becomes finding (c_μ, d_μ, c_g, d_g) to minimize

$$\Phi(d_\mu, c_\mu, d_g, c_g) \equiv (y - Td_\mu - Kc_\mu)' D (y - Td_\mu - Kc_\mu) + e'g + n\lambda_\mu c_\mu' K c_\mu + n\lambda_g c_g' K c_g, \tag{2.3.4}$$

where D is a diagonal matrix with the ith diagonal element $e^{-g(x_i)}$ and $e = (1, \cdots, 1)'$. A clue for minimizing (2.3.4) is that it is a convex function of each of μ and g by fixing the other. Thus, we can minimize (2.3.4) with respect to μ and g in an iterative fashion.

Fixing g, (2.3.4) reduces to a penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n} d_{ii}(y_i - \mu(x_i))^2 + \lambda_\mu J(\mu), \tag{2.3.5}$$

where d_{ii} is the (i, i) entry of diagonal matrix D. The solution to (2.3.5) is

$$\widehat{\mu} = A_\mu(\lambda_\mu)y, \tag{2.3.6}$$

where A_μ is the so-called hat matrix (see Wahba, 1990). In this step, only smoothing parameter λ_μ is involved. A commonly used smoothing parameter tuning technique is to choose a λ_μ which minimizes the GCV score:

$$V(\lambda_\mu) = \frac{n^{-1}y'D^{1/2}\left(I - A_\mu(\lambda_\mu)\right)^2 D^{1/2}y}{\left[n^{-1}tr\left(I - A_\mu(\lambda_\mu)\right)\right]^2}. \tag{2.3.7}$$

Fixing μ, the objective functional becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left(z_i e^{-g(x_i)} + g(x_i)\right) + \lambda_g J(g), \tag{2.3.8}$$

where $z_i = (y_i - \tilde{\mu}(x_i))^2$ and $\tilde{\mu}$ is the current estimate of μ. It is worth noting that (2.3.8) has the form of a penalized Gamma likelihood as if $z_i, i = 1, \ldots, n$ were independent samples from Gamma distributions with shape parameter 1 and scale parameters $e^{g(x_i)}, i = 1, \ldots, n$. This connection makes it possible to apply the general methodology for solving penalized likelihood problems with

responses from exponential family. For a fixed smoothing parameter λ_g, (3.8) is strictly convex in g and thus can be minimized using the Newton iteration. In this step, the objective functional only contains λ_g. We tune λ_g by minimizing the generalized approximate cross validation (GACV) technique developed by Xiang and Wahba (1996):

$$GACV(\lambda_g) = \frac{1}{n}\sum_{i=1}^{n}\left(z_i e^{-g(x_i)} + g(x_i)\right) + \frac{tr\left(A_g(\lambda_g)\right)}{n}\sum_{i=1}^{n}\frac{z_i e^{-\widehat{g}(x_i)}\left(z_i - e^{-\widehat{g}(x_i)}\right)}{1 + tr\left(HA_g(\lambda_g)\right)}, \tag{2.3.9}$$

where H is a diagonal matrix with (i,i) entry $e^{-\widehat{g}(x_i)}$ and $A_g(\lambda_g)$ is the influence matrix of (2.3.8) (see Xiang and Wahba, 1996).

Summing up, an algorithm which chooses the smoothing parameters automatically is as follows.

---

1. *Initialize c_g, d_g.*
2. *(i) Given the current c_g, d_g, choose λ_μ such that $V(\lambda_\mu)$ is minimized.*
   (ii) Given the current c_g, d_g and λ_μ, update c_μ, d_μ by minimizing (2.3.4) with respect to c_μ and d_μ.
3. *(i) Given the current c_μ, d_μ, choose λ_g such that $GACV(\lambda_g)$ is minimized.*
   (ii) Given the current c_μ, d_μ and λ_g, update c_g, d_g by minimizing (2.3.4) with respect to c_g and d_g.
4. *Iterate Step 2 and Step 3 till convergence.*

In our simulations, we found that extremely large negative intermediate estimates of g in the iteration often lead to overfitting and numerical instability. To avoid this problem, one could slightly perturb those near zero residuals. For example, in our simulation, we used $z_i = \max(0.00001, (y_i - \tilde{\mu}(x_i))^2)$.

The algorithm converges fairly fast according to our experience. Usually, it converges after 4-5 iterations between Step 2 and 3 if one starts from $c_g = 0$ and $d_g = 0$. It is also worth noting that the estimate of μ after the first iteration is just the usual unweighted μ estimation if we choose $c_g = 0$ and $d_g = 0$ as the initial value.

2.4 Bayesian Confidence Interval

Approximating Bayesian inference is often used to construct a confidence interval for penalized likelihood estimators. Usually, the penalty term can be interpreted as a prior and the penalized likelihood estimator is then equivalent to a maximum posterior estimator. Applying the Laplace approximation to the corresponding posterior distribution of the unknown function around this posterior mode, we can get an approximating Bayesian confidence interval. For details, see Wahba (1990) and Gu (2002).

Similarly, here we can regard penalty terms as priors for μ and g respectively.

More precisely, let the prior distributions be

$$\begin{aligned} d_\mu &\sim N(0, \tau_1^2 I_{(M)}), \qquad c_\mu \sim N\left(0, b_1^2 K\right) \\ d_g &\sim N(0, \tau_2^2 I_{(M)}), \qquad c_g \sim N\left(0, b_2^2 K\right) \end{aligned}$$

Using the argument for smoothing splines (Wahba, 1978), we can characterize our DPLE as a maximum a posteriori estimator.

Lemma 2.4.1 *Let $\tau_1^2, \tau_2^2 \to \infty$, the DPLE defined in the last section is a posterior mode of (μ, g) given y with*

$$\lambda_\mu = 1/nb_1^2 \qquad \textit{and} \qquad \lambda_g = 1/nb_2^2.$$

To construct the approximate Bayesian confidence intervals for each of μ and g, we will let the other be fixed. Let us first consider the Bayesian formulation for μ. For fixed smoothing parameters, the approximate solution for μ is

$$\mu(\cdot) = \phi(\cdot)' d_\mu + \xi(\cdot)' c_\mu$$

where $\phi = (\phi_1, \ldots, \phi_M)'$ and $\xi(\cdot) = (K(x_1, \cdot), \ldots, K(x_n, \cdot))'$. Using the improper prior for (c, d) as in Lemma 2.4.1, we have

$$p(c_\mu, d_\mu) \propto \exp\left(-\frac{n\lambda_\mu}{2} c_\mu' K c_\mu\right).$$

Thus, given g, the posterior distribution of c_μ, d_μ would be

$$p(c_\mu, d_\mu | y) \propto \exp\left(-\frac{1}{2}(y - Td_\mu - Kc_\mu)' D (y - Td_\mu - Kc_\mu) - \frac{n\lambda_\mu}{2} c_\mu' K c_\mu\right).$$

This suggest that the posterior distribution of (c_μ, d_μ) given y is a multivariate normal distribution. Hence

$$E\left(\mu(x)|y,\lambda_\mu\right) = (\xi(x)',\phi(x)')Q_\mu^{-1}\begin{pmatrix} K \\ T \end{pmatrix} Dy; \qquad (2.4.1)$$

$$Cov\left(\mu(x),\mu(x^1)|y,\lambda_\mu\right) = (\xi(x)',\phi(x)')Q_\mu^{-1}\begin{pmatrix} \xi(x^1) \\ \phi(x^1) \end{pmatrix} \qquad (2.4.2)$$

where

$$Q_\mu = \begin{pmatrix} KDK + n\lambda_\mu K & KDT \\ T'DK & T'DT \end{pmatrix}.$$

Similarly, for fixed smoothing parameters, the approximate solution for g is

$$g(\cdot) = \phi(\cdot)'d_g + \xi(\cdot)'c_g.$$

Given μ, the posterior distribution of (c_g, d_g) could be approximated by

$$p(c_\mu, d_\mu|y) \propto \exp\left(-\frac{1}{2}(\tilde{z} - Td_g - Kc_g)'W(\tilde{z} - Td_g - Kc_g) - \frac{n\lambda_g}{2}c_g'Kc_g\right),$$

where $\tilde{z}$ is the response used for the last Newton iteration to minimize (2.3.8) (see Gu, 2002) and W is a diagonal matrix with the (i,i)th entry $\tilde{z}_i e^{-g(x_i)}$. Thus,

$$E\left(g(x)|y,\lambda_g\right) \approx (\xi(x)',\phi(x)')Q_g^{-1}\begin{pmatrix} K \\ T \end{pmatrix} W\tilde{z}; \qquad (2.4.3)$$

$$Cov\left(g(x),g(x^1)|y,\lambda_g\right) \approx (\xi(x)',\phi(x)')Q_g^{-1}\begin{pmatrix} \xi(x^1) \\ \phi(x^1) \end{pmatrix} \qquad (2.4.4)$$

where

$$Q_g = \begin{pmatrix} KWK + n\lambda_g K & KWT \\ T'WK & T'WT \end{pmatrix}.$$

2.5 An Example

In this section, we will demonstrate the proposed method through a simple example. In this example, we consider three different sample sizes: $n = 125, 250, 500$. For each sample size, one hundred datasets were generated using the following setup.

$$x_i = (i - 0.5)/n, \qquad i = 1, \ldots, n; \tag{2.5.1}$$

$$y_i \sim N\left(\mu(x_i), exp(g(x_i))\right), \tag{2.5.2}$$

where

$$\begin{aligned} \mu(x) &= 2\left[\exp\left(-30(x-0.25)^2\right) + \sin(\pi x^2)\right] - 2, \\ g(x) &= \sin(2\pi x). \end{aligned} \tag{2.5.3}$$

For each dataset, the DPLE was calculated using the algorithm given in the last section. The comparative Kullback-Leibler (CKL) distance is used to measure the performance of the estimators. More precisely, for an estimator $\widehat{\mu}$ of the mean function μ, the CKL distance from $\widehat{\mu}$ to μ is defined as

$$CKL(\widehat{\mu}, \mu) = \frac{1}{n}\sum_{i=1}^{n}\left(\mu(x_i) - \widehat{\mu}(x_i)\right)^2 \sigma(x_i)^{-2}.$$

Similarly, the CKL distance from an estimator $\widehat{g}$ to g is defined as

$$CKL(\widehat{g}, g) = \frac{1}{n}\sum_{i=1}^{n}\left[e^{g(x_i)-\widehat{g}(x_i)} + \widehat{g}(x_i) - g(x_i)\right],$$

which differs from the KL distance between $\widehat{g}$ and g only by a constant not depending on $\widehat{g}$. Figure 2.1 gives the true test functions together with their $5th, 50th$ and $95th$ best fits for different sample sizes.

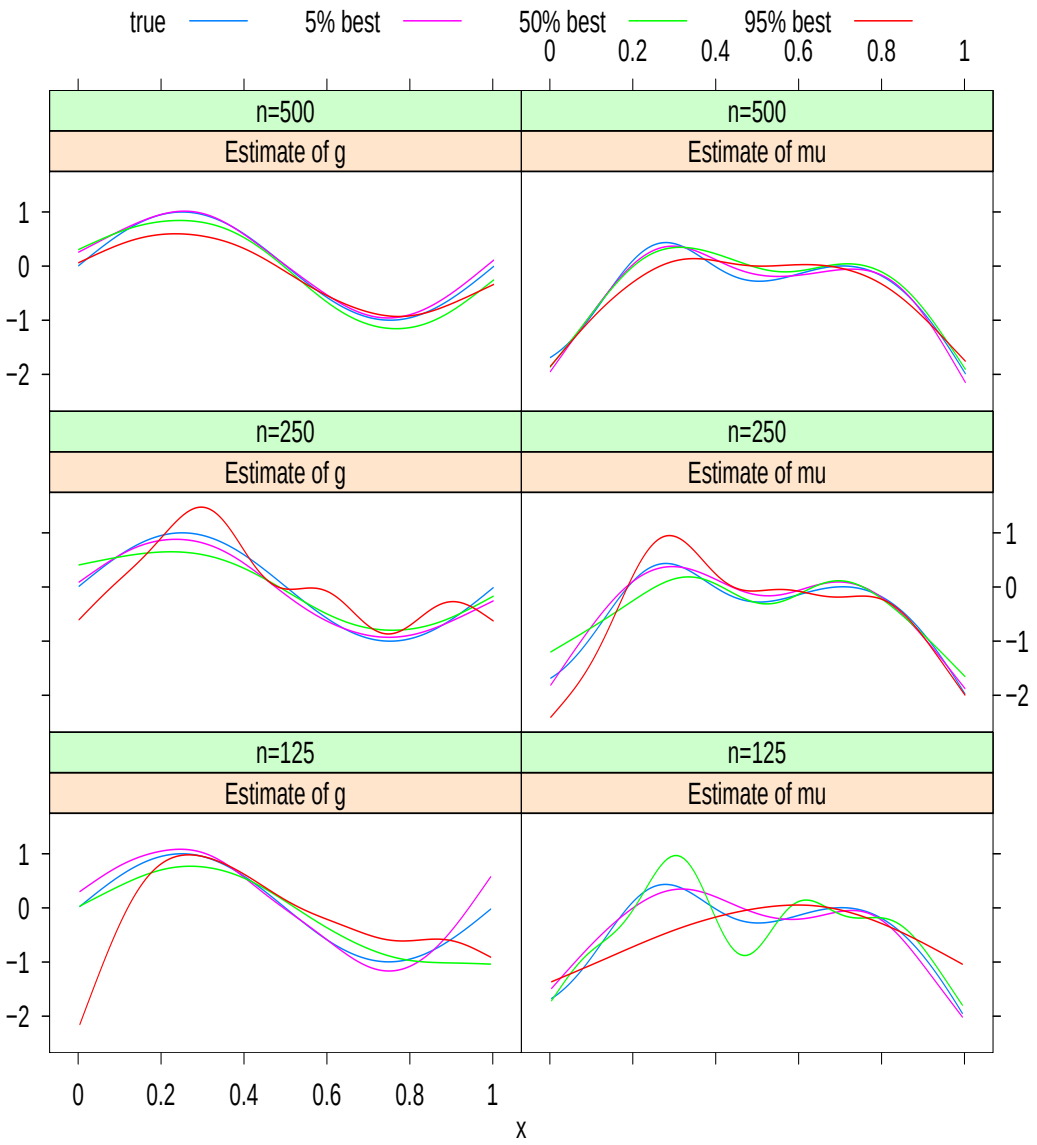

Figure 2.1: Finite-sample properties of DPLE

From Figure 2.1, we observe that the DPLE performs very well in most examples for sample size as small as 125. When the sample size increases, the performance improves. For sample size 500, even the 95th best fits are quite close to the truth.

Next, we investigate the empirical coverage of the Bayesian confidence intervals derived in Section 2.4. To this end, we repeated the above simulation with a fixed sample size 200. One hundred datasets were generated. For each dataset, we computed the Bayesian confidence interval at each sample point $x_i, i = 1, \ldots, n$. We define the empirical coverage for each sample point as the percentage of datasets, for which the Bayesian confidence intervals successfully cover the true test functions.

Figure 2.2 gives the estimators and their 95% Bayesian confidence interval for a typical dataset.

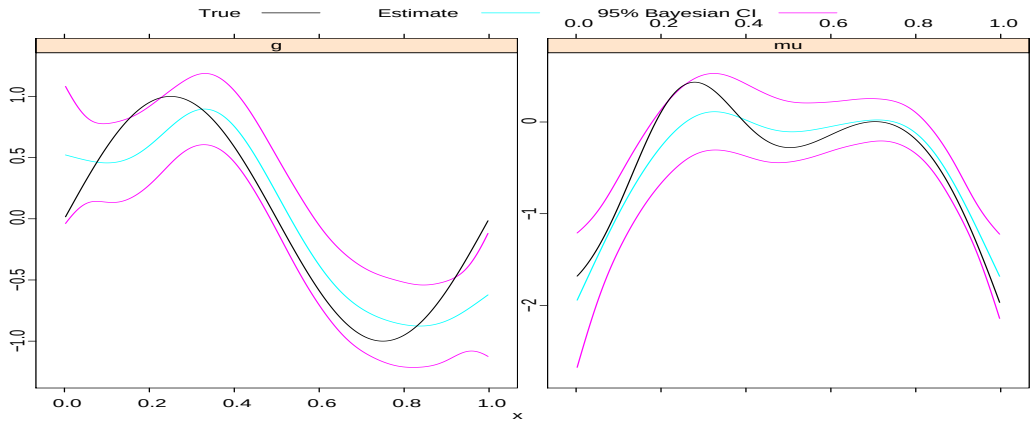

Figure 2.2: Bayesian confidence intervals for DPLE

We also plotted the true test functions together with the empirical coverage probabilities on each sample points in figure 2.3. We find that 95% coverage is not achieved for each sample point. Instead, the overall coverage is approximately 95%. This phenomenon is called across-the-functions coverage, which is well understood for the smoothing splines (see Wahba, 1990 and Gu, 2002).

Over all the sample points, define the average coverage proportion by

$$\begin{aligned} ACP_\mu(\alpha) &= \frac{1}{n}\#\{i : |\widehat{\mu}(x_i) - \mu(x_i)| < z_{\alpha/2}\sqrt{\widehat{var}\,(\widehat{\mu}(x_i)|y)}\} \\ ACP_g(\alpha) &= \frac{1}{n}\#\{i : |\widehat{g}(x_i) - g(x_i)| < z_{\alpha/2}\sqrt{\widehat{var}\,(\widehat{g}(x_i)|y)}\}, \end{aligned}$$

where $\widehat{var}$ stands for the approximate Bayesian posterior variance obtained in Section 2.4 and $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution.

In the current example, we have

$$ACP_\mu(0.05) = 93.1\%, \qquad ACP_g(0.05) = 92.5\%.$$

These coverages are slightly lower than 95%. This is reasonable because the approximate Bayesian confidence interval derived in Section 2.4 assumes that the smoothing parameters are fixed. But in our example, we automatically tuned the smoothing parameters instead of fixing them, which increases the variability of the DPLE.

These observations confirm that the Bayesian confidence interval constructed in Section 2.4 is valid.

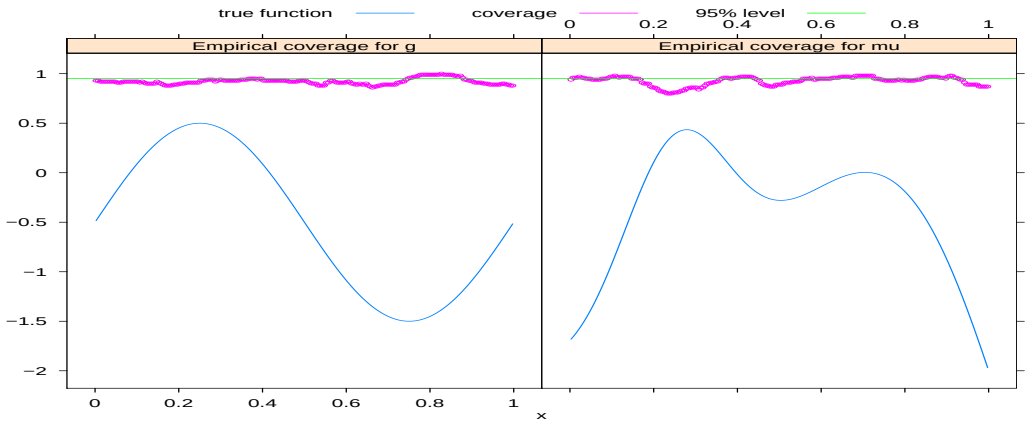

Figure 2.3: Coverage of Bayesian confidence intervals for DPLE

Chapter 3

Automatic Smoothing for Poisson Regression

3.1 Introduction

Poisson regression is widely used to model the event count data (see Vermunt, 1996). A nonparametric estimate of the canonical parameter of a Poisson process based on penalized likelihood smoothing spline models was proposed by O'Sullivan, Yandell and Raynor (1986). In this chapter, we are concerned with the adaptive choice of smoothing parameters in their smoothing spline models for nonparametric Poisson regression.

Suppose that $y_1, ..., y_n$ are n independent observations from a Poisson process with density of the form

$$f\left(y_i, \eta_0\left(x_i\right)\right) = \exp\left[y_i \eta_0\left(x_i\right) - e^{\eta_0(x_i)} - \ln\left(y_i!\right)\right], \tag{3.1.1}$$

where η_0 is the so-called canonical parameter and $\{x_i, i = 1, 2, ..., n\}$ are either univariate or multivariate covariates. We are interested in estimating $\eta_0\left(\cdot\right)$. When η_0 is of a linear form, it is the usual generalized linear model with a Poisson

likelihood (McCullagh and Nelder, 1983). In this case, since a parametric form of η_0 is assumed, maximum likelihood methods may be used to estimate and asses the fitted models. Although the generalized linear model has been proved to be a very useful modeling approach in many applications, its linear assumption on η_0 is still very strict in many cases. Various parametric approaches have been proposed to allow more flexibility than this simple linear model (Wei, 1998). We will not review the general literature on parametric modeling of the count data. In contrast, we will focus on the nonparametric estimate of η_0 which allows η_0 to take on a more flexible form by only assuming that it is an element of some reproducing kernel Hilbert space $\mathcal{H}$ of smooth functions.

The smoothing spline model for Poisson regression was first introduced by O'Sullivan, Yandell and Raynor (1986). They used the penalized likelihood method to estimate η_0. Their smoothing spline estimator $\widehat{\eta}_\lambda(\cdot)$ of $\eta_0(\cdot)$ is defined as the minimizer in $\mathcal{H}$ of

$$-\sum_{i=1}^{n} l\left(y_i, \eta\left(x_i\right)\right) + \frac{n\lambda}{2} J\left(\eta\right). \tag{3.1.2}$$

where the smoothing parameter $\lambda \geq 0$ balances the tradeoff between minimizing the negative log likelihood function

$$L(y, \eta) \equiv -\sum_{i=1}^{n} l\left(y_i, \eta\left(x_i\right)\right) \equiv \sum_{i=1}^{n} \left[-y_i \eta\left(x_i\right) + e^{\eta(x_i)} - \ln\left(y_i!\right)\right], \tag{3.1.3}$$

and the "smoothness" $J(\eta)$. Here $J(\eta)$ is a quadratic penalty functional defined on $\mathcal{H}$.

Like the other smoothing spline models, the choice of the smoothing parameter λ considerably affects the performance of the estimate of η_0. If λ is very small, the estimator $\widehat{\eta}_\lambda$ would be very close to "interpolating" the observations. On the other hand if $J(\eta) = \int(\eta'')^2$ and $\lambda = \infty$, the estimator will boil down to an estimate of a generalized linear model. Thus, smoothing parameter selection is one of the most important practical issue for the nonparametric Poisson regression.

For regression with Gaussian type response, various criteria for selecting an optimal smoothing parameter have been proposed and extensively studied (see Wahba, 1990). Among which, cross-validation (CV, for short; Stone, 1974), generalized cross validation (GCV, for short; Craven and Wahba, 1979), Mallows C_p (Mallows, 1973) and their randomized versions (Girard, 1989,1991) are most commonly used. But for Poisson response, the situation is much more complicated because not only the mean but also the variance of the response are involved with the unknown function η_0. Another difficulty each smoothing parameter selection method should face is the computational problem. Even for a fixed smoothing parameter, we need iterations to get $\widehat{\eta}_\lambda$. Consequently, the tuning method must be computationally efficient.

Although many criteria for selecting smoothing parameters for Gaussian regression models have been extensively studied, the corresponding literature for generalized nonparametric regression is rather sparse. Two of the most often used techniques are a GCV based iterative procedure (Gu, 1990) and GACV

(Xiang and Wahba, 1996). The first proposal for tuning smoothing parameters in non-Gaussian regression models is due to Gu (1990). He suggested adapting the GCV in an iterative procedure of solving (3.1.2). A major criticism of this approach is that it is not guaranteed to converge although in most applications it does. For that reason, the study here does not incorporate Gu's idea.

Since its introduction by Xiang and Wahba (1996), GACV has become the most popular tool to optimally choose the smoothing parameter for generalized spline-based regression. Although GACV was derived under a general setting where the conditional distribution of response variables given the covariates falls in the exponential family, most of its later applications focus on the Bernoulli data. Despite its great success for Bernoulli data, we are unclear about whether these good properties can be extended to Poisson regression.

An alternative approach to tune the smoothing parameter for the nonparametric Poisson regression is based on the unbiased risk (UBR, for short) estimate of a penalized likelihood estimator. Unlike the logistic regression where a UBR does not exist according to Ye and Wong (1997), Poisson regression has an exact UBR. However, the problem is that naively using the UBR for tuning smoothing parameters is very computationally expensive.

The current chapter has two purposes. First we want to derive a computable approximation for the UBR so that it can be used to optimally choose smoothing parameters. We also want to answer the question how GACV can be applied to Poisson data, and how it compares to the AUBR for the Poisson regression.

In the next section, GACV for Poisson regression is briefly reviewed. An approximation of the unbiased risk estimate (AUBR) for Poisson regression is derived in Section 3.3. The computational problem for applying GACV and AUBR to the penalized likelihood smoothing spline model for Poisson regression is then formulized in Section 3.4. Several sets of simulations are conducted in Section 3.5 to evaluate both smoothing parameter tuning techniques. Finally, a real life example is given to illustrate the methods.

3.2 GACV

Denote $\widehat{\eta}$ as an estimator of η_0. The discrepancy between $\widehat{\eta}$ and the true parameter η_0 can be measured by the sample Kullback-Leibler (KL) distance

$$\begin{aligned} KL\left(\eta_0, \widehat{\eta}\right) &= \frac{1}{n}\sum_{i=1}^{n} E_z\left[z_i\left(\eta_0(x_i) - \widehat{\eta}(x_i)\right) - \left(e^{\eta_0(x_i)} - e^{\widehat{\eta}(x_i)}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[e^{\eta_0(x_i)}\left(\eta_0(x_i) - \widehat{\eta}(x_i)\right) - \left(e^{\eta_0(x_i)} - e^{\widehat{\eta}(x_i)}\right)\right], \quad (3.2.1) \end{aligned}$$

where z is an independent copy of $y = (y_1, \ldots, y_n)'$. Our goal is to adaptively choose an estimator such that the KL distance from $\widehat{\eta}$ to η_0 is minimized. Noting that term

$$\frac{1}{n}\sum_{i=1}^{n}\left[\eta_0(x_i)e^{\eta_0(x_i)} - e^{\eta_0(x_i)}\right]$$

is independent of the estimator $\widehat{\eta}$, we can simply consider the remaining terms of (3.2.1), which are the so-called comparative Kullback-Leibler (CKL) distance

$$CKL\left(\eta_0, \widehat{\eta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[-e^{\eta_0(x_i)}\widehat{\eta}(x_i) + e^{\widehat{\eta}(x_i)}\right]. \quad (3.2.2)$$

The involvement of the unknown parameter η_0 in CKL makes directly minimizing (2.2) infeasible. A commonly used technique to get around this problem is cross validation.

Define the leaving-out-one cross validation function $CV(\lambda)$ as

$$CV(\eta_0, \widehat{\eta}_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\widehat{\eta}_\lambda^{(-i)}(x_i) + e^{\widehat{\eta}_\lambda(x_i)}\right], \tag{3.2.3}$$

where $\widehat{\eta}_\lambda^{(-i)}$ is the minimizer of (3.1.2) with the ith data point omitted. If η_0 is "smooth" and the data are dense, the leaving-out-one cross validation can be expected to be at least roughly unbiased for the CKL loss. Let

$$y^{(-i)} = \left(y_1, \ldots, y_{i-1}, e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}, y_{i+1}, \ldots, y_n\right). \tag{3.2.4}$$

The leaving-out-one lemma (Xiang and Wahba, 1996) tells us that the minimizer of (3.1.2) with response vector $y^{(-i)}$ is also $\widehat{\eta}_\lambda^{(-i)}$. A linearization argument yields

$$\widehat{\eta}_\lambda(x_i) - \widehat{\eta}_\lambda^{(-i)}(x_i) \approx \frac{\partial\widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}). \tag{3.2.5}$$

Thus,

$$\begin{aligned}
CV(\eta_0, \widehat{\eta}_\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\widehat{\eta}_\lambda^{(-i)}(x_i) + e^{\widehat{\eta}_\lambda(x_i)}\right] \\
&= L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\left(\widehat{\eta}_\lambda(x_i) - \widehat{\eta}_\lambda^{(-i)}(x_i)\right) \\
&\approx L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial\widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}) \\
&= L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial\widehat{\eta}_\lambda(x_i)}{\partial y_i}\frac{(y_i - e^{\widehat{\eta}_\lambda(x_i)})}{1 - \frac{e^{\widehat{\eta}_\lambda(x_i)} - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}}{y_i - e^{\widehat{\eta}_\lambda^{(-i)}(x_i)}}} \\
&\approx L(y, \widehat{\eta}_\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i\frac{\partial\widehat{\eta}_\lambda(x_i)}{\partial y_i}(y_i - e^{\widehat{\eta}_\lambda(x_i)})\frac{1}{1 - \frac{\partial\widehat{\eta}_\lambda(x_i)}{\partial y_i}e^{\widehat{\eta}_\lambda(x_i)}}
\end{aligned} \tag{3.2.6}$$

The last quantity of the above equation is called the approximate cross validation (ACV). Replacing $\partial\widehat{\eta}_\lambda(x_i)/\partial y_i$ by $tr(A(\widehat{\eta}_\lambda))$ and $e^{\widehat{\eta}_\lambda(x_i)}\partial\widehat{\eta}_\lambda(x_i)/\partial y_i$ by $tr(V(\widehat{\eta}_\lambda)^{1/2}A(\widehat{\eta}_\lambda)V(\widehat{\eta}_\lambda)^{1/2})$, we get the GACV for the Poisson regression.

$$GACV(\eta_0,\widehat{\eta}_\lambda) = L(y,\widehat{\eta}_\lambda) + \frac{tr(A(\widehat{\eta}_\lambda))}{n}\frac{\sum_{i=1}^n y_i(y_i - \exp(\widehat{\eta}_\lambda(x_i)))}{n - tr(V(\widehat{\eta}_\lambda)^{1/2}A(\widehat{\eta}_\lambda)V(\widehat{\eta}_\lambda)^{1/2})}, \quad (3.2.7)$$

where $A(\widehat{\eta}_\lambda) = \partial\widehat{\eta}_\lambda/\partial y$ and $V(\widehat{\eta}_\lambda)$ is a diagonal matrix with the $(i,i)th$ entry $e^{\widehat{\eta}_\lambda(x_i)}$. We usually call A the influence matrix or hat matrix for (3.1.2).

3.3 AUBR

Another way to avoid minimizing (3.2.2) is to minimize an unbiased estimator of CKL instead. Note that for any function f, the following identity holds (Hudson, 1978),

$$\begin{aligned} E_{y_i}\left[y_i f(y_i-1)\right] &= \sum_{k=0}^{\infty} e^{-\mu_i}\frac{\mu_i^k}{k!}kf(k-1) \\ &= \sum_{k=1}^{\infty} e^{-\mu_i}\frac{\mu_i^{k-1}}{(k-1)!}\mu_i f(k-1) \\ &= \mu_i\sum_{k=0}^{\infty} e^{-\mu_i}\frac{\mu_i^k}{k!}f(k) \\ &= \mu_i E_{y_i}\left[f(y_i)\right], \end{aligned} \quad (3.3.1)$$

where $\mu_i = E(y_i) = \exp(\eta_0(x_i))$. Consequently, replacing f in (3.3.1) by $\widehat{\eta}$, an unbiased estimator for (3.2.2) is

$$UBR\left(\eta_0,\widehat{\eta}\right) = \frac{1}{n}\sum_{i=1}^n\left[-y_i\widehat{\eta}^i(x_i) + e^{\widehat{\eta}(x_i)}\right] = L(y,\widehat{\eta}) + \sum_{i=0}^n y_i\left(\widehat{\eta}(x_i) - \widehat{\eta}^i(x_i)\right), \quad (3.3.2)$$

where $\widehat{\eta}^i$ is estimated in the same way as $\widehat{\eta}$ with response vector

$$(y_1, \ldots, y_{i-1}, y_i - 1, y_{i+1}, \ldots, y_n)'.$$

This unbiased risk estimator was first derived by Ye and Wong (1997) in a technical report.

Although the unbiased risk estimator (3.3.2) is very elegant, it is not directly applicable because it is usually very expensive to compute. For example, consider the penalized estimator. For each smoothing parameter λ, we must calculate $n+1$ penalized likelihood estimators $\widehat{\eta}_\lambda, \widehat{\eta}_\lambda^1, \ldots, \widehat{\eta}_\lambda^n$ so that $UBR\,(\eta_0, \widehat{\eta}_\lambda)$ could be obtained. This is impracticable even for medium sample sizes. Here, we shall try to find a good approximation of this UBR which will considerably reduce the computation.

Regard $\widehat{\eta}(x_i)$ as a continuously differentiable function of y_i. Using the mean value theorem, we have

$$\widehat{\eta}(x_i) - \widehat{\eta}^i(x_i) = \frac{\partial \xi(x_i)}{\partial y_i}$$

where ξ is the same estimator as $\widehat{\eta}$ with response vector $y - p1_i$ for some $0 \leq p \leq 1$ and 1_i is a vector with the ith entry being 1 and all the other elements being 0. Approximating $\partial\xi(x_i)/\partial y_i$ by $\partial\widehat{\eta}(x_i)/\partial y_i$, we get the following approximation of the UBR

$$AUBR(\eta_0, \widehat{\eta}) \equiv L\,(y, \widehat{\eta}) + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{\partial \widehat{\eta}(x_i)}{\partial y_i}. \tag{3.3.3}$$

Unlike GACV, the derivation of both UBR and AUBR is not limited to penalized likelihood spline estimator. In the next section, we shall restrict our

attention to the penalized likelihood smoothing spline model. More details of the computational procedure will be formulized for this model and for different smoothing parameter tuning methods.

3.4 Smoothing Parameter Selection

For convenience, we shall use vector notation hereafter. For example, $y = (y_1, \ldots, y_n)'$, and $\eta_0 = (\eta_0(x_1), \ldots, \eta_0(x_n))'$. η, $\widehat{\eta}$, $\widehat{\eta}^i$ and other vectors are defined in the same manner.

The problem of adaptive choice of smoothing parameter λ can be described as finding the best approximation of η_0 within the class of penalized likelihood estimators $\mathcal{M} \equiv \{\widehat{\eta}_\lambda : \lambda \in \Lambda\}$. The difficulty is that we do not know η_0. Instead, we only observe random variables related to η_0. Either GACV or AUBR can serve as a proxy of the CKL distance from $\mathcal{M}$ to η_0.

Given λ, the computational problem is then to find η to minimize

$$I_\lambda(y, \eta) = -\frac{1}{n}\sum_{i=1}^{n} l(y_i, \eta_i) + \frac{\lambda}{2}\eta'\Sigma\eta, \tag{3.4.1}$$

Let $\widehat{\eta}_\lambda^{i,\delta}$ minimize $I_\lambda(y - \delta 1_i, \eta)$. Note that $\widehat{\eta}_\lambda$ minimizes $I_\lambda(y, \eta)$. Therefore,

$$\frac{\partial I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta} = \frac{1}{n}(-y + \exp(\widehat{\eta}_\lambda)) + \lambda\Sigma\widehat{\eta}_\lambda = 0 \tag{3.4.2}$$

$$\frac{\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^{i,\delta})}{\partial \eta} = \frac{1}{n}\left(-y + \delta 1_i + \exp(\widehat{\eta}_\lambda^{i,\delta})\right) + \lambda\Sigma\widehat{\eta}_\lambda^{i,\delta} = 0 \tag{3.4.3}$$

Using a first order Taylor expansion to $\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^i)/\partial\eta$ at $(y, \widehat{\eta}_\lambda^i)$, we have

the following equation:

$$0 = \frac{\partial I_\lambda(y - \delta 1_i, \widehat{\eta}_\lambda^{i,\delta})}{\partial \eta} = \frac{\partial I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta} + \frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta' \partial \eta} \left(\widehat{\eta}_\lambda^{i,\delta} - \widehat{\eta}_\lambda \right) - \delta \frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial y' \partial \eta} 1_i + O\left(\left|\left| \widehat{\eta}_\lambda^{i,\delta} - \widehat{\eta}_\lambda \right|\right|^2 + \delta^2 \right) \quad (3.4.4)$$

If $\delta \to 0$, we get

$$\begin{aligned} \frac{\partial \widehat{\eta}_\lambda}{\partial y_i} &= -\left(\frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial \eta' \partial \eta} \right)^{-1} \frac{\partial^2 I_\lambda(y, \widehat{\eta}_\lambda)}{\partial y' \partial \eta} 1_i \\ &= (V(\widehat{\eta}_\lambda) + n\lambda\Sigma)^{-1} 1_i. \end{aligned} \quad (3.4.5)$$

As defined before, $V(\widehat{\eta}_\lambda)$ is a diagonal matrix with the $(i,i)th$ element $\exp(\widehat{\eta}_\lambda(x_i))$. Thus, to compute GACV or AUBR, it suffices to evaluate Σ.

The representer theorem (Kimeldorf and Wahba, 1971) tells us that the exact minimizer of (3.1.2) has a finite representation when $J(\eta)$ is a semi norm in a reproducing kernel Hilbert space $\mathcal{H}$. A popular example is $J(\eta) = \int (\eta'')^2$. If $\mathcal{H}$ is decomposed into $\mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is the null space of J, then the minimizer of (3.1.2) in $\mathcal{H}$ has the following form

$$\widehat{\eta}_\lambda(x) = \sum_{v=1}^{m} d_v \phi_v(x) + \sum_{i=1}^{n} c_i K(x, x_i), \quad (3.4.6)$$

where $\{\phi_v\}$ is the basis of $\mathcal{H}_0$, and it is being assumed that an $n \times m$ matrix S with (i,v) entry $\phi_v(x_i)$ is of full column rank. $c = (c_1, \ldots, c_n)'$ satisfies $S'c = 0$, and K is the reproducing kernel for $\mathcal{H}_1$. For example, if we take $\mathcal{H}$ as the second order Sobolev space and $J(\eta) = \int (\eta'')^2$, then

$$K(u,v) = k_2(u)k_2(v) - k_4([u-v]),$$

where $n!k_n(u)$ is the nth Bernoulli polynomial and $[\tau]$ is the fractional part of τ. Furthermore, $J(\widehat{\eta}_\lambda) = c'Qc$ where Q is an $n \times n$ matrix with the $(i,j)th$ entry $K(x_i, x_j)$. Thus to minimize (1.2), it suffices to find c and $d = (d_1, \ldots, d_m)'$ that minimizes

$$-L(y, \sum_{v=1}^{m} d_v \phi_v(x_i) + \sum_{j=1}^{n} c_i K(x_i, x_j)) + \frac{n\lambda}{2} c'Qc. \tag{3.4.7}$$

In order to apply (4.4) to compute GACV and AUBR, we need to find Σ such that $\widehat{\eta}_\lambda' \Sigma \widehat{\eta}_\lambda = c'Qc$. It has been shown by Xiang and Wahba (1996) that such Σ has the form

$$\Sigma = \Delta \left(\Delta Q \Delta'\right)^{+} \Delta', \tag{3.4.8}$$

where Δ is any $n \times (n-m)$ matrix of orthonormal vectors whose columns are perpendicular to the columns of S, and $+$ represents the Moore-Penrose generalized inverse. In the case where Q is of full rank, we can also write

$$\Sigma = Q^{-1} - Q^{-1} S (S' Q^{-1} S)^{-1} S' Q^{-1}. \tag{3.4.9}$$

(3.4.5), (3.4.8) and (3.4.9) offer us a direct way to compute GACV and AUBR. However, when the sample size gets larger, the computation of Σ by (3.4.7) or (3.4.8) may not be stable. As an alternative, we suggest using the following procedure to approximate $\partial \widehat{\eta}_\lambda / \partial y$.

With a non quadratic penalized negative log likelihood (3.1.2), iterations are needed to calculate the penalized likelihood fit for a fixed smoothing parameter. $\theta = (\theta_1, ..., \theta_N)'$ to minimize

$$I_\lambda = -\sum_{i=1}^{n} l\left(y_i, \eta_i\left(\theta\right)\right) + \frac{n\lambda}{2} \theta' \Sigma_\theta \theta, \tag{3.4.10}$$

where

$$\eta_i(\theta) = \sum_{j=1}^{N} \theta_j B_j(x_i)$$

and Σ_θ is defined by

$$\theta' \Sigma_\theta \theta = J\left(\sum_{j=1}^{n} \theta_j B_j\right).$$

Since I_λ is a convex function of θ, we may compute θ via a Newton-Ralphson iteration. Define $w_i = \exp(\eta(x_i)), u_i = y_i - w_i$. $\theta^{[k]}$ (the kth estimate of θ) can be updated by:

$$\theta^{[k+1]} = \min_\theta \left[\frac{1}{n}\sum_{i=1}^{n} w_i^{[k]}\left(y_i^{[k]} - \eta_i(\theta)\right)^2 + \frac{\lambda}{2}\theta'\Sigma_\theta\theta\right], \tag{3.4.11}$$

where $y_i^{[k]} = \eta_i(\theta^{[k]}) - u_i^{[k]}/w_i^{[k]}$ and $u_i^{[k]}, w_i^{[k]}$ are the values of u, w evaluated at $\theta^{[k]}$. The influence matrix for (3.4.10) is defined as an $n \times n$ matrix $A^{[k+1]}$ such that $\eta\left(\theta^{[k+1]}\right) = A^{[k+1]} y^{[k]}$. Denote $A^{[\infty]}$ the influence matrix after the convergence of this algorithm. $A^{[\infty]}$ is a natural approximation of $\partial\theta^{[\infty]}/\partial y^{[\infty]}$. The chain rule gives us

$$A(\widehat{\eta}_\lambda) = \frac{\partial \widehat{\eta}_\lambda}{\partial y} = \frac{\partial \widehat{\eta}_\lambda}{\partial y^{[\infty]}} \frac{\partial y^{[\infty]}}{\partial y} \approx A^{[\infty]} V(\widehat{\eta}_\lambda)^{-1}, \tag{3.4.12}$$

Then, A can be evaluated via an efficient and stable algorithm for computing $A^{[k]}$ (Wahba, 1990).

3.5 Simulations

Several sets of simulations will be presented in this section to evaluate the performance of GACV and AUBR as smoothing parameter tuning techniques

for nonparametric Poisson regression.

3.5.1 GACV and AUBR as Approximations to CKL

As argued before, AUBR approximates an unbiased estimator of CKL. The main motivation for GACV is also to approximate the CKL. Asymptotically, if η_0 is "smooth", one may expect the GACV curve and AUBR curve ($GACV(\eta_0, \widehat{\eta}_\lambda)$ or $AUBR(\eta_0, \widehat{\eta}_\lambda)$ vs λ) to be close to the CKL curve. The first set of simulations is conducted to evaluate how close the GACV and AUBR computed via either (3.4.9) or (3.4.12) are to the CKL curve. For this purpose, we consider the following test function

$$\eta_0(x_i) = 2\sin(2\pi x_i), \qquad i = 1, \ldots, 100,$$

where $x_i = (i - 0.5)/100$. $y_i, i = 1, \ldots, 100$ are independently sampled from a Poisson distribution with intensity $\exp(\eta_0(x_i))$. The above experiment has been repeated for 50 times. For each simulated dataset, the CKL curve, the GACV and AUBR curves computed via both (3.4.9) and (3.4.12) have been recorded.

The top left panel of figure 3.1 gives the average curves from 50 replications of the above experiment. AUBR, calculated via either (3.4.9) or (3.4.12), provides an unbiased estimator of the CKL from this figure. Although the average GACV curves are not so close to CKL as the AUBR curves, they do capture the main shape of CKL curves in average. To get a better understanding of the individual behavior of each criteria, the remaining five panels in Figure 3.1 give the CKL curve, the GACV and AUBR curve computed by either (3.4.9) or (4.12) for

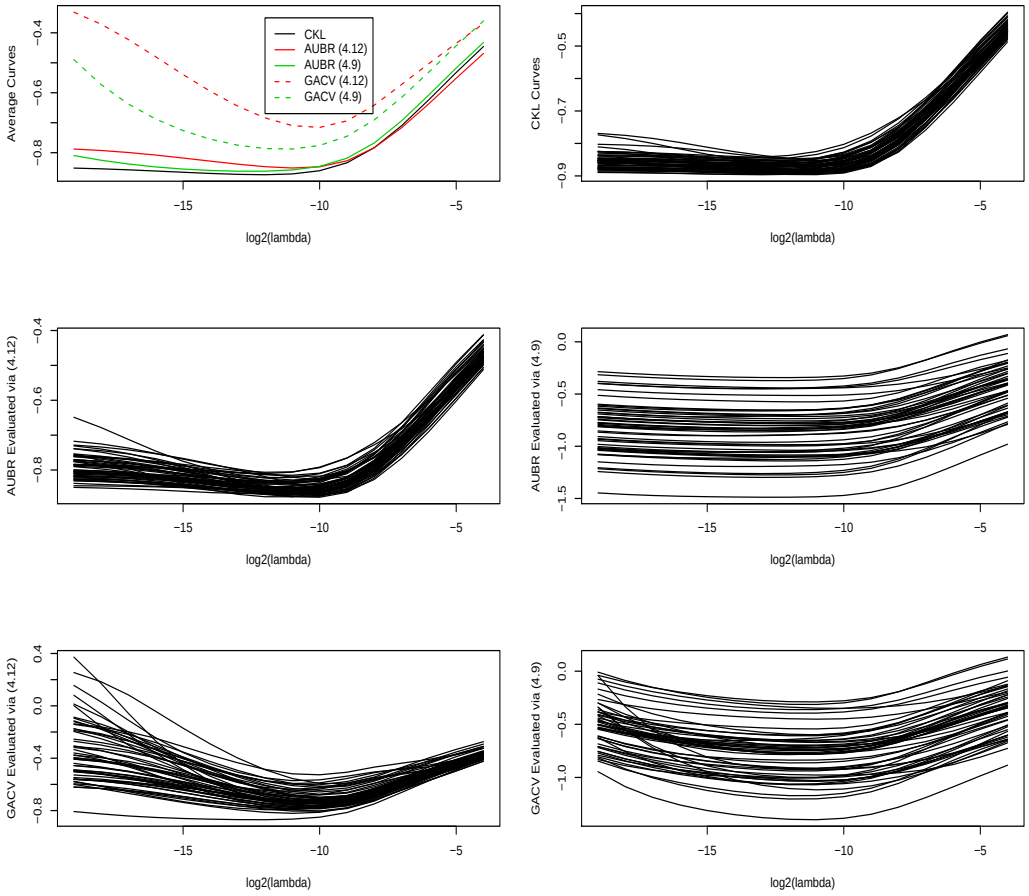

Figure 3.1: AUBR curve, GACV curve and CKL curve for Poisson regression

each of the 50 replications. An interesting feature is that GACV and AUBR from (3.4.9) are much more variable than those from (3.4.12). For this reason, we shall use (3.4.12) to evaluate GACV and AUBR in the rest of the chapter. Another interesting observation from Figure 3.1 is that GACV penalizes small smoothing parameters too much.

To investigate the possible dependence on the magnitude of η_0 of the performance of AUBR and GACV, we also conduct the following simulation. Consider the following 5 test functions

$$\begin{aligned}
\eta_0(x_i) &= 2\sin(2\pi x_i) + 3; \\
\eta_0(x_i) &= 2\sin(2\pi x_i) + 2; \\
\eta_0(x_i) &= 2\sin(2\pi x_i) + 1; \\
\eta_0(x_i) &= 2\sin(2\pi x_i); \\
\eta_0(x_i) &= 2\sin(2\pi x_i) - 1.
\end{aligned}$$

Again, the sample size is 100 and covariate x is defined as before. For each test function, the previous experiment is repeated 50 times. The average CKL curve, AUBR curve and GACV curve are depicted in Figure 3.2.

In each panel, black solid line corresponds to average CKL curve, red solid line stands for average AUBR curve and red dashed line for average GACV curve. One can see that both AUBR and GACV are very close to the average CKL for the first three test functions. For the last two test functions, however, biases appear for both criteria. AUBR tends to give too small penalty for small

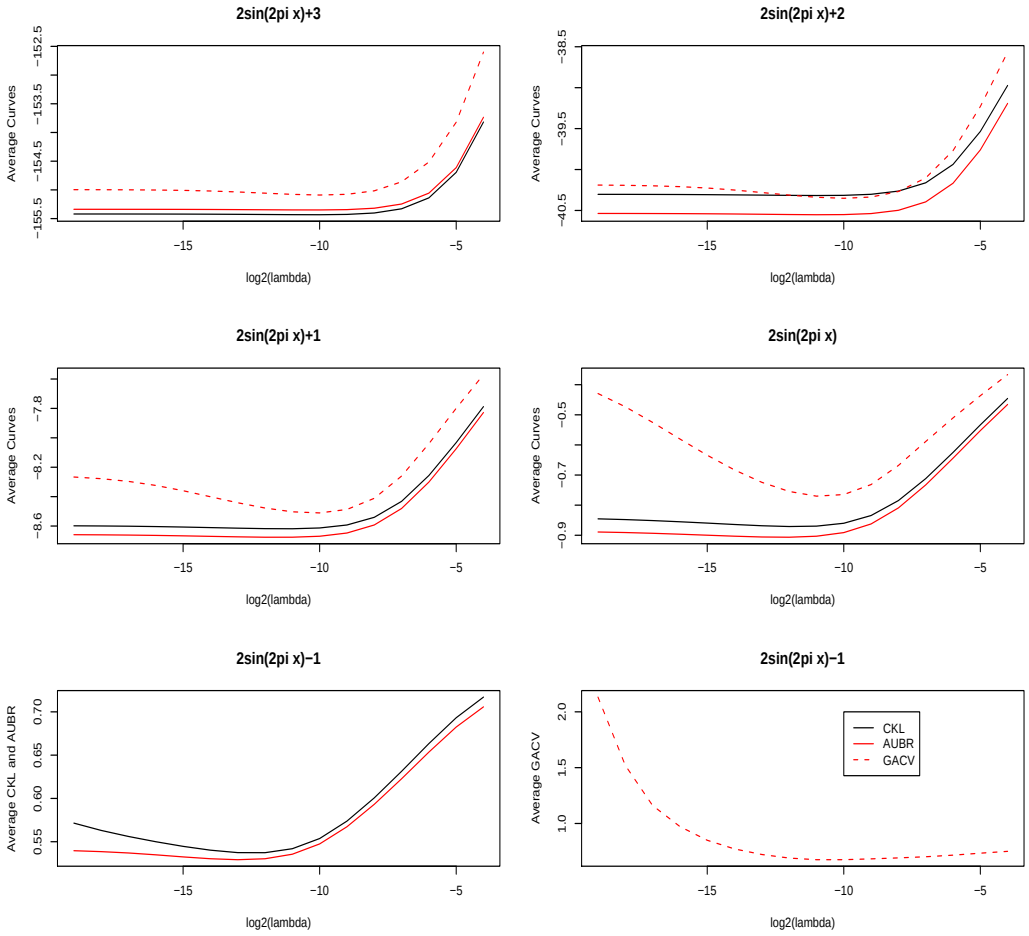

Figure 3.2: Both AUBR and GACV approximate CKL

smoothing parameters. GACV, on the other hand, gives too much penalty for small smoothing parameters. This observation might suggest a tendency of GACV to oversmooth when the intensity function is small.

3.5.2 Tuning Smoothing Parameters with GACV and AUBR

The performance of GACV and AUBR based smoothing parameter tuning methods depends on how well they approximate the CKL around the optimal smoothing parameter. Fortunately, although GACV and AUBR are biased estimators of CKL when the intensity function is very small, they are still reasonably unbiased around the minima of the CKL curve in general. To compare the smoothing parameter tuning method based on GACV and AUBR, we conducted two sets of simulations.

The first set consists of 4 test functions of different magnitude. To get a better impression of the magnitude of the intensity functions, we present them in terms of the intensity function $\mu = exp(\eta_0)$.

$$\begin{aligned}
\mu_1(x) &= \exp\left(\frac{sin(2\pi x)}{2 - sin(2\pi x)}\right) \\
\mu_2(x) &= 7 + 7x^5 + 7(x-1)^5 \\
\mu_3(x) &= 10000\left(x^8(1-x)^2 + x^2(1-x)^8\right) \\
\mu_4(x) &= 5\exp\left(-100(x-0.75)^2\right) + 480(x-0.75)^2
\end{aligned}$$

The last three test functions were used by Climov, Hart and Simar (2002) and they have different signal to noise ratios, which are defined as

$$SNR = \frac{\max_{0\le x\le 1}\mu(x) - \min_{0\le x\le 1}\mu(x)}{\sqrt{E\mu(X)}},$$

where the expectation is taken over covariate X, in our example, it is uniformly distributed over $[0,1]$. The values of SNR are 9,30,5 for μ_2, μ_3, μ_4, respectively. All 4 test functions are displayed in Figure 3.3.

For each test function, the following procedure was repeated for one hundred times:

(1) One hundred $x's$ are drawn from $U[0,1]$.

(2) For each covariate $x_i, i = 1, \ldots, 100$, y_i is sampled from a Poisson distribution with intensity $\mu_k(x_i)$ for $k = 1, \ldots, 4$.

(3) Estimate η by $\widehat{\eta}_\lambda$ with λ chosen to minimize the GACV score, compute the true KL distance from the estimator of the true function.

(4) Estimate η by $\widehat{\eta}_\lambda$ with λ chosen to minimize the AUBR score, compute the true KL distance from the estimator of the true function.

Figure 3.4 presents the scatter plots of $KL\left(\eta_0, \widehat{\eta}_{\lambda_{GACV}}\right)$ versus $KL\left(\eta_0, \widehat{\eta}_{\lambda_{AUBR}}\right)$.

For test functions μ_2 and μ_4, AUBR performs better than GACV. For test functions μ_1 and μ_3, AUBR and GACV perform similarly. But AUBR still is slightly better than GACV in terms of outperforming GACV more often.

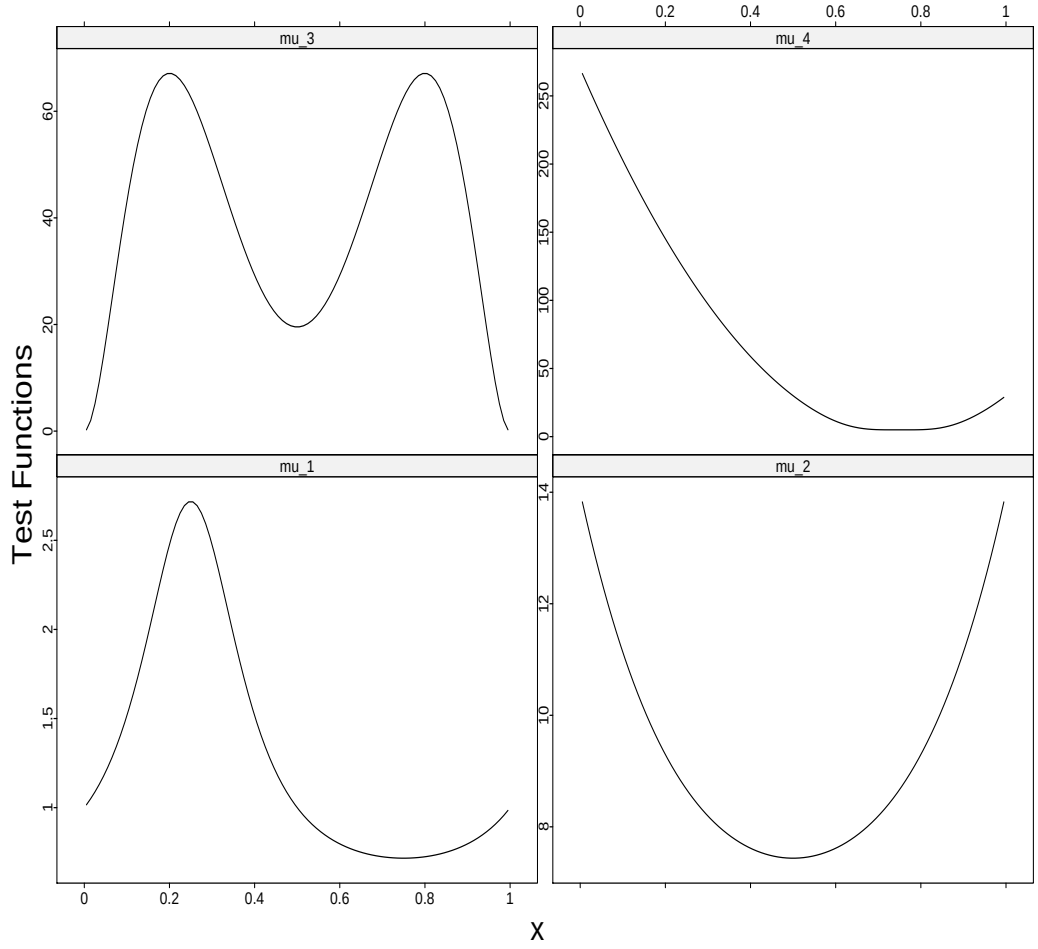

Figure 3.3: Test functions

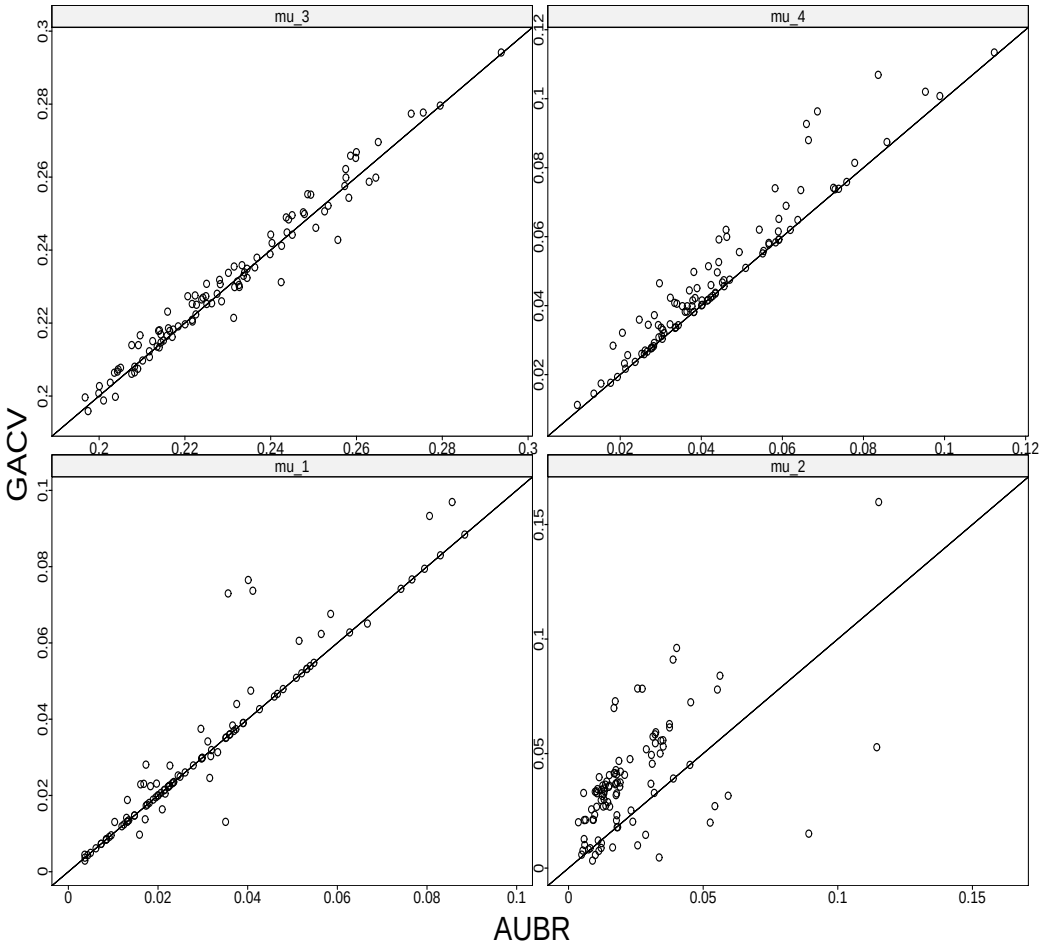

Figure 3.4: KL comparison using different tuning methods

In another set of simulations, we want to extend our comparison to two-dimensional test functions. Two test functions were used:

$$\begin{aligned}\mu_5(x_1, x_2) &= \exp(2\sin(2\pi x_1) - \sin(2\pi x_2)) \\ \mu_6(x_1, x_2) &= \exp(8(\exp(-\frac{1}{3.38\left((x_1-2)^2+x_2^2\right)}) + \exp(-\frac{1}{3.38\left((x_1+2)^2+x_2^2\right)})) - 1) - 46\end{aligned}$$

200 triples of (x_1, x_2, y) are sampled according to the following laws: (x_1, x_2) is independently sampled from $U[0,1]^2$, then y is sampled from a Poisson distribution with intensity $\mu(x_1, x_2)$. After getting these 200 observations, we estimate the log intensity function by a penalized likelihood estimator with thin plate splines where

$$J(\eta) = \int_0^1 \int_0^1 \left[\left(\frac{\partial^2 \eta}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 \eta}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 \eta}{\partial x_2^2}\right)^2 \right] dx_1 dx_2$$

and smoothing parameter λ chosen to minimize either AUBR or GACV. We repeat this experiment one hundred times.

The range of μ_5 is from 0.05 to 20, while the range of μ_6 is from 0.6 to 8.7. To investigate the possible effects of the magnitude of the intensity functions, we also consider two other test functions with intensities

$$\mu_7(x_1, x_2) = 7.4 \times \mu_5(x_1, x_2), \qquad \mu_8(x_1, x_2) = 7.4 \times \mu_6(x_1, x_2).$$

We refer to μ_5 and μ_6 as test functions with small intensities and to μ_7 and μ_8 as test functions with large intensities. The top left panel of Figure 3.5 depicts the common shape of μ_5 and μ_7.

The top right panel gives the common shape of μ_6 and μ_8. The KL distance based comparisons between GACV and AUBR are given in the remaining four

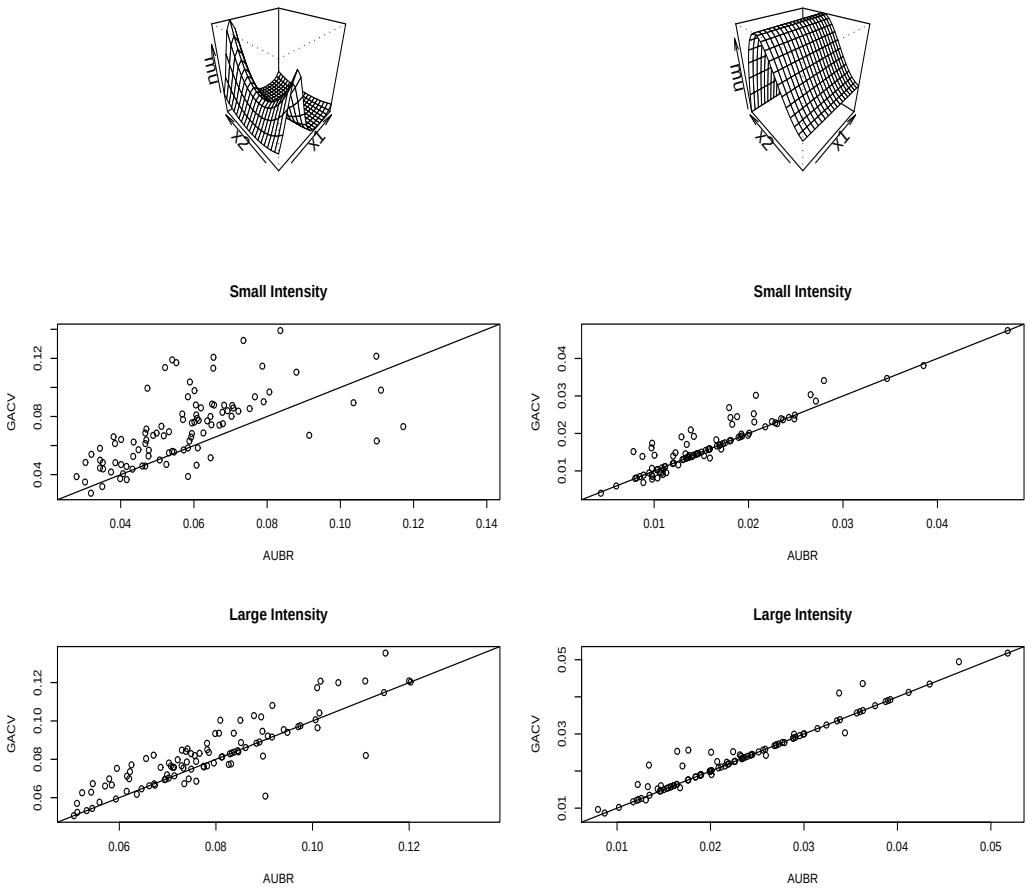

Figure 3.5: Two-dimensional KL comparison using different tuning methods

panels. When the intensity functions are relatively small, AUBR outperforms GACV. When the intensity gets larger, this difference diminishes.

3.5.3 Summary of Simulation Results

From the above simulations, we find that both GACV and AUBR are good approximations of the CKL loss and both can be applied to choose smoothing parameters for nonparametric Poisson regression if the intensity function is large. However, when the intensity functions are small, GACV has a tendency to overestimate the CKL loss for small smoothing parameters. As a consequence, we conjectured that oversmoothing might occur for small intensity functions if we choose smoothing parameters by GACV. This is confirmed by the observation that AUBR outperforms GACV when intensity functions are small.

3.6 Application

In this section, we apply the smoothing parameter tuning methods based on GACV and AUBR to a real dataset. The dataset represents a time series of medical count data. The surveyors recorded the number of (new) polio infections per month during the time from January 1st, 1970 till December 12th, 1987. The data were recorded by the US Ministry for Health. The dataset contains the number of infections per month, and thus a total of $18 \times 12 = 216$ count data. We are interested in how the incidence rate evolves over time.

For this purpose, we consider a Poisson regression where the canonical parameter η_0 is a function of time. Figure 3.6 gives the AUBR curve and GACV curve for the Poisson regression. Also provided are the observed counts and the fitted intensity functions using the smoothing parameter chosen by minimizing the AUBR or GACV. From the figure, we observe a suspicious oversmoothing of GACV based tuning technique. This further confirms what we discovered in the simulations.

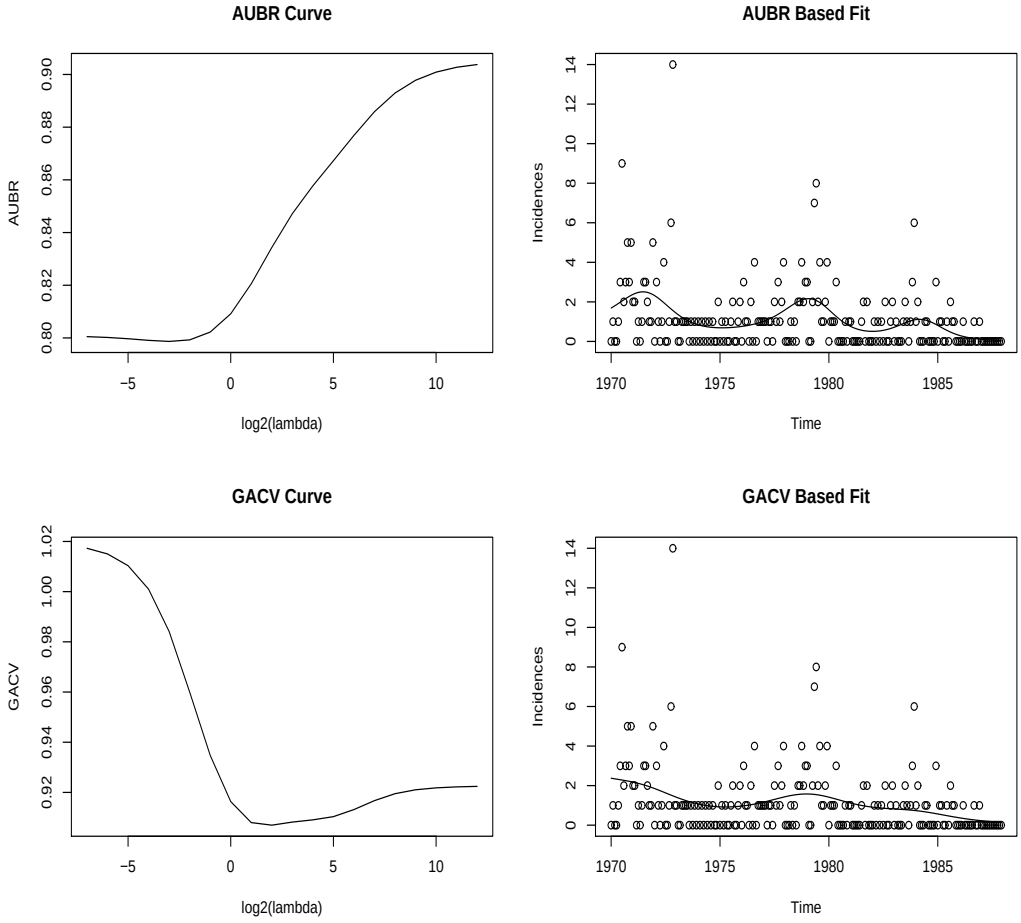

Figure 3.6: Polio infection data

Chapter 4

Efficient Empirical Bayes Variable Selection and Estimation in Linear Models

4.1 Introduction

We consider the problem of variable selection and coefficient estimation in the common normal linear regression model where we have n observations on a dependent variable Y and p predictors $(x_1, x_2, ..., x_p)$, and

$$Y = X\beta + \epsilon, \tag{4.1.1}$$

where $\epsilon \sim N_n(0, \sigma^2 I)$, and $\beta = (\beta_1, ..., \beta_p)'$. Throughout this chapter, we center each input variable so that the observed mean is zero, and scale each predictor so that the sample standard deviation is one.

The underlying notion behind variable selection is that some of the predictors are redundant and therefore only an unknown subset of the β coefficients are nonzero. By effectively identifying the subset of important predictors, variable

selection can improve estimation accuracy and enhance model interpretability. Classical variable selection methods, such as C_p, AIC, and BIC, choose among possible models using penalized sum of squares criteria, with the penalty being a constant multiple of the model dimension. George and Foster (2000) showed that these criteria correspond to a hierarchical Bayes model selection procedure under a particular class of priors. This gives a new perspective of various earlier model selection methods and put them in a unified framework. The hierarchical Bayes formulation put a prior on the model space, and then put a prior on the coefficients given the model. This is conceptually attractive. George and Foster (2000) proposed to estimate the hyperparameters of the hierarchical Bayes formulation with a marginal maximum likelihood criterion or a conditional maximum likelihood criterion. The resulting empirical Bayes criterion uses an adaptive dimensionality penalty and compares favorably with the penalized least squares criteria with fixed dimensionality penalty. However, even after the hyperparameters are estimated, the resulting model selection criterion has to be evaluated on each candidate model to select the best model. This is impractical for even moderate number of predictors since the number of candidate models grows exponentially as the number of predictors increases. In practice, this type of methods are implemented in a stepwise fashion, through forward selection or backward elimination. In doing so, one content oneself with the locally optimal solution instead of the globally optimal solution.

A number of other variable selection methods have been introduced in recent year (George and McCulloch, 1993; Foster and George, 1994; Breiman, 1995; Tibshirani, 1996; Fan and Li, 2001; Shen and Ye, 2002; and Efron, Johnston, Hastie and Tibshirani, 2004). In particular, Efron et al. (2004) proposed an effective variable selection algorithm LARS (least angle regression) that is extremely fast, and showed that with slight modification the LARS algorithm can be used to efficiently compute the popular LASSO estimate for variable selection, which is defined as:

$$\widehat{\beta}^{LASSO}(\lambda) = \arg\min_{\beta} \left(||Y - X\beta||^2 + \lambda \sum |\beta_i| \right), \tag{4.1.2}$$

where $\lambda > 0$ is a regularization parameter. By using the L_1 penalty, minimizing (4.1.2) yields a sparse estimate of β if λ is chosen appropriately. Consequently, a submodel of (4.1.1) which contains only the covariates corresponding to the nonzero components in $\widehat{\beta}^{LASSO}(\lambda)$ is selected as the final model. The LARS algorithm computes the whole path of the LASSO with a computational load in the same magnitude of the ordinary least square. Therefore the computation is extremely fast, and this facilitates the choice of the tuning parameter with criteria such as C_p or GCV.

In this chapter we adopt a hierarchical Bayes framework similar to that of George and McCulloch (1993) and George and Foster (2000), but with new prior specifications. We show that the resulting empirical Bayes estimator is closely related to the LASSO estimator and is quickly computable. We propose a conditional maximum likelihood criterion for the choice of the hyperparameter,

which in turn leads to an alternative method for choosing the tuning parameter in the LASSO. Unlike earlier methods including that in George and Foster (2000) and the LASSO tuned with C_p, where the error variance σ^2 is assumed known or fixed at the estimate from the saturated model, in our method it is estimated together with the other parameters using our conditional maximum likelihood criterion. Therefore our method can potentially be used in situations when the dimension is larger than the sample size.

The rest of the chapter is structured as follows. In Section 4.2 we introduce a hierarchical Bayes formulation for variable selection. In Section 4.3 we present an analytic approximation to the posterior probabilities in the Bayesian formulation which turns out to be connected to the LASSO estimate. This connection is further justified theoretically under the condition of orthogonal design in Section 4.4. In Section 4.5, we propose an empirical Bayes method to choose the hyperparameters via a conditional maximum likelihood criterion. In Section 4.6, we conduct a simulation study to compare our method with some related model selection methods. Section 4.7 presents a method to estimate the posterior probabilities of the models and the variables. This makes it possible to assess the model and variable uncertainty. We illustrate the performance of our method on a real dataset in Section 4.8. A summary is given in Section 4.9.

4.2 Hierarchical Model Formulation

A hierarchical model formulation for variable selection in linear models consists of the following three main ingredients:

(i) a prior probability $P(\mathcal{M})$ for each candidate model $\mathcal{M}$;

(ii) a prior $P(\theta_{\mathcal{M}}|\mathcal{M})$ for parameter $\theta_{\mathcal{M}}$ associated with model $\mathcal{M}$;

(iii) a data generating mechanism conditional on $(\mathcal{M}, \theta_{\mathcal{M}})$, $P(Y|\mathcal{M}, \theta_{\mathcal{M}})$.

Once these three components are specified, one can combine data and priors to form posterior

$$P(\mathcal{M}|Y) = \frac{P(\mathcal{M}) \int P(Y|\mathcal{M}, \theta_{\mathcal{M}}) P(\theta_{\mathcal{M}}|\mathcal{M}) d\theta_{\mathcal{M}}}{\sum_{\mathcal{M}'} \int P(Y|\mathcal{M}', \theta_{\mathcal{M}'}) P(\theta_{\mathcal{M}'}|\mathcal{M}') d\theta_{\mathcal{M}'} P(\mathcal{M}')}. \tag{4.2.1}$$

We begin by indexing each candidate model with one binary vector $\gamma = (\gamma_1, \ldots, \gamma_p)'$. An element γ_i takes value 0 or 1 depending on whether the i-th predictor is excluded from the model or not. Adopting this notation, under model γ, (4.1.1) can be written as

$$Y|\gamma, \beta \sim N\left(X_\gamma \beta_\gamma, \sigma^2 I_{(n)}\right), \tag{4.2.2}$$

where subscript γ indicates that only those columns or elements with the corresponding γ element being 1 are included. Notice that β_γ is of dimension $|\gamma|$, where $|\gamma|$ denotes $\sum \gamma_i$.

Now we specify the priors for β and γ. By the definition of γ, it is natural to force $\beta_i = 0$ if $\gamma_i = 0$. On the other hand, if $\gamma_i = 1$, we give a double exponential

prior for β_i. That is,

$$\beta_i|\gamma_i = (1-\gamma_i)\delta(0) + \gamma_i DE(0,\tau), \quad j = 1, ..., p, \tag{4.2.3}$$

where $DE(0,\tau)$ has density function $\tau \exp(-\tau|x|)/2$. The prior is chosen so that we can take advantage of the efficient LARS algorithm later on, but it is an attractive choice of prior on its own. In contrast to the commonly used normal prior $\beta_i|\gamma_i = 1 \sim N(0,\tau^2)$, the double exponential prior can better accommodate large regression coefficients due to its heavier tail probability. In a wavelet setup, Johnston and Silverman (2002) argued that the double exponential prior can achieve the adaptive minimax convergence rates which is not obtainable for normal prior.

For γ, a widely used prior is $P(\gamma) = q^{|\gamma|}(1-q)^{p-|\gamma|}$ with a prespecified q. This prior assumes that each predictor enters the model independently with a prior probability q, whether the predictors are correlated or not. The prior models the prior information on the model sizes but does not distinguish models with the same size. However, it is often the case that the presence of highly correlated predictors are to be avoided simply because those predictors are providing similar information on the response. To achieve this, we propose the following prior for γ

$$P(\gamma) \propto q^{|\gamma|}(1-q)^{p-|\gamma|}\sqrt{\det\left(X_\gamma' X_\gamma\right)} \tag{4.2.4}$$

where $det\left(X_\gamma' X_\gamma\right) = 1$ if $|\gamma| = 0$. To see the effect of correlation between predictors in our prior specification, consider the conditional prior odds ratio

for $\gamma_j = 1$:

$$\frac{P(\gamma_j = 1|\gamma^{[-j]})}{P(\gamma_j = 0|\gamma^{[-j]})} = \frac{q}{1-q}\sqrt{\frac{\det\left(X'_{\gamma^{[-j]},\gamma_j=1}X_{\gamma^{[-j]},\gamma_j=1}\right)}{\det\left(X'_{\gamma^{[-j]},\gamma_j=0}X_{\gamma^{[-j]},\gamma_j=0}\right)}}, \tag{4.2.5}$$

where superscript $[-j]$ indicates that the jth component is removed. If X_j is highly correlated with the current covariates, $X_{\gamma^{[-j]},\gamma_j=0}$, the second factor of the right hand side of (4.2.5) will be small. Therefore, it is more likely that X_j will be removed from the full model. This is desirable since X_j does not contain much "additional" information.

Our Bayesian formulation consists of (4.2.2), (4.2.3), and (4.2.4). Three parameters need to be specified for this formulation, namely q, τ and σ^2. From a hierarchical Bayesian point of view, one can either use prespecified values or put a higher level prior for them. Both of these approaches require human expertise. To avoid the need for expert information, we take the empirical Bayes approach and use an automatic default prior parameter choice. The automatic choice of the hyperparameters will be introduced in Section 4.5.

With our formulation, the joint distribution $P(\gamma, \beta_\gamma, Y)$ is

$$P(\gamma, \beta_\gamma, Y) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}} \exp\left(-\frac{||Y - X_\gamma\beta_\gamma||^2 + \lambda\sum_{i\in\gamma}|\beta_i|}{2\sigma^2}\right)(1-q)^p w^{|\gamma|}$$

Therefore,

$$P\left(\gamma|Y\right) = C(Y)w^{|\gamma|}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}}\exp\left(-\frac{||Y - X_\gamma\beta_\gamma||^2 + \lambda\sum_{i\in\gamma}|\beta_i|}{2\sigma^2}\right)d\beta_\gamma, \tag{4.2.6}$$

where

$$w = \left(\frac{q}{1-q} \frac{\tau}{2} \sqrt{2\pi\sigma^2} \right)$$

and $\lambda = 2\sigma^2\tau$. We will pick the model γ which maximizes $P(\gamma|Y)$. In principle, exact evaluation of the posterior probability $P(\gamma|Y)$ could be obtained. However, this task can not be performed in closed form. To compute the high dimensional integrals involved in $P(\gamma|Y)$, analytical or numerical approximation methods are needed.

4.3 Analytical Approach

The major difficulty of the posterior inference for our Bayesian model comes from the high dimensional integration in (4.2.6). Since no analytically tractable solution to this integral exists in general, approximations would be our only resort. Because the posterior probability is expected to spread over a large number of possible models, it is not possible to construct analytical approximations which do uniformly well for all candidate models. Our proposal here is to focus on a subset of candidate models containing the model with the highest posterior probability, whose posterior probabilities can be approximated very well.

Let

$$\beta_\gamma^* = \arg\min_{\beta_\gamma} \left(||Y - X_\gamma \beta_\gamma||^2 + \lambda \sum_{i \in \gamma} |\beta_i| \right).$$

Denote $\beta_\gamma = \beta^*_\gamma + u$. We can rewrite (4.2.6) as

$$P(\gamma|Y) = C(Y)w^{|\gamma|}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}}\exp\left(-\frac{||Y-X_\gamma\beta_\gamma||^2+\lambda\sum_{i\in\gamma}|\beta_i|}{2\sigma^2}\right)d\beta_\gamma$$

$$\begin{aligned}
= \; & C(Y)w^{|\gamma|}\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}} \\
& \times\exp\left(-\frac{||X_\gamma u||^2-2\tilde{Y}'_\gamma X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta^*_i+u_i|-|\beta^*_i|\right)}{2\sigma^2}\right)du \\
& \times\exp\left(-\frac{\min_{\beta_\gamma}\left(||Y-X_\gamma\beta_\gamma||^2+\lambda\sum_{i\in\gamma}|\beta_i|\right)}{2\sigma^2}\right) \qquad (4.3.1)
\end{aligned}$$

where $\tilde{Y}_\gamma = Y - X_\gamma\beta^*_\gamma$ and hereafter, we will omit the subscript γ if no confusion occurs. Our main task would be the evaluation of

$$\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}}\exp\left(-\frac{||X_\gamma u||^2-2\tilde{Y}'X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta^*_i+u_i|-|\beta^*_i|\right)}{2\sigma^2}\right)du \qquad (4.3.2)$$

Define

$$f(u) \equiv \frac{||X_\gamma u||^2-2\tilde{Y}'X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta^*_i+u_i|-|\beta^*_i|\right)}{\sigma^2}. \qquad (4.3.3)$$

Note that the definition of f depends on γ implicitly because it has $|\gamma|$ dimensional argument. Hereafter we omit this dependence for notational convenience. From the definition of u, we see that $f(u)$ is minimized at $u^* = \mathbf{0}$.

Now we consider the following two types of models separately.

Definition 4.3.1 *For a dataset (X, Y) and a given regularization parameter λ*

(i) a model γ is called regular if and only if β^*_γ does not contain zeroes or $|\gamma| = 0$;

(ii) a model γ is called nonregular if β^*_γ contains at least one zero component.

4.3.1 Regular Models

For regular models, $f(u)$ is differentiable at $u = u^*$, and

$$\left.\frac{\partial^2 f(u)}{\partial u \partial u^T}\right|_{u=u^*} = \frac{1}{\sigma^2} X'_\gamma X_\gamma. \tag{4.3.4}$$

Applying the Laplace approximation to (4.3.2), we get

$$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \frac{\sqrt{\det\left(X'_\gamma X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}} \exp\left(-\frac{||X_\gamma u||^2 - 2\tilde{Y}' X_\gamma u + \lambda \sum_{i\in\gamma}\left(|\beta^*_i + u_i| - |\beta^*_i|\right)}{2\sigma^2}\right) du \approx 1. \tag{4.3.5}$$

Thus from (3.1), we have

$$P(\gamma|Y) \approx C(Y) w^{|\gamma|} \exp\left(-\frac{\min_{\beta_\gamma}\left(||Y - X_\gamma\beta_\gamma||^2 + \lambda\sum_{i\in\gamma}|\beta_i|\right)}{2\sigma^2}\right), \tag{4.3.6}$$

where $C(Y)$ is a constant not depending on γ.

4.3.2 Nonregular Models

Although (4.3.6) provides a computationally efficient approximation to $P(\gamma|Y)$ for regular models, it does not apply to nonregular models since for these models $f(u)$ is not differentiable at $u = u^*$. However, we show in the following that in our model selection procedure, we can concentrate on the regular models.

It is beneficial to exclude complex models which do not receive more support from the data than their simpler counterparts. For this reason, we compare a nonregular model γ with a regular submodel of γ. Without loss of generality, assume that γ is of the form $(1, ..., 1, 0, ..., 0)$ where the first $|\gamma|$ components are ones, and that only the first s components of the $|\gamma|$ dimensional vector β^*_γ are nonzero. By the definition of nonregular models, $s < |\gamma|$. Let γ^* be the p dimensional binary vector representing a submodel of γ with only the first s elements being 1. Our task here is to compare $P(\gamma|Y)$ and $P(\gamma^*|Y)$. Since $f(u)$ is minimized at $u = \mathbf{0}$, for any $i \leq s$,

$$\left.\frac{\partial f}{\partial u_i}\right|_{u=\mathbf{0}} = 0,$$

which leads to

$$2\tilde{Y}'X_i = \lambda sign(\beta^*_{\gamma,i}), \qquad \text{if } i \leq s. \tag{4.3.7}$$

On the other hand, for $s < i \leq |\gamma|$, the ith component of β^*_γ is zero, and we have

$$\left.\frac{\partial f}{\partial u_i}\right|_{u_i=0^+;u_j=0,\forall j\neq i} \geq 0, \qquad \left.\frac{\partial f}{\partial u_i}\right|_{u_i=0^-;u_j=0,\forall j\neq i} \leq 0.$$

This implies that

$$\left|2\tilde{Y}'X_i\right| \leq \lambda, \quad \text{and} \quad \beta^*_{\gamma,i} = 0, \qquad \text{if } s < i \leq |\gamma|. \tag{4.3.8}$$

By (4.3.7) and (4.3.8), from simple calculations we get,

$$\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X_\gamma' X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}}\exp\left(-\frac{||X_\gamma u||^2-2\tilde{Y}' X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta_i^*+u_i|-|\beta_i^*|\right)}{2\sigma^2}\right)du$$

$$<\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\frac{\sqrt{\det\left(X_\gamma' X_\gamma\right)}}{\left(\sqrt{2\pi\sigma^2}\right)^{|\gamma|}}\exp\left(-\frac{||X_\gamma u||^2}{2\sigma^2}\right)du=1.$$

Thus,

$$P(\gamma|Y)<C(Y)w^{|\gamma|}\exp\left(-\frac{\min_{\beta_\gamma}\left(||Y-X_\gamma\beta_\gamma||^2+\lambda\sum_{i\in\gamma}|\beta_i|\right)}{2\sigma^2}\right). \quad (4.3.9)$$

Now because $\tilde{Y}_\gamma=\tilde{Y}_{\gamma^*}$ and $\beta^*_{\gamma,i}=\beta^*_{\gamma^*,i}$ for any $i\le s$, applying (4.3.6) to the regular model γ^* and (4.3.9) to the nonregular model γ, we conclude that asymptotically,

$$\frac{P(\gamma|Y)}{P(\gamma^*|Y)}\le w^{|\gamma|-s}.$$

If $w\le 1$, the data do not give more support to the bigger model γ than γ^* and thus we would pick γ^*. Consequently, we can avoid computing $P(\gamma|Y)$ for nonregular model γ.

4.4 Connection between the LASSO and the Bayesian framework

Summarizing the above analysis, we find that if w is set to 1, then

(i) To search for the model with the highest posterior probability, we can concentrate on the regular models.

(ii) For regular models, the posterior probability $P(\gamma|Y)$ can be approximated by

$C(Y)\exp[-h(\gamma)/(2\sigma^2)]$, where

$$h(\gamma) = \min_{\beta_\gamma} \left(||Y - X_\gamma \beta_\gamma||^2 + \lambda \sum_{i \in \gamma} |\beta_i| \right)$$

In general, these conclusions are good approximations; for the special case of orthogonal design matrix X, they can be proved rigorously:

Theorem 4.4.1 *Suppose that $w = 1$. Under orthogonal design, i.e. $X'X = (n-1)I_p$*

(i) If model γ is nonregular, then there exists a γ^* such that $P(\gamma|Y) < P(\gamma^*|Y)$

(ii) Suppose that $\lambda = o(\sqrt{n})$. If model γ is regular, then as $n \to \infty$,

$$P(\gamma|Y) \to C(Y)\exp\left(-\frac{||Y - X_\gamma \beta^*_\gamma||^2 + \lambda \sum_{i\in\gamma} |\beta^*_i|}{2\sigma^2} \right).$$

We note the $(n-1)$ factor in the condition $X'X = (n-1)I_p$ is just to conform to our convention we made at the beginning of the chapter that the data be scaled to have sample standard deviation one.

By (i) and (ii), we can now focus on searching for the regular model γ with the smallest $h(\gamma)$. A straightforward search involves going through each of the large number of candidate models to identify regular models and to minimize h over all the regular models. This would be computationally very demanding. Fortunately, such a search is not necessary, and the following proposition provides us with the key to a simple and explicit recipe for find a regular model that minimizes $h(\gamma)$.

Proposition 4.4.1 *Let* $\widehat{\beta} = \arg\min_\beta (\|Y - X\beta\|^2 + \lambda \sum_{i=1}^p |\beta_i|)$*, and let model* $\widehat{\gamma}$ *be such that* $\widehat{\gamma}_i = I(\widehat{\beta}_i \neq 0)$*, where* $I(\cdot)$ *is the indicator function. Then* $\widehat{\gamma}$ *is the regular model that minimizes* $h(\gamma)$*.*

Interestingly, $\widehat{\gamma}$ is exactly the same model selected by the LASSO. In other words, the model selected by the LASSO has the highest posterior probability under our Bayesian model (4.2.4) with $w = 1$. Therefore we can use the LASSO algorithm to select a model with approximately the largest posterior probability when $w = 1$. The LASSO also gives the maximum a posteriori estimate of the regression coefficients for the selected model at the same time. Thus, if our goal is to select a model and estimate the regression coefficient, we can use the LASSO to fulfill the task. This equivalence allows us to take advantage of the recently developed fast LASSO algorithm to compute the solution for our Bayesian formulation. The connection with the LASSO estimator also highlights a distinction between our empirical Bayes methods and earlier proposals such as that in George and Foster (2000) which can only be implemented in a stepwise fashion in most practical situation. Using the LASSO algorithm to calculate the Bayesian solution would save tremendous computational effort and make our procedure suitable for large datasets with high dimensionality.

The close relationship also gives a new Bayesian interpretation to the LASSO. Tibshirani (1996) mentioned that the LASSO has another Bayesian interpretation with independent double exponential prior on each regression coefficient. Tibshirani's formulation is somehow less natural as a Bayesian variable selection

procedure since it puts prior probability one on the full model. Consequently, the corresponding posterior probability for the full model will also be one even if the posterior modal estimates of some regression coefficients are zero. Therefore, it is not practical to assess the uncertainty of a selected model. In contrast, the posterior probabilities of our formulation provide a measure of the model uncertainty, see Section 4.7.

4.5 Prior Elicitation

To start the analysis, we need to specify σ^2, q and τ or equivalently, σ^2, λ and w. To take the computational advantage of the LASSO, we set $w = 1$. The remaining hyperparameters we need to take care of are σ^2 and λ. The later is exactly the tunning parameter selection problem faced by the LASSO. Tibshirani (1996) proposed a GCV score to select λ. In the following, we adopt an empirical Bayesian approach for selecting both σ^2 and λ.

From an empirical Bayesian point of view, one could choose σ^2 and λ by maximizing the marginal likelihood of Y, which is

$$f(Y|\sigma^2, \lambda) = \sum_{\gamma} \int_{-\infty}^{\infty} P(Y, \gamma, \beta_\gamma) d\beta_\gamma.$$

This can be implemented when the number of variables is small. However, in situations where the number of variables is moderately large, the summation is over a large number of items, and is not practical for large datasets. In such situations we can consider maximizing condition likelihood as proposed by

George and Foster (2000).

For a give λ, denote $\widehat{\gamma}_\lambda$ the selected model. The conditional likelihood of Y given the selected model is

$$\begin{aligned} f\left(Y|\widehat{\gamma}_\lambda,\sigma^2,\lambda\right) &= \int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{||Y-X_{\widehat{\gamma}_\lambda}\beta_{\widehat{\gamma}_\lambda}||^2}{2\sigma^2}\right) \\ &\quad \times\left(\frac{\lambda}{4\sigma^2}\right)^{|\widehat{\gamma}_\lambda|}\exp\left(-\frac{\lambda\sum_{i\in\widehat{\gamma}_\lambda}|\beta_i|}{2\sigma^2}\right)d\beta_{\widehat{\gamma}_\lambda} \\ &\approx \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-|\widehat{\gamma}_\lambda|}\left(\frac{\lambda}{4\sigma^2}\right)^{|\widehat{\gamma}_\lambda|}\left(\det\left(X'_{\widehat{\gamma}_\lambda}X_{\widehat{\gamma}_\lambda}\right)\right)^{-1/2}\times \\ &\quad \times\exp\left(\frac{-\min_\beta\left(||Y-X_{\widehat{\gamma}_\lambda}\beta_{\widehat{\gamma}_\lambda}||^2+\lambda\sum_{i\in\widehat{\gamma}_\lambda}|\beta_i|\right)}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-|\widehat{\gamma}_\lambda|}\left(\frac{\lambda}{4\sigma^2}\right)^{|\widehat{\gamma}_\lambda|}\left(\det\left(X'_{\widehat{\gamma}_\lambda}X_{\widehat{\gamma}_\lambda}\right)\right)^{-1/2}\times \\ &\quad \times\exp\left(-\frac{\min_\beta\left(||Y-X\beta||^2+\lambda\sum_{i=1}^p|\beta_i|\right)}{2\sigma^2}\right) \end{aligned}$$

We used (4.3.1) and (4.3.5) in the above approximation, which holds since the selected model is regular. We choose σ^2 and λ as the maximizer of conditional likelihood $f\left(Y|\widehat{\gamma}_\lambda,\sigma^2,\lambda\right)$. Simple calculations show that this is equivalent to choosing λ by minimizing

$$\begin{aligned} CML(\lambda) &\equiv (n+|\widehat{\gamma}_\lambda|)\left[\ln\left(\frac{\min_\beta\left(||Y-X\beta||^2+\lambda\sum_{i=1}^p|\beta_i|\right)}{n+|\widehat{\gamma}_\lambda|}\right)+1\right] \\ &\quad +\ln\left(\det\left(X'_{\widehat{\gamma}_\lambda}X_{\widehat{\gamma}_\lambda}\right)\right)-2|\widehat{\gamma}_\lambda|\ln\left(\sqrt{2\pi}\lambda/4\right), \end{aligned}$$

and the estimate of σ^2 is $\min_\beta\left(||Y-X\beta||^2+\lambda\sum_{i=1}^p|\beta_i|\right)/(n+|\widehat{\gamma}_\lambda|)$.

Remark Both C_p used in Efron *et. al.* (2004) and the empirical Bayes approach

previously proposed by George and Foster (2000) need to estimate σ^2 by fitting the full model and therefore can only be applied in the situation $p < n$. Our proposal here avoids this problem.

Remark It is interesting to notice that the derivation for CML does not work for the Bayesian interpretation given by Tibshirani (1996) mentioned at the end of Section 4.4, since the approximation only applies to regular models. However, it is tempting to maximize $f\left(Y|\widehat{\beta}_\lambda, \widehat{\gamma}_\lambda, \sigma^2, \lambda\right)$ instead. Unfortunately, it leads to a criterion

$$\min_{\beta}\left(||Y - X\beta||^2 + \lambda\sum_{i=1}^{p}|\beta_i|\right)$$

which is trivially minimized at $\lambda = 0$.

4.6 Simulations

In this section, we compare the proposed CML based empirical Bayes procedure with several other popular approaches for variables selection and estimation. The methods compared include

(i) (CML) Our empirical Bayes estimate with hyper-parameters selected by CML;

(ii) (C_p) The LASSO with λ selected by C_p;

(iii) (GCV) The LASSO with λ selected by GCV;

(iv) (GFF) The empirical Bayes approach proposed George and Foster (2000)

and implemented in a forward selection fashion. George and Foster proposed a conditional maximum likelihood method to choose the hyperparameters in their Bayes formulation.

We compare these methods in terms of the size of selected models, model error, and the computation time on a Pentium III 750M computer. All simulations were conducted using *R*. The path of LASSO estimate was computed using the *LARS* package. The model error of an estimate $\widehat{\beta}$ is given by

$$ME(\widehat{\beta}) = \left(\widehat{\beta} - \beta\right)' V \left(\widehat{\beta} - \beta\right),$$

where $V = E(X'X)$ is the population covariance matrix of X. The models in our simulation example have also been used in Tibshirani (1996).

Example 4.6.1 *Consider the following four models*

I. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ *and* $\sigma = 3$. *The correlation between* $X_{\cdot i}$ *and* $X_{\cdot j}$ *is* $\rho^{|i-j|}$ *with* $\rho = 0.5$.

II. Same as (I) except that $\beta_j = 0.85$ *for all* j.

III. Same set-up as before, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)'$ *and* $\sigma = 2$.

IV. Forty correlated predictors are considered. $x_{ij} = z_{ij} + w_i$, *where* z_{ij}, w_i *are independent standard normal random variables. The true regression coefficients are* 2 *for the first 20 predictors and* 0 *for the other predictors.*

For the first three models, two hundred datasets with sample size 20 *were generated. For the fourth model, two hundred datasets with sample size 100 were generated.*

Table 4.1: Comparisons on the Simulated Datasets – Model I

	CML	C_p	GCV	GFF
Average Size	5.145	5.290	7.370	5.655
	(1.188)	(1.593)	(0.704)	(3.029)
Median ME	3.216	3.669	3.452	5.535
Average ME	3.994	5.188	4.522	6.370
	(2.179)	(2.084)	(1.902)	(4.746)
Average CPU Time	0.110	0.048	0.232	0.283
(secs)	(0.014)	(0.006)	(0.028)	(0.017)

Table 4.2: Comparisons on the Simulated Datasets – Model II

	CML	C_p	GCV	GFF
Average Size	5.675	5.700	7.225	6.795
	(1.177)	(1.456)	(0.798)	(2.515)
Median ME	4.490	4.566	3.978	5.601
Average ME	4.950	5.601	4.763	6.551
	(3.199)	(4.475)	(3.472)	(4.662)
Average CPU Time	0.110	0.048	0.233	0.276
(secs)	(0.025)	(0.011)	(0.046)	(0.043)

Tables 4.1-4.4 give the means/medians and standard deviations (in the parentheses) over the 200 simulated datasets. We see from the tables that CML tends to select models with relatively smaller sizes than the other methods. To see whether the choice of sparse models comes with sacrificing the prediction accuracy, we provide pairwise prediction accuracy comparison between CML and the other methods for Model I-IV in Figure 4.1.

Model I has a signal to noise ratio approximately 5.7. The first row of Figure 4.1 gives the pairwise comparison between CML and the other three methods based on the model errors for the 200 simulated datasets. We can see that CML

Table 4.3: Comparisons on the Simulated Datasets – Model III

	CML	C_p	GCV	GFF
Average Size	4.235	4.060	7.265	1.315
	(1.315)	(2.200)	(0.740)	(1.246)
Median ME	0.928	1.003	1.307	0.150
Average ME	1.191	1.817	1.698	0.635
	(1.180)	(2.403)	(1.399)	(1.836)
Average CPU Time	0.113	0.050	0.236	0.290
(secs)	(0.012)	(0.007)	(0.025)	(0.019)

Table 4.4: Comparisons on the Simulated Datasets – Model IV

	CML	C_p	GCV	GFF
Average Size	25.450	27.260	33.430	6.890
	(2.248)	(4.586)	(3.150)	(1.275)
Median ME	60.246	74.340	84.191	183.319
Average ME	61.184	80.216	87.183	183.472
	(14.894)	(32.016)	(27.942)	(37.666)
Average CPU Time	0.870	0.296	5.022	8.869
(secs)	(0.091)	(0.031)	(0.361)	(0.299)

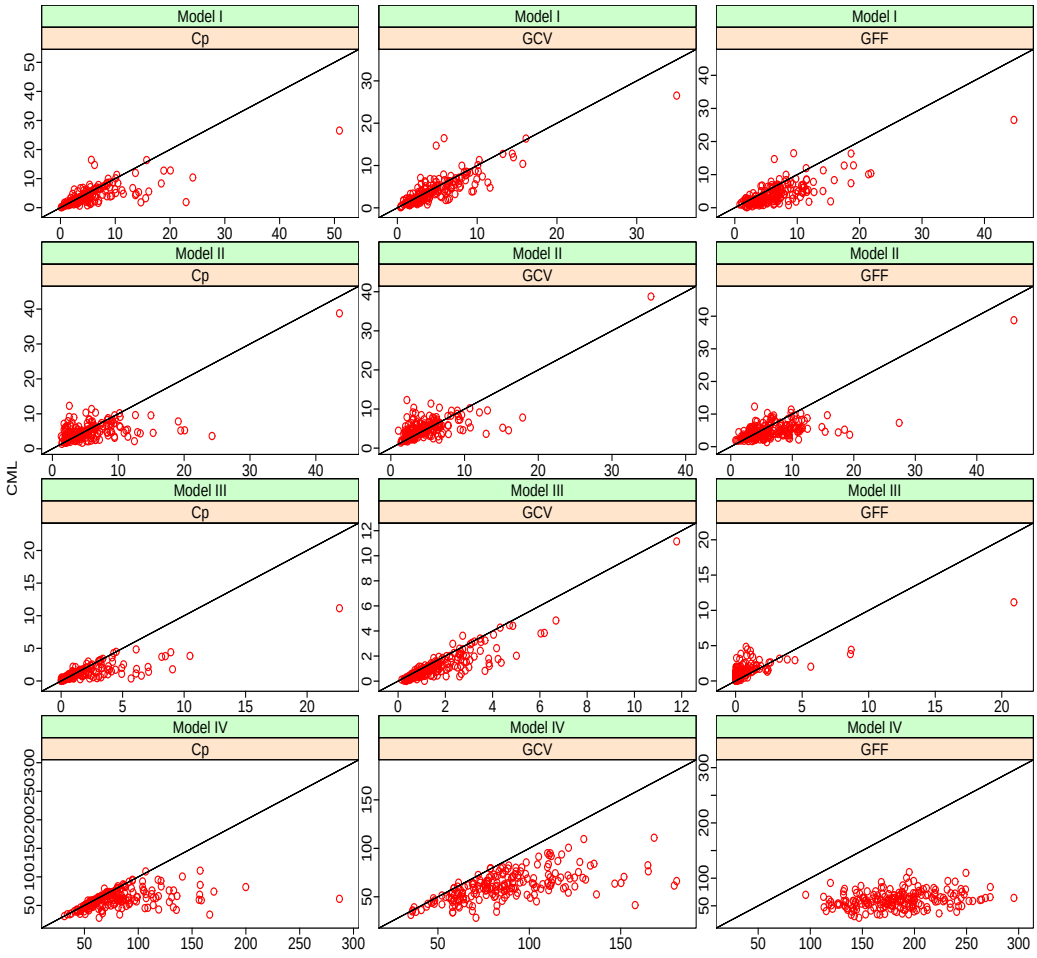

Figure 4.1: Pairwise Prediction Accuracy Comparison between CML and Other Methods for Example 4.6.1

performs the best for this model.

Model II has a lower signal to noise ratio, which is approximately 1.8. As one may see from Table 4.2 and the second row of Figure 4.1, CML achieves good prediction accuracy with a small model size.

Model III represents a set-up well suited for stepwise subset selection with signal to noise ratio about 7. In this case GFF, which uses a stepwise procedure, performs the best, followed by CML.

Model IV is a bigger model with signal to noise ratio about 9. CML performs the best. GFF selected models with too few predictors and consequently has a larger bias than the other methods.

In summary, GCV tends to select models with large sizes and its prediction performance is better in the situation where most predictors are in the true model. Since the empirical Bayes approach proposed by George and Foster (2000) can only be implemented through a stepwise greedy search, it inherits both the advantages and disadvantages of the greedy search methods. Because the algorithm is myopic, as noted in earlier studies (Chen, Donoho and Sauders, 1999), it might work perfectly if the size of the true model is small but in other cases, it might make suboptimal choices in the first few iterations and end up spending most of its time correcting the mistakes made in the first few terms. In general, CML compares favorably with other methods.

The simulation also indicates that CML and C_p enjoy favorable computation speed. GCV is slower mostly because of the evaluation of the trace of the

information matrix. The forward selection method is slower than the other methods, especially in the larger model.

4.7 Estimating Posterior Probabilities

Another potential advantage of our Bayesian formulation is that it enables us to estimate the the posterior probability for each candidate model, providing a natural means of assessing the model uncertainty. Although computationally efficient, the aforementioned analytic approach does not provide us with these quantities. In order to obtain an estimate of the posterior probability, we need an alternative strategy – stochastic search.

The posterior probabilities of the models depend on the hyperparameters in the Bayesian setup, which we estimate with the method introduced in Section 4.5. We propose a Gibbs sampling scheme to compute the estimated posterior probabilities, and use the resulting posterior probabilities as an approximate assessment of model uncertainty. For a fully Bayesian approach, one would choose the prior hyperparameters with expert knowledge or put another level of prior on the hyperparameters. Our algorithm can still be used for these setting with mild modifications. In our application here, we adopt a Gibbs sampler which generates a stochastic process moving through the model space. By summarizing the samples from the stochastic processes, various statistics of the posteriori such as the posterior probability of a selected model and the posterior mean of the regression coefficient can be obtained. Alternatively,

Bayesian model averaging techniques could also be applied (Raftery, Madigan and Hoeting, 1996).

Rewrite (4.2.2), (4.2.3) and (4.2.4) as

$$\begin{aligned} Y|\gamma,\beta &\sim N\left(X_\gamma\beta_\gamma,\sigma^2 I_{(n)}\right) \\ \beta_j|\gamma_j,v_j &\sim (1-\gamma_j)\delta(0)+\gamma_j N\left(0,v_j\eta^2\right), \qquad j=1,\cdots,p \\ v_j &\sim Exp(1), \qquad j=1,\cdots,p \\ P(\gamma) &\propto q^{|\gamma|}(1-q)^{p-|\gamma|}\sqrt{\det\left(X_\gamma' X_\gamma\right)}, \end{aligned}$$

where $\eta=\sqrt{2}/\tau$. A Gibbs sampler can be devised to sample from the above formulation:

(i) Choose initial values for γ,β and v

(ii) Generate the block $\gamma_j,\beta_j|Y,v,\gamma^{[-j]},\beta^{[-j]}$ by

  (a) Generate $\gamma_j|Y,v,\gamma^{[-j]},\beta^{[-j]}=\gamma_j|Y,v_j,\gamma^{[-j]},\beta^{[-j]}$

  (b) Generate $\beta_j|Y,v,\gamma,\beta^{[-j]}=\beta_j|Y,v_j,\gamma,\beta^{[-j]}$

(iii) Generate $v_j|Y,v^{[-j]},\gamma,\beta=v_j|Y,\gamma_j,\beta$

(iv) Repeat (ii) and (iii) for $j=1,\ldots,p$

The proposed sampling scheme is irreducible and aperiodic, because it is readily to be checked that in one step the sampler can reach any point in the parameter space from any other point. Each of the three sampling steps has closed-form density and can be implemented efficiently.

4.7.1 Sampling Scheme

Sample from $\gamma_j|Y, v_j, \gamma^{[-j]}, \beta^{[-j]}$

To sample $\gamma_j|Y, v_j, \gamma^{[-j]}, \beta^{[-j]}$, we note that

$$P(\gamma_j = 1|Y, v_j, \gamma^{[-j]}, \beta^{[-j]}) = \frac{1}{1 + \frac{f(Y|\beta^{[-j]}, v, \gamma^{[-j]}, \gamma_j=0)P(\gamma^{[-j]}, \gamma_j=0)}{f(Y|\beta^{[-j]}, v, \gamma^{[-j]}, \gamma_j=1)P(\gamma^{[-j]}, \gamma_j=1)}},$$

where

$$f(Y|\beta^{[-j]}, v, \gamma^{[-j]}, \gamma_j = 0) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{z'z}{2\sigma^2}\right),$$

and

$$\begin{aligned}
& f(Y|\beta^{[-j]}, v, \gamma^{[-j]}, \gamma_j = 1) \\
= & \int_{-\infty}^{\infty} f(Y|\beta, v, \gamma^{[-j]}, \gamma_j = 1) f(\beta_j|v_j, \gamma_j = 1) d\beta_j \\
= & \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{z'z - 2z'X_j\beta_j + X_j'X_j\beta_j^2}{2\sigma^2}\right) \times \\
& \times \frac{1}{\sqrt{2\pi v_j\eta^2}} \exp\left(-\frac{\beta_j^2}{2v_j\eta^2}\right) d\beta_j \\
= & \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{z'z}{2\sigma^2}\right) \exp\left(\frac{(z'X_j)^2 v_j\eta^2}{2\sigma^2(\sigma^2 + X_j'X_j v_j\eta^2)}\right) \sqrt{\frac{\sigma^2}{\sigma^2 + X_j'X_j v_j\eta^2}}
\end{aligned}$$

and

$$z = Y - X_{\gamma^{[-j]}, \gamma_j=0}\beta_{\gamma^{[-j]}, \gamma_j=0}.$$

Sample from $\beta_j|Y, v_j, \gamma, \beta^{[-j]}$

Once γ_j is generated, the sampling of β_j is straightforward.

$$\begin{aligned}
\beta_j|Y, \gamma^{[-j]}, \beta^{[-j]}, v, \gamma_j = 0 &= 0 \\
\beta_j|Y, \gamma^{[-j]}, \beta^{[-j]}, v, \gamma_j = 1 &\propto \prod_{i=1}^{n} \phi(Y_i| \left(X_{\gamma^{[-j]},\gamma_j=1}\beta_{\gamma^{[-j]},\gamma_j=1}\right)_i, \sigma^2)\phi(\beta_j|0, 2v_j\eta^2) \\
&\sim N\left(\frac{v_j\eta^2(z'X_j)}{v_j\eta^2(X_j'X_j)+\sigma^2}, \frac{v_j\eta^2\sigma^2}{v_j\eta^2(X_j'X_j)+\sigma^2}\right),
\end{aligned}$$

where $\phi(\cdot|\mu, \alpha^2)$ stands for the density function of a normal distribution with mean μ and variance α^2.

Sample from $v_j|Y, \beta, \gamma_j$

Depending on the value of γ_j, the conditional distribution of v_j is given by

$$\begin{aligned}
v_j|Y, \beta, v^{[-j]}, \gamma_j = 0 &\sim Exp(v_j|1) \\
v_j|Y, \beta, v^{[-j]}, \gamma_j = 1 &\sim \phi(\beta_j|0, 2v_j\eta^2)Exp\left(v_j|1\right),
\end{aligned}$$

where $Exp(\cdot|a)$ is the density function of a exponential distribution with scale parameter a. When $\gamma_j = 1$, the sampling distribution is not of a standard form, and we used the ratio of uniform method (Wakefield, Gelfand and Smith, 1992) to sample from it. It has previously been demonstrated by Choy and Smith (1997) that this method is very efficient for sampling from densities of the current form.

4.7.2 Summarizing the Samples

After the burn-in period, we can collect a sufficiently large number of samples from the Gibbs sampler. The basic idea behind this stochastic search approach is that these samples approximately follow the posterior distribution $u, \gamma, \beta|Y$. Therefore, Bayesian inferences can be done by summarizing these samples. For example, if we are interested in the posterior probability for a given model $\widehat{\gamma}$, we can estimate it by the empirical frequency of $\widehat{\gamma}$ in the samples. The posterior probability for a specific element of γ to be 1 can also be estimated in the same fashion. Like other MCMC type methods, the main difficulty of implementing the Gibbs sampler is to determine the length of burn-in period and to decide how many samples to collect. These issues are beyond the scope of the current chapter and interested readers are referred to Carlin and Louis (1996) or Gelman, Carlin, Stern and Rubin (2004).

To demonstrate the usage of the stochastic approach, we conducted a small simulation example which is similar to Example 4.1 of George and McCulloch (1993).

Example 4.7.1 *Consider $n = 60$ observations simulated from the model:*

$$Y = X\beta + \sigma\varepsilon,$$

where $\beta = (0, 0, 0, 1, 1.2)'$, $\sigma = 2.5$ and ε follows a standard normal distribution. In order that we can compute the "exact" posterior probabilities defined

as (4.2.1) for each candidate model, we chose an orthogonal design. The design matrix was generated by centering and orthogonalizing 60 *samples from* $N_5(\mathbf{0}, I_5)$.

There are a total of 32 candidate models, of which 8 have exact posterior probabilities larger than 1 percent. Table 4.5 reports the eight models with highest posterior probabilities together with their exact posterior probabilities and their estimated posterior probabilities based on $20,000$ samples from the Gibbs sampler output after $5,000$ burn-in samples.

Table 4.5: Estimate of Posterior Probabilities

Variables	Exact $P(\mathcal{M}\|Y)$	MCMC Estimate
x_4, x_5	0.16930	0.18330
x_2, x_4, x_5	0.12856	0.13270
x_3, x_4, x_5	0.12856	0.13065
x_1, x_4, x_5	0.12856	0.12570
x_2, x_3, x_4, x_5	0.09762	0.10060
x_1, x_2, x_4, x_5	0.09762	0.09350
x_1, x_3, x_4, x_5	0.09762	0.08905
x_1, x_2, x_3, x_4, x_5	0.07412	0.06770

From Table 4.5, we see that the Gibbs sampler provides fairly accurate estimate to the posterior probabilities. Also, the first model was picked by the LASSO estimate which means that the LASSO selects the model with the highest posterior probability.

In order to check the coefficient estimates, Figure 4.2 gives the kernel density estimates of posterior distributions for β_4 and β_5 given $\gamma = (0, 0, 0, 1, 1)$. The vertical line in each panel corresponds to the LASSO estimate. Since the LASSO

picks the model with the highest posterior probability, the LASSO estimate corresponds to the exact posterior modal estimate. Figure 4.2 confirms that the Gibbs sampler also provides accurate coefficient estimates.

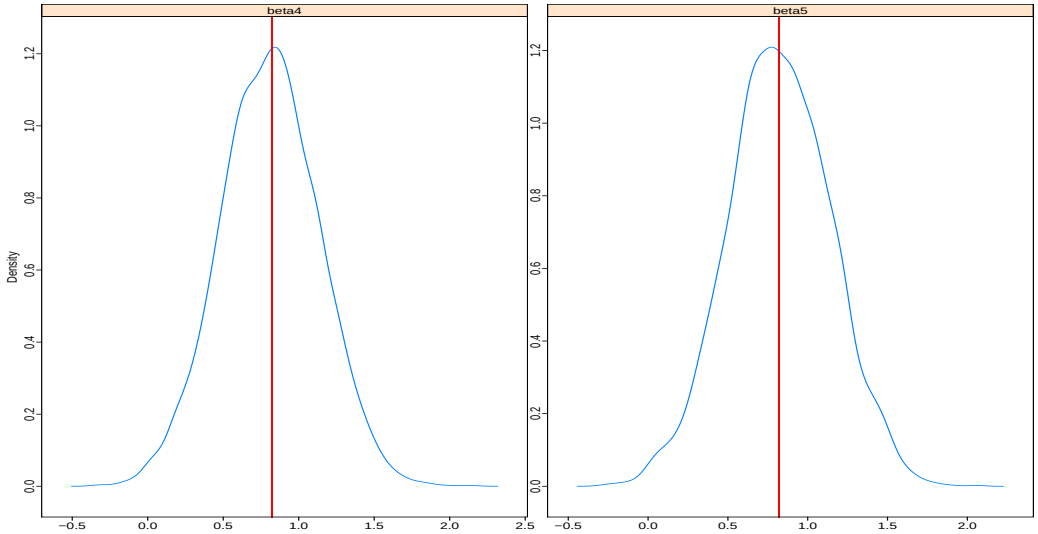

Figure 4.2: Densities of the Posterior Distributions of Regression Coefficients from Example 4.7.1

4.8 Real Example

We apply our method to the prostate dataset from the study by Stamey *et. al.* (1989). This dataset, previously used in Tibshirani (1996), consists the medical records of 97 male patients who were about to receive a radical prostatectomy. The response variable is the level of prostate specific antigen. The predictors are eight clinical measures: log(cancer volume) (lcavol), log(prostate weight)

(lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45).

We first compare the prediction performance of the four different methods from Section 4.5. The prediction performance was estimated using 10 fold cross validation. Table 4.6 provides the average model sizes and prediction errors (PE) for different methods.

Table 4.6: Prediction Accuracy

Method	CML	C_p	GCV	GFF
Size	6.5	6.4	7.7	6.5
	(0.71)	(0.97)	(0.48)	(2.4)
PE	0.55	0.57	0.55	0.59
	(0.23)	(0.24)	(0.24)	(0.22)

Table 4.6 shows that CML gives a very competitive performance. CML and GCV have the best overall prediction. But GCV tends to select much bigger models than CML. Applying CML on the full dataset, **lcp** and **gleason** are excluded from the selected model. It is worth noting that as reflected by the standard deviation, GFF is much less stable than the other methods and selected models with very different sizes. Table 4.7 reports the posterior mode estimate of the regression coefficients given the selected model.

In order to evaluate the model uncertainty for the selected model. We ran the Gibbs sampler and collected $20,000$ samples after a $5,000$ burn-in period. The selected model turned out to be the most frequently visited model with an

Table 4.7: Posterior Mode Coefficient Estimate

Variables	Coefficient Estimate
lcavol	0.735
lweight	0.097
age	-0.433
lbph	0.159
svi	0.103
lcp	0
gleason	0
pgg45	1.918

estimated posterior probability 6.08%. A total of 108 models were visited in the $20,000$ samples, thirty two of which had estimated posterior probabilities larger than 1%. Table 4.8 reports the estimated posterior probabilities for the seven most frequently visited models.

Table 4.8: Seven Most Frequently Visited Models

Excluded Variables	$P(\mathcal{M}\|Y)$
lcp, gleason	6.080%
age, lcp, gleason	5.310%
lcp, gleason, pgg45	5.020%
lcp, pgg45	5.015%
age, lcp, gleason, pgg45	4.875%
age, lcp, pgg45	4.635%
lcp	4.105%

Although the 6-predictor model is the model with the highest posterior probability, its posterior probability is only 6.08%. This is reasonable because we have a large number of candidate models and consequently the confidence in any specific model would not be high. This phenomenon suggests looking at the

posterior distribution from another angle. Instead of looking at the posterior probability of the models, one can look at each variable individually. We estimated $P(\gamma_i|Y)$ for $i = 1, \ldots, 8$ using the same output from the Gibbs sampler. Table 4.9 reports these posterior probabilities.

Table 4.9: Probability of Inclusion for Individual Variables

Variables	Probability of Inclusion
lcavol	100.00%
lweight	94.151%
age	50.395%
lbph	68.175%
svi	97.535%
lcp	38.740%
gleason	43.870%
pgg45	49.625%

From this table we see that the most important predictors are **lcavol**, **lweight** and **svi**. These variables have posterior probabilities close to 100% and should be included in the final model. There is also moderate evidence showing that **lbph** has some influence on the response. It has posterior probability 68.135%. Other predictors do not seem to have a strong effect on the response. These include **age**, **pgg45**, **lcp** and **gleason**. In fact, based on output from the Gibbs sampler, the probability that at least one of these four variables is present in the model is 62.605%.

4.9 Summary

We developed an empirical Bayes method for variable selection and estimation in linear regression models. The method is based on a particular hierarchical Bayesian formulation and the parameters including the error variance in the linear model are estimated through a conditional maximum likelihood method. Analytical approximations to the posterior probabilities reveal the intimate relationship between the estimator from our Bayesian formulation and the LASSO. This connection allows us to compute the Bayesian estimate with the quick LASSO algorithm. The empirical Bayes choice of the hyperparameters also provides a new way to select the tuning parameter for the LASSO. The Bayesian formulation also provides a natural means to assess model and variable uncertainty through estimated posterior probabilities. We presented a Gibbs sampler to generate samples from the posterior distribution, which were used to assess model and variable uncertainty.

Chapter 5

Discussions and Future Works

In this thesis, we investigate the penalized likelihood estimate in three different settings. Potential generalization of the content of this thesis could be pursued in a few important directions.

The work presented in Chapter 2 was originally motivated by the covariance modeling problem. Modeling of covariances is an important problem in numerical weather prediction (NWP). See, for example Derber and Bouttier (1999), Dee *et. al.* (1999), Hollingsworth and Lönnberg (1986), Gong *et. al.* (1998), Purser and Parrish (2003) and Chapnik *et. al.*(2003). Note that although the example in Chapter 2 is univariate, the method applies to spatial data. A first generalization of the method in this paper would suppose that the ε_i in (2.1.1) have mean 0 and variance 1, but have a correlation matrix $C(x, x')$ which is known up to a small number of estimable parameters θ, and the challenge is to estimate $\sigma(x)$ and θ simultaneously. Generalizing further to a problem important in NWP, copious observations are available on the difference between observations and forecast. Since observation and forecast errors are generally modelled as independent, the y_i could be modeled as coming from a spatial process with covariance $\sigma_B(x)\sigma_B(x')C_B(x, x') + \sigma_R(x)\sigma_R(x')C_R(x, x')$, where B

refers to forecast and R to observations, and the $C's$ are correlation matrices known up to possibly a few parameters. μ is the difference between the observation and forecast biases, but under the commonly made assumption that observations are unbiased, μ would be the forecast bias. In practice C_B and C_R will have very different structure, so that it may be possible in some cases to estimate the σ's by generalization of the techniques in this paper. Chapnik *et. al.* discuss the case where σ_B and σ_R are constants to be estimated. It is believed that this and other generalizations have potential for a variety of important applications.

In Chapter 3, we compared computable approximations to the cross-validation and unbiased risk estimate as smoothing parameter tuning methods for nonparametric Poisson regression. A more fundamental question is to what extent this comparison can be applied to the exact unbiased risk estimate and the leave-out-one cross-validation. It would also be interesting to know whether or not this superiority of UBR based tuning methods can be extended for other likelihood models.

In Chapter 4, we developed a general empirical Bayes framework for variable selection and estimation in linear regression models. Two complementary ways of conducting the posterior analysis have also been introduced. Our attention has been mainly paid to the analytical approach which reveals the intimate connection between the Bayes method and the LASSO estimator. It is quite interesting to see if this connection can be generalized to nonparametric variable

selection problems. There are also quite a few issues that can be addressed to make the stochastic approach applicable in more general situations. For example, in some applications, slow convergence of the Gibbs sampler might be encountered. In these cases, acceleration of the convergence could be considered. One such example is the case where highly correlated predictors exist. The sampler might get stuck for a long time. A possible remedy is to sample a group of βs together instead of individually. Alternatively, an additional step which shuffles these predictors with certain probability can be added to the sampler. For more general situations, it would be of great practical importance to check how the techniques discussed in Gelfand and Sahu (1994) can be applied in the current setting. Different from other existing methods, our procedure has potential to be applied to the cases where we have less observations than covariates. One such situation is the microarray data analysis. It would be great to further explore this application of our method.

Bibliography

[1] Andersen, T.G. and Lund, J. (1997), Estimating continuous time stochastic volatility function models of the short interest rate. *J. Economet.* **77**, 343-377.

[2] Andrews, D. F. and Mallows, C. L. (1974), Scale mixtures of normal distributions, *J. Royal. Statist. Soc. B.*, **36**, 99-102.

[3] Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373-384.

[4] Carlin, B. P. and Louis, T. A. (1996) *Bayes and empirical Bayes methods for data analysis*, Chapman & Hall, London.

[5] Carroll, R.J. (1982), Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10**, 1224-1233.

[6] Chapnik, B., Desroziers, G., Rabier, F. and Talagrand, O. (2003), Properties and first application of an error statistics tuning method in variational assimilation, to appear in *Monthly Weather Review*.

[7] Chen, S. S., Donoho, D. L. and Sauders, M. A. (1999), Atomic Decomposition by Basis Pursuit, *SIAM J. Scientific Computing*, **20**, 33-61.

[8] Chipman, H., George, E. I. and McCulloch, R. E. (2001), The practical

implementation of Bayesian model selection (with discussion), *IMS Lecture Notes Monogr. Ser., 38, Model selection*, 65-134.

[9] Choy, S. T. B. and Smith, A. F. M. (1997) Hierchical models with scale mixtures of normal distributions, *Test*, **6**, 205-221.

[10] Craven, P., Wahba, G. (1979), Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* **31(4)**, 377–403.

[11] Dee, D., Gaspari, G., Redder, C., Rukhovets, L. and A. da Silva (1999), Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications, *Monthly Weather Review*, **8**, 1835-1849.

[12] Derber, J. and Bouttier, F. (1999), A reformulation of the background error covariance in the ECMWF global data assimilation system, *Tellus*, **51A**, 195-221.

[13] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004), Least angle regression, *Ann. Statist.*, **32** 407-499.

[14] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96** 1348-1360.

[15] Fan, J.Q. and Yao, Q.W. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.

[16] Foster, D. P. and George, E. I. (1994), The risk inflation criterion for multiple regression, *Ann. Statist.*, **22**, 1947-1975.

[17] Gallant, A. R. amd Tauchen, G. (1997), Estimation of continuous time models for stock returns and interest rates. *Macroeconomic Dynamics*, **1**, 135-168.

[18] Gelfand, A. and Sahu, S. K. (1994), On Markov chain Monte Carlo acceleration, *J. Comput. Graph. Statist.*, **3**, 261-276.

[19] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian data analysis (second edition)*, Chapman & Hall/CRC, Boca Raton, FL.

[20] George, E. I. (2000), The variable selection problem, *J. Amer. Statist. Assoc.*, **95**, 1304-1308.

[21] George, E. I. and Foster, D. P. (2000), Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731-747.

[22] George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88**, 881-889.

[23] Girard, D. A. (1989), A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data, *Numer. Math.* **56(1)**, 1–23.

[24] Girard, D. A. (1991), Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization, *Ann. Statist.* **19(4)**, 1950–1963.

[25] Gong, J., Wahba, G., Johnson, D. and Tribbia, J. (1998), Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters, *Monthly Weather Review*, **126**, 210-231.

[26] Gu, C. (1990), Adaptive spline smoothing in non-Gaussian regression models, *J. Amer. Statist. Assoc.* **85**, 801–807.

[27] Gu, C. (2002), *Smoothing spline ANOVA models*. Springer-Verlag, New York.

[28] Hall, P. and Carroll, R.J. (1989), Variance function estimation in regression: the effect of the estimation of the mean. *J. Roy. Statist. Soc. B* **51**, 3-14.

[29] Hollingsworth, A. and Lönnberg, P. (1986), The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field, *Tellus*, **38A**, 111-136.

[30] Hudson, H. M. (1978), A Natural Identity for Exponential Families with Application in Multiparameter Estimation, *Ann. Statist.* **6**, 473-484.

[31] Johnston, I. and Silverman, B. W. (2002), Empirical Bayes selection of wavelet thresholds.

[32] Kimeldorf, G. and Wahba, G. (1971), Some Results on Techebycheffian Spline Functions, *J. Math. Anal. Applic.*, **33**, 82-95.

[33] Mallows, C. L. (1973), Some Comments on C_p, *Technometrics* **15**, 661–675.

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

[35] Jobson, J. D. and Fuller, W. A. (1980), Least Squares Estimation When the Covariance Matrix and Parameter Vector are Functionally Related. *J. Amer. Statist. Assoc.*, **75**, 176-181.

[36] Muller, H.G. and Stadtmuller, U. (1987), Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–625.

[37] O'Sullivan, F., Yandell, B. S. and Raynor, W. J., Jr. (1986), Automatic smoothing of regression functions in generalized linear models, *J. Amer. Statist. Assoc.* **81**, 96–103.

[38] Purser, J. and Parrish, D. (2003), A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation, *Meteor. Atmos. Physics*, **82**, 209-226.

[39] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1996), Bayesian model averaging for linear regression models, *J. Amer. Statist. Assoc.*, **92**, 179-191.

[40] Ruppert, D., Wand, M.P., Holst, U. and Hossjer, O. (1997), Local polynomial variance-function estimation. *Technometrics* **39**, 262-273.

[41] Shen, X. and Ye, J. (2002) Adaptive model selection, *J. Amer. Statist. Assoc.*, **97**, 210-221.

[42] Smith, M. S. (1996), Nonparametric regression – A Markov chain Monte Carlo approach, *unpublished Ph.D. thesis*, the Australian Graduate School Management at the University of New South Wales, Australia.

[43] Stadtmuller, V. and Tsybakov, A.B. (1995), Nonparametric recursive variance estimation. *Statistics* **27**, 55-63.

[44] Stone, M. (1974), Cross-validation and multinomial prediction, *Biometrika* **61**, 509–515.

[45] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.

[46] Vermunt, J. K. (1996), *Log-linear event history analysis*, **Series on Work and Organization**, Tilburg: Tilburg University Press.

[47] Wahba, G. (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. B*, **40(3)**, 364-372.

[48] Wahba, G. (1990), Spline models for observational data. *CBMS-NSF Regional Conference Series in Applied Mathematics, 59.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

[49] Wakefield, J. C., Gelfand, A. and Smith, A. F. M. (1992), Efficient generation of random variates via the ratio-of-uniforms method, *Statist. & Computing*, **1**, 129-133.

[50] Wei, B. C. (1998), *Exponential family nonlinear models*, **Lecture Notes in Statistics, 130**, Singapore: Springer-Verlag Singapore.

[51] Xiang, D. and Wahba, G. (1996), A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica*, **6**, 675–692.

[52] Yau, P. and Kohn, R. (2003), Estimation and variable selection in nonparametric heteroscedastic regression, *Statistics and Computing* **13**, 191 - 208.

[53] Ye, J. and Wong, W. H. (1997), Evaluation of Highly Complex Modeling Procedures with Binomial and Poisson Data, *Technical Report*, Department of Statistics, University of Chicago.

Appendix A

Proof of Theorem 4.1

Proof. We use the same notation as those used in Section 3. Under orthogonal design, (4.3.1) can be written as

$$\begin{aligned}
& C(Y)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{\frac{n-1}{2\pi\sigma^2}}\right)^{|\gamma|}\exp\left(-\frac{||Y-X_\gamma\beta_\gamma||^2+\lambda\sum_{i\in\gamma}|\beta_i|}{2\sigma^2}\right)d\beta_\gamma \\
= \; & C(Y)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{\frac{n-1}{2\pi\sigma^2}}\right)^{|\gamma|} \\
& \times\exp\left(-\frac{||X_\gamma u||^2-2\tilde{Y}_\gamma' X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta_i^*+u_i|-|\beta_i^*|\right)}{2\sigma^2}\right)du \\
& \times\exp\left(-\frac{\left|\left|Y-X_\gamma\beta_\gamma^*\right|\right|^2+\lambda\sum_{i\in\gamma}|\beta_i^*|}{2\sigma^2}\right)
\end{aligned}$$

where $\tilde{Y}_\gamma = Y - X_\gamma\beta_\gamma^*$. Denote

$$\begin{aligned}
Q \; \equiv \; & \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\sqrt{\frac{n-1}{2\pi\sigma^2}}\right)^{|\gamma|}\exp\left(-\frac{||X_\gamma u||^2-2\tilde{Y}_\gamma' X_\gamma u+\lambda\sum_{i\in\gamma}\left(|\beta_i^*+u_i|-|\beta_i^*|\right)}{2\sigma^2}\right)du \\
= \; & \prod_{i\in\gamma}\int_{-\infty}^{\infty}\sqrt{\frac{n-1}{2\pi\sigma^2}}\exp\left(-\frac{(n-1)u_i^2-2\tilde{Y}_\gamma' x_i u_i+\lambda\left(|\beta_i^*+u_i|-|\beta_i^*|\right)}{2\sigma^2}\right)du_i \\
\equiv \; & \prod_{i\in\gamma}Q_i. \qquad\qquad (A.0.1)
\end{aligned}$$

(i) Without loss of generality, suppose that $j \in \gamma$ and $\beta_j^* = 0$. Let γ^* be the submodel of γ with the j-th predictor variable excluded, then $\tilde{Y}_\gamma = \tilde{Y}_{\gamma^*}$

and $\beta^*_{\gamma^*,i} = \beta^*_{\gamma,i}$, $\forall i \in \gamma^*$. By (4.3.1) and (A.0.1), we have,

$$\frac{P(\gamma|Y)}{P(\gamma^*|Y)} = Q_j.$$

By (4.3.7) and (4.3.8),

$$\left|2\tilde{Y}' x_j\right| \leq \lambda.$$

This gives

$$\frac{P(\gamma|Y)}{P(\gamma^*|Y)} = Q_j < \int_{-\infty}^{\infty} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_j^2}{2\sigma^2}\right) du_j = 1.$$

The proof of (i) is now completed.

(ii) By (4.3.1) and (A.0.1), we only need to show that $Q_i \to 1$, $\forall i \in \gamma$. Without loss of generality, we assume that $sgn(\beta_i^*) > 0$. Similar to the derivation of (4.3.7), we have $2\tilde{Y}' x_i = \lambda$.

$$\begin{aligned}
Q_i &= \int_{-\infty}^{\infty} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2 - 2\tilde{Y}' x_i u_i + \lambda\left(|\beta_i^* + u_i| - |\beta_i^*|\right)}{2\sigma^2}\right) du_i \\
&= \int_{-\infty}^{\infty} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2 + \lambda\left(|\beta_i^* + u_i| - \beta_i^* - u_i\right)}{2\sigma^2}\right) du_i \\
&= \int_{-\beta_i^*}^{\infty} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i \\
&\quad + \int_{-\infty}^{-\beta_i^*} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2 - 2\lambda\left(\beta_i^* + u_i\right)}{2\sigma^2}\right) du_i \\
&= \Phi\left(\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) + \exp\left(\frac{\lambda^2/(n-1) + 2\lambda\beta_i^*}{2\sigma^2}\right) \Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right).
\end{aligned}$$

By the mean value theorem, for some ξ between $-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}$ and $-\frac{\beta_i^*}{\sigma/\sqrt{n-1}}$,

$$\Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right) - \Phi\left(-\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) = -\phi(\xi)\lambda/\sigma\sqrt{n-1} \to 0$$

given that $\lambda = o(n^{1/2})$. Thus,

$$Q_i \geq \Phi\left(\frac{\beta_i^*}{\sigma/\sqrt{n-1}}\right) + \Phi\left(-\frac{\beta_i^* + \lambda/(n-1)}{\sigma/\sqrt{n-1}}\right) \to 1.$$

On the other hand,

$$\begin{aligned} & \int_{-\infty}^{-\beta_i^*} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2 - 2\lambda\left(\beta_i^* + u_i\right)}{2\sigma^2}\right) du_i \\ \leq & \int_{-\infty}^{-\beta_i^*} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i. \end{aligned}$$

Therefore,

$$Q_i \leq \int_{-\infty}^{\infty} \sqrt{\frac{n-1}{2\pi\sigma^2}} \exp\left(-\frac{(n-1)u_i^2}{2\sigma^2}\right) du_i = 1.$$

Thus, $Q_i \to 1$ as $n \to \infty$. ∎

Appendix B

Proof of Proposition 4.1

Proof. The proposition follows from two observations on h.

(i) h is an decreasing function of γ. More specifically, if γ_1 is a submodel of γ_2, then $h(\gamma_1) \geq h(\gamma_2)$;

(ii) $h(\widehat{\gamma}) = h(\mathbf{1})$, where $\mathbf{1} = (1, ..., 1)$ stands for the full model.

Combining (i) and (ii), we see that for any regular model γ,

$$h(\widehat{\gamma}) = h(\mathbf{1}) \leq h(\gamma).$$

Now the proof is completed by the fact that $\widehat{\gamma}$ is a regular model. ∎