DEPARTMENT OF STATISTICS

University of Wisconsin

1210 West Dayton St.

Madison, WI 53706

TECHNICAL REPORT NO. 1094

July 12, 2004

# SOME PROBLEMS IN MODEL SELECTION [1]

Chenlei Leng

chenlei@stat.wisc.edu

http://www.stat.wisc.edu/~chenlei

[1]This research is partially supported by NSF Grants DMS 0072292, DMS 0134987 and NIH grant EYO9946.

# SOME PROBLEMS IN MODEL SELECTION

By

**Chenlei Leng**

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 dissertation consists of three parts: the first two parts are related to smoothing spline ANOVA models; the third part concerns the Lasso and its related procedures in model selection.

In Part I, by adopting the Cox proportional hazard model to quantify the hazard function, we propose a novel nonparametric model selection technique to analyze time to event data, within the framework of smoothing spline ANOVA models. Instead of the usual squared norms in the traditional smoothing spline ANOVA, our method employs a regularization with the penalty functional being the sum of the component norms. This method shrinks functional components and produces some components that are exactly zeros. It is an extension of the "COSSO" proposal by Lin and Zhang (2002) in Gaussian regression to hazard regression. To compute the estimate when the smoothing parameter is fixed, we develop an efficient algorithm based on a reformulation of the penalized partial likelihood. Approximations to the leave-out-one likelihood cross validation score are derived to choose the smoothing parameters. Both simulations and real examples suggest that our proposal is very powerful for model selection and component estimation.

Part II of the thesis concerns penalized likelihood density estimation. We introduce a randomized Generalized Approximate Cross Validation score to

estimate the smoothing parameters. A three dimensional example illustrates the usefulness of the method.

Part III studies the consistency of several recent linear model selection proposals. The Lasso, the Forward Stagewise regression and the Lars are closely related procedures recently proposed for linear regression problems. Each of them can produce sparse models and can be used both for estimation and variable selection. We show, however, that the dual goal of accurate estimation and consistent variable selection can not be achieved simultaneously: when the tuning parameter is chosen to minimize the prediction error, as is commonly done in practice, in general these procedures are not consistent in terms of variable selection. That is, the sets of variable selected by the procedures are not consistent at finding the true set of important variables. In particular, we show that for any sample size $n$, when there are superfluous variables in the linear regression model and the design matrix is orthogonal, the probability of the procedures correctly identifying the true set of important variables is less than a constant (smaller than one) not depending on $n$.

# Acknowledgements

I am most grateful for the guidance of my advisor Professor Grace Wahba. Being a great researcher herself, Grace inspired my interest in studying nonparametric statistics and gave me the opportunity to conduct further research. Moreover, she provided an encouraging and critical atmosphere for my research and the development of this dissertation. It was a great pleasure for me to work under her supervision. I would also like to thank Professor Yi Lin who shared his many ideas and provided constructive comments. This dissertation would be still in infancy without advice and help from Grace and Yi Lin.

I thank Professors Kjell Doksum, Chunming Zhang and Robert Nowak, for their service on my defense committee.

I enjoyed meeting with the thursday spline meeting group, led by Grace and Yi Lin. I want to thank them for all their help, interest and valuable hints. Especially, I am obliged to Hao Helen Zhang for her collaboration on the first part of this thesis. Being involved in such an active research group is itself fun and rewarding.

I must not forget to thank the fellow students, the professors and the staff members in the statistics department. During my stay in Madison, they helped me in various ways and made my life and study enjoyable.

I owe so much to my parents and my sister, who have always been supportive

of my pursuing an advanced degree in science. Last but not least, I thank my wife Hui for giving me continuous encouragements and the most patient love. I dedicate this work to her.

This research is partially supported by NSF Grants DMS 0072292, DMS 0134987 and NIH grant EYO9946.

# List of Figures

# List of Tables

# Contents

# Chapter 1

# Introduction for Part I

*"Everything should be made as simple as possible, but not simpler."*

Albert Einstein

## 1.1 Motivation

The first part of this dissertation focuses on developing an automatic model selection and model estimation procedure for survival data.

The problem of analyzing time to event data arises in a number of applied fields, such as biology, engineering, economics and epidemiology. A distinct feature of such data sets is that they usually contain censored observations. Censoring occurs when an object's life length, or survival time, is known to occur in a certain period of time, but the exact time is unknown. The main interest in time to event data analysis is to study the dependence of the survival time on some explanatory variables. Without prior knowledge what variables may contribute to the survival time, it is a common practice for scientists to collect many covariates at the beginning of such studies. The critical tasks are to identify the subset of the important covariates and to assess their effects on

the survival time. The first objective is often referred to as model selection or variable selection, and the second is estimation.

Model selection serves to reduce the dimensionality of the covariates and is related to the parsimony of the model. For a given set of observations, there exist infinite many possible models explaining the data. Simpler models are usually preferred for the sake of interpretibility and scientific insight. It coincides with the so called "Occam's razor" principle which states the simplest model is more likely to be correct. This property is especially important in medical studies, since the investigators are usually interested in detecting variables which can explain certain outcomes. By excluding noisy or irrelevant covariates, model selection provides a better understanding of the data generating mechanism. And identifying risk factors can greatly facilitate the goal of further scientific investigation of the important variables.

The other goodness of measure is in terms of prediction performance. It is desirable that the estimator follows closely with empirical evidence and generalizes well for future observations. Intuitively, better prediction could be obtained by pruning out the superfluous variables which do not contribute to prediction.

A vast majority of the literature study the model selection problem in the context of multivariate linear models, in which the underlying multivariate function assumes a parametric form known *a priori*. Such simplification often leads to simple and interpretable models. Furthermore, it is relatively easy to develop inference procedures for parametric models. However, since parametric models

rely heavily on the model assumption, misspecification of the model can lead to misleading results in real practice. As a result, linear modeling procedures should be exercised with caution unless such claim is supported by strong empirical evidence. An immediate advantage of nonlinear models is that little prior information on model structure is needed, and in turn, it may offer insight to more appropriate forms for parametric modeling. Thanks to cheap and reliable computing power, nonparametric methods have emerged as effective alternatives to parametric modeling in a wide variety of statistical problems. This part of the dissertation aims to develop an automatic nonparametric model selection method for survival data where censoring occurs.

## 1.2 Literature Review

The problem of model selection has drawn the attention of applied and theoretical statisticians for a long time. Parametric modeling techniques are usually used at the beginning of such analysis since they are well studied. Many linear model selection methods utilize the maximum likelihood method while penalizing the number of non-zero components. Popular penalties include AIC, BIC and Mallow's $C_p$. Breiman (1995) showed that such traditional linear model selection algorithms, which include forward, backward and best subset selection, suffer from instability and relatively lack of accuracy. In an attempt to stabilize the estimate and to improve prediction accuracy, Breiman (1995) proposed a nonnegative garrote variable selection criterion by constrained optimization,

after an ordinary linear model is fitted. Tibshirani (1996) proposed a linear model selection method called the Lasso (Least Absolute Shrinkage and Selection Operator). The Lasso employs an $L_1$ type of penalty on the regression coefficients which tends to produce sparse models. Its application in survival analysis is studied in Tibshirani (1997). It has been shown that the Lasso has very good performance in terms of model selection and prediction accuracy. Fan and Li (2001, 2002) proposed a variant of the Lasso called the SCAD (Smoothly Clipped Absolute Deviation). It was shown that the SCAD enjoys the oracle property, namely, the true regression coefficients are estimated as if correct sub-model is known. Despite the attractiveness of linear model selection tools, parametric methods are useful only in the setting of standard linear models, where a parametric form is known *a priori*. In practice, some of the covariates could act nonlinearly. Linear modeling for such data can incur a large bias. One way to overcome this challenge is to use a large number of parametric terms for such covariates. Without prior knowledge, however, it is not clear how to specify the parametric terms in advance. A nice alternative is to let the data find nonlinear features, which is the main advantage of nonlinear methods.

The Cox's proportional hazard model is among the most popular techniques in modeling survival data since its introduction by Cox (1972, 1975) . A thorough exposure to this topic can be found in Kalbfleisch and Prentice (2002),

Klein and Moeschberger (1997), among others. The hazard assumes a semi-parametric model where the baseline hazard is totally unspecified and a parametric form is hypothesized for the covariates. In this model, the covariates act multiplicatively on the hazard function. The partial likelihood formulation is the standard technique for parameter estimation and statistical inference. To avoid possible model misspecification in a parametric setup, an alternative approach is to allow the covariate effects to vary in a high-dimensional function space, leading to various nonparametric and semi-parametric estimation methods. A popular choice in the nonparametric modeling is via the minimization of a penalized likelihood. O'Sullivan (1998) and Gu (1996) studied penalized partial likelihood estimation. An upper bound on the rate of convergence is given by O'Sullivan (1993). Zucker and Karr (1990) considered a generalization of the proportional hazard model which allows time varying coefficients. For model selection purposes, the existing nonparametric techniques are usually hypothesis testing and heuristic search type of methods. Hastie and Tibshirani (1990) section 9.4 considered several nonlinear model selection methods in the same spirit as the stepwise selection, where the familiar additive models are entertained. Gray (1992, 1994) used splines with fixed degrees of freedom as an exploratory tool to assess the effect of covariates. The actual model selection is dealt with using hypothesis testing. Kooperberg, Stone and Truong (1995) employed a heuristic search algorithm using polynomial splines to model the

hazard function. Additionally, Wood *et al.* (2002) derived a Bayesian formulation and the corresponding Gibbs sampler to select a model in nonparametric spline regression. Although posterior information can be used for model selection, the authors commented that the Bayesian method in the paper can only be efficiently implemented for up to 5 to 6 variables.

The Reproducing Kernel Hilbert Space (RKHS) theory provides an elegant framework to study functional approximation and conduct statistical analysis. A thorough exposure on RKHS can be found in Aronszajn (1950). Smoothing spline analysis of variance models (SS-ANOVA), a modeling procedure based on RKHS theory, is a popular choice to investigate functional relations. SS-ANOVA models are widely used for Gaussian regression, generalized regression, density estimation and hazard rate regression, see, for example, Wahba (1990), Wahba *et al.* (1995), Lin *et al.* (2000), Gao *et al.* (2001) and Gu (2002). For model selection purposes, Zhang *et al.* (2002) proposed a likelihood basis pursuit method. In this proposal, each nonparametric component is expanded as a linear combination of basis functions. They employ an $L_1$ penalty on the coefficients to encourage sparsity in the estimated coefficients. This gives a possible generalization of the Lasso method to nonparametric regression. However, a separate model selection has to be applied after model fitting, since sparsity in coefficients does not guarantee the sparsity in components. A sequential Monte Carlo bootstrap test algorithm is developed for model selection purpose. Recently, Lin and Zhang (2002) proposed an automatic nonparametric

model selection for Gaussian observations in the context of SS-ANOVA models, which is called the Cosso (COmponent Selection and Smoothing Operator). By imposing a penalty on the sum of the component norms instead of the usual squared norms, the method does automatic model selection and component estimation. The Cosso has very good properties in terms of selecting the right components and estimating. It is shown that the Cosso can model reasonably high dimensional data. A closely related procedure is proposed by Gunn and Kandola (2002) in the machine learning literature.

In this dissertation, a unified framework for model selection and model estimation is developed for time to event data analysis. The method is an extension of the Cosso methodology for Gaussian data to survival data.

## 1.3 Outline of Part I

This part considers model selection for Cox's proportional hazard models. The method is a natural extension of the Gaussian Cosso proposal to survival data analysis. The problem is formulated by using penalized partial likelihood with the penalty being the sum of the norms, while the usual practice is to use the sum of the squared norms as the penalty. We show that using the sum of the norms has the advantage in generating exact zero estimates for some components. We derive a prediction based cross validation criterion to facilitate adaptive selection of the smoothing parameter, which governs the fit to the data and the roughness of the estimate.

The part is organized as follows. Chapter 2 derives the partial likelihood and reviews SS-ANOVA models. In Chapter 3, the model selection problem is formulated via the method of regularization, where a penalty on the sum of the component norms is imposed on the partial likelihood. The formulation is a natural extension of the Gaussian Cosso proposal to the context of survival analysis. We show that the Cosso can be formulated as a traditional SS-ANOVA problem with added constraints on some parameters. The connection enables us to use the existing fitting algorithm for the traditional SS-ANOVA models. We then derive an efficient algorithm to compute the estimate when the smoothing parameter is fixed. In Chapter 4, we further develop computationally efficient tuning criteria to choose the smoothing parameter by approximating the leave-out-one likelihood cross validation score. Our proposal involves no further computation once the estimate is obtained. We demonstrate the usefulness of our method via some simulations in Chapter 5. The method is applied to several real data sets in Chapter 6, including the lung cancer data, primary biliary cirrhosis data and mouse leukemia data. Chapter 7 of this dissertation gives conclusion remarks and future direction of research.

# Chapter 2

# Nonparametric Model Selection

In this chapter, we review the partial likelihood method for the estimation of the baseline hazard function. We then discuss the decomposition of the relative risk function and introduce the corresponding Reproducing Kernel Hilbert Space.

## 2.1 The Partial Likelihood

To fix notations, suppose that there are $n$ independent objects under study. Let $T,\ C,\ x$ be respectively survival time, censoring time and their associated covariates, where $x = (x^{(1)}, ..., x^{(d)})^T$ is a d-dimensional vector in the domain $\mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(d)}$, also known as explanatory variables or confounders. $x$ may contain some continuous variables and some categorical variables. Without loss of generality, it is assumed that each continuous covariate is in the range of $[0, 1)$, otherwise each covariate is scaled to the interval $[0, 1)$. Our data consist of the triple $(Z_i, \delta_i, x_i),\ i = 1, ...n$, where $Z_i = \min\{T_i, C_i\}$ is the time on study for the $i$th subject and $\delta_i = I_{(T_i \leq C_i)}$ is the censoring indicator ($\delta_i = 1$ if the event has occurred and $\delta_i = 0$ if the lifetime is right censored). $x_i = (x_i^{(1)}, ..., x_i^{(d)})^T$ is the vector of covariates associated with the $i$th subject which may affect the survival

distribution of $T$. It is assumed that $T$ and $C$ are conditionally independent given $x$ and that the censoring mechanism is uninformative. We are interested in studying the dependence of the survival time on the covariates.

To facilitate discussions to follow, let $f(t|x), S(t|x)$ and $h(t|x)$ be respectively the conditional density function, the conditional survival function and the conditional hazard function. It is easy to see that the full likelihood is

$$\prod_{i=1}^{n} f(Z_i|x_i)^{\delta_i} S(Z_i|x_i)^{1-\delta_i} = \prod_{i=1}^{n} h(Z_i|x_i)^{\delta_i} \prod_{i=1}^{n} S(Z_i|x_i).$$

Without loss of generality, we assume that there are no ties in the observed failure time. Presence of ties is dealt with using the technique in Breslow (1974). Let $t_1^0 < ... < t_N^0$ be the ordered observed failure time. Denote $(j)$ as the label for the item falling at $t_j^0$ so the covariates associated with the $N$ failures are $x_{(1)}, ..., x_{(N)}$. Let $R_j$ be the risk set right before the time $t_j^0$:

$$R_j = \{i : Z_i \geq t_j^0\}.$$

For the family of proportional hazard models due to Cox (1972), the conditional hazard rate of an individual with covariate $x$ is

$$h(t|x) = h_0(t) \exp\{\eta(x)\},$$

where $h_0(t)$ is an arbitrary baseline hazard function and $\eta(x)$ is the logarithm of the relative risk function. Note that in the above expression for $h(t|x)$, if one multiplies $h_0(t)$ by a constant and divides $\exp\{\eta(x)\}$ by the same constant, $h(t|x)$ does not change. In order for $\eta(x)$ to be identifiable, we assume $\int \eta = 0$.

Cox models are also called multiplicative hazard models, since the hazard rates of two subjects with distinct covariates are proportional.

In the parametric Cox model, it is assumed $\eta(x) = x\beta$, where $\beta = (\beta_1, ..., \beta_d)^T$ is the parameter vector. In this dissertation, we relax the parametric assumption and consider the situation where $\eta(x)$ is a nonparametric function of $x$. Our major concern is to identify and study the structure of $\eta(x)$.

The log likelihood can be written as

$$\sum_{i=1}^{n} \{\delta_i [\log h_0(Z_i) + \eta(x_i)] - H_0(Z_i) \exp[\eta(x_i)]\}, \tag{2.1.1}$$

where $H_0(t)$ is the cumulative baseline hazard function.

Following Fan and Li (2002), p 79-80 and Breslow's idea, denote the cumulative hazard function as a piecewise constant function with possible jumps at the observed failure times, that is $H_0(t) = \sum_{j=1}^{N} h_j I_{[t_j^0 \le t]}$. Then

$$H_0(Z_i) = \sum_{j=1}^{N} h_j I_{i \in R_j}.$$

Substituting the cumulative baseline hazard into (2.1.1), one obtains

$$\sum_{j=1}^{N} \log h_j + \sum_{j=1}^{n} \delta_j \eta(x_j) - \sum_{i=1}^{n} [\exp(\eta(x_i)) \sum_{j=1}^{N} h_j I_{i \in R_j}]. \tag{2.1.2}$$

Maximum likelihood method is used to estimate $h_j$'s. Taking derivative with respect to $h_j$ and solving for $h_j$ in (2.1.2), one gets

$$\hat{h}_j = \{\sum_{i \in R_j} \exp(\eta(x_i))\}^{-1}.$$

Plugging $\hat{h}_j$'s back in (2.1.2) and dropping a constant $-N$, we get the so-called partial likelihood

$$\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\}. \tag{2.1.3}$$

The partial likelihood is treated as a usual likelihood for the purpose of parameter estimation and statistical inference.

## 2.2 The Model Selection Problem

From a mathematical perspective, estimating the multivariate function $\eta$ is a functional approximation problem. An important aspect of statistical modeling, which distinguishes it from mere functional approximation, is the interpretibility of the results. It is of great interest to decompose multivariate functions in a way similar to the classical analysis of variance (ANOVA).

As the first step to approximate the relative risk function, $\eta(x)$ can be estimated by a linear combination of parametric terms,

$$\eta(x) = \sum_{i=1}^{p}\beta_i h_i(x) = \beta^T \mathbf{h}(x),$$

where $\mathbf{h}(x) = [h_1(x), ..., h_p(x)]^T$ is a vector of $p$ fixed functions of input. In the simplest linear regression case, $\mathbf{h}(x) = [x^{(1)}, ..., x^{(d)}]$ with $p = d$. Maximum likelihood method can be used to estimate $\beta_i$'s. For model selection purposes, one can use best subset, forward selection and backward selection to eliminate some of the coefficients. The criteria such as AIC and BIC can be employed to control the number of non-zero parameters. A different variable selection for

linear model selection is the shrinkage method, which includes the nonnegative garrote (Breiman 1995), the Lasso (Tibshirani 1996) and the SCAD (Fan and Li, 2001). A distinct feature of these methods is that a penalized likelihood score is minimized, where the penalty functions usually encourage sparsity in the estimated coefficients.

Although sometimes linear models provide adequate fit, it is not always appropriate to make linear or even quadratic or cubic assumptions. When the linear assumption is far from the truth, the estimate under such assumptions may be very misleading. In response to this, people turn to more flexible nonlinear models. Recent years have witnessed unprecedented advance in nonlinear modeling. A wide variety of techniques have been proposed to allow data adaptive fitting.

Similar to the classical ANOVA in designed experiments, the $d$ dimensional function $\eta$ can be decomposed as

$$\eta(x) = \eta_0 + \sum_{j=1}^{d} \eta_j(x^{(j)}) + \sum_{j<k} \eta_{j,k}(x^{(j)}, x^{(k)}) + ... + \eta_{1,...,d}(x^{(1)}, ..., x^{(d)}),$$

where $\eta_0$ is a constant, $\eta_j$'s are the main effects, and the $\eta_{j,k}$'s are the two way interactions and so on. The identifiability condition is assured by certain side conditions on the $\eta_j$'s, $\eta_{j,k}$'s and so on. Higher order terms in this decomposition are often excluded to control the model complexity. The truncated series is written as

$$\eta(x) = \eta_0 + \sum_{\alpha=1}^{p} \eta_\alpha(x). \tag{2.2.1}$$

Excluding all the interactions yields the familiar additive models which have been studied by Hastie and Tibshirani (1990). Including two way interactions and lower order terms yields the two way interaction model

$$\eta(x) = \eta_0 + \sum_{j=1}^{d} \eta_j(x^{(j)}) + \sum_{j<k} \eta_{j,k}(x^{(j)}, x^{(k)}).$$

Lower order approximations contribute to faster convergence rates, as investigated by Lin (2000). The family of the low dimensional ANOVA decompositions represents a nonparametric compromise in an attempt to overcome the "curse of dimensionality", since estimating a more general function $\eta(x^{(1)}, ..., x^{(d)})$ requires very large data sets for even moderate $d$.

The model selection problem in the functional ANOVA setup is to identify a suitable subset of $\{\eta_\alpha\}$'s which are important for the sake of estimating survival time.

## 2.3 Smoothing Spline ANOVA Models

The idea of the method of regularization is to minimize a penalized partial likelihood criterion

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \tau J(\eta), \tag{2.3.1}$$

where $J(\eta)$ is a roughness penalty. The parameter $\tau$ controls the smoothness of the estimator. As $\tau$ increases, the estimated log relative risk function is forced toward a function which lies in the null space of $J(\eta)$, i.e., the estimator goes to $\eta_0$ satisfying $J(\eta_0) = 0$.

In SS-ANOVA models, the estimate is associated with a metric space such that penalized likelihood score is continuous in $\eta$. More precisely, it is assumed that $\eta$ lies in a reproducing kernel Hilbert space (RKHS) corresponding to the decomposition (2.2.1). An RKHS $\mathcal{H}$ is a Hilbert space of functions on a domain with all the evaluation functionals $t : \eta \to \eta(t)$ bounded. An RKHS possesses a reproducing kernel $R(\cdot, \cdot)$, a non-negative definite function satisfying $(R(t, \cdot), f(\cdot)) = f(t), \forall f \in \mathcal{H}$, where $(\cdot, \cdot)$ is an inner product in $\mathcal{H}$. For a thorough exposure to RKHS, see Aronszajn (1950) and Wahba (1990) for details.

For the function space over $X^{(j)}$ on $[0, 1]$, we use the second order Sobolev Hilbert space, namely,

$$W^{(j)}[0,1] = \{f : f(x^{(j)}), f'(x^{(j)}) \text{ are absolutely continuous and } f''(x^{(j)}) \in L_2[0,1]\}.$$

When endowed with the inner product

$$(f,g)_{W^{(j)}} = \{\int_0^1 f(t)dt\}\{\int_0^1 g(t)dt\} + \{\int_0^1 f'(t)dt\}\{\int_0^1 g'(t)dt\} + \int_0^1 f''(t)g''(t)dt,$$

$W^{(j)}$ is an RKHS with a reproducing kernel

$$K^{(j)}(s,t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|).$$

Here

$$k_1(s) = s - 0.5,$$

$$k_2(s) = [k_1^2(s) - 1/12]/2,$$

$$k_4(s) = [k_1^4(s) - k_1^2(s)/2 + 7/240]/24,$$

for $s, t \in \mathcal{X}^{(j)}$. This is a special case of equation (10.2.4) in Wahba (1990) with $m = 2$.

$W^{(j)}$ can be decomposed into the direct sum of two orthogonal subspaces as $W^{(j)} = 1^{(j)} \oplus W_1^{(j)}$, where $1^{(j)}$ is the "mean" space and $W_1^{(j)}$ is the "contrast" space generated by the kernel

$$K_1^{(j)}(s,t) = k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|).$$

For a categorical variable $X^{(i)}$ on the discrete domain $\mathcal{X}^{(i)} = \{1, ..., D\}$, a function is a vector of length $D$ and evaluation is simply coordinate extraction. We decompose $W^{(i)}$ as $1^{(i)} \oplus W_1^{(i)}$, where

$$1^{(i)} = \{\eta : \eta(1) = \cdots \eta(D)\}$$

and

$$W_1^{(i)} = \{\eta : \sum_{j=1}^{D} \eta(j) = 0\}$$

is generated by the reproducing kernel

$$K_1^{(i)}(s,t) = DI_{(s=t)} - 1, \;\; s, t \in \{1, ..., D\}.$$

This kernel defines a shrinkage estimate being shrunk towards the mean, as discussed in Gu (2002) section 2.2.

A convenient approach to the construction of RKHS on a product domain $\bigotimes_{\alpha=1}^{d} \mathcal{X}^{(\alpha)}$ is by taking the tensor product of spaces constructed on the marginal

domain $\mathcal{X}^{(\alpha)}$. The tensor product has the form

$$\begin{aligned}\bigotimes_{j=1}^{d} W^{(j)} &= \bigotimes[\{1^{(j)}\} \bigoplus \{W_1^{(j)}\}] \\ &= [\{1\}] \bigoplus \sum_{j=1}^{d} [\{W_1^{(j)}\}] \bigoplus \sum_{j<k} [\{W_1^{(j)}\} \bigotimes \{W_1^{(k)}\}] + ....,\end{aligned} \tag{2.3.2}$$

where the dependence of the additive components on $\{1^{(j)}\}$ is suppressed, that is,

$$\{1\} = \bigotimes_{j=1}^{d} \{1^{(j)}\},$$

$$\{W_1^{(j)}\} = \{1^{(1)}\} \bigotimes \cdots \bigotimes \{1^{(j-1)}\} \bigotimes \{W_1^{(j)}\} \bigotimes \{1^{(j+1)}\} \cdots \bigotimes \{1^{(d)}\},$$

and so on. Each functional component in the decomposition (2.2.1) falls in the corresponding subspace of $\bigotimes_{j=1}^{d} W^{(j)}$. Therefore, the tensor sum in (2.3.2) is truncated using the same functional ANOVA argument. We will denote the truncated series as

$$\mathcal{H} = \{1\} \bigoplus \mathcal{H}_1 = \{1\} \bigoplus \sum_{\alpha=1}^{p} \{\mathcal{W}^{\alpha}\}. \tag{2.3.3}$$

To obtain additive spline models, one retains the mean space $\{1\}$ and only those subspaces of the form $\{W_1^{(j)}\}$. Retaining all $[\{W_1^{(j)}\} \bigotimes \{W_1^{(k)}\}]$ and lower order terms yield two way interaction models. The reproducing kernel of $\bigotimes \mathcal{X}^{(j)}$ is

the product of the $d$ reproducing kernels on the marginal domains

$$\begin{aligned}
&\prod_{j=1}^{d}\{1+K_1^{(j)}(s^{(j)},t^{(j)})\}\\
=&1+\sum_{j=1}^{d}K_1^{(j)}(s^{(j)},t^{(j)})+\sum_{j<k}K_1^{(j)}(s^{(j)},t^{(j)})K_1^{(k)}(s^{(k)},t^{(k)})+...\\
&+\prod_{j=1}^{d}K_1^{(j)}(s^{(j)},t^{(j)}),
\end{aligned}$$

where $s,t \in \bigotimes \mathcal{X}^{(\alpha)}$. Using the property that the reproducing kernel for a product term is the product of individual reproducing kernels, we can obtain a truncated expansion of the product of the reproducing kernels, which corresponds to the functional decomposition of $\eta(x)$. With some abuse of the notation, we use $K_\alpha(s,t)$ as the reproducing kernel for the term $\eta_\alpha$ in the functional decomposition of $\eta$, where $K_\alpha$ applies to the corresponding coordinate of $s$ and $t$, and $\eta_\alpha$ can be any main effect, or any interaction term. For example, for the term $\eta_{i,j}$, its reproducing kernel is simply $K_\alpha(s,t)=K_1^{(i)}(s^{(i)},t^{(i)})K_1^{(j)}(s^{(j)},t^{(j)})$. Thus, the reproducing kernel corresponding to $\mathcal{H}$ is conveniently written as $K=1+\sum_{\alpha=1}^{p}K_\alpha$.

To further encompass the linear model, we may make a further orthogonal decomposition of $W_1^{(j)}$ into parametric and nonparametric terms. This issue is not explored in this article since our emphasis is on the selection of functional components of the SS-ANONA.

# Chapter 3

# Model Formulation

This chapter further studies the model selection problem in SS-ANOVA. A Cosso type of penalty is introduced for this purpose. We show that the minimization problem can be reformulated as a usual SS-ANOVA model with added constraints on a set of parameters. We present a one step update algorithm to compute the estimate while the smoothing parameter is fixed. The important issue of choosing the smoothing parameter will be discussed in the next chapter.

## 3.1 The Cosso Estimate in Survival Analysis

The usual smoothing spline ANOVA penalizes the squared norm of each component, namely

$$J(\eta) = \sum_{\alpha=1}^{p} \theta_\alpha^{-1} ||P^\alpha \eta||^2,$$

where $P^\alpha \eta$ is the projection of $\eta$ onto $W^\alpha$ and $\theta_\alpha$'s are nonnegative smoothing parameters. With high dimensional covariates, fitting a model with $p$ smoothing parameters is computationally intensive but not infeasible. The algorithm of Gu and Wahba (1991) has been used to optimally tune the smoothing parameters via multi dimensional minimization. However, since the algorithm operates on $\tau$

and $\log(\theta_\alpha)$, none of the functional components is estimated as zero. Therefore, ad hoc variable selection technique has to be applied after the estimation. Gu (1992) introduced some geometric diagnostics for the identifiability and the practical significance of the fitted terms. The Cosso, however, does simultaneous model selection and model estimation.

The Cosso extension to survival data aims at minimizing a penalized partial likelihood score

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \tau J(\eta) \text{ with } J(\eta) = \sum_{\alpha=1}^{p}\|P^\alpha\eta\|. \quad (3.1.1)$$

The penalty functional is a sum of RKHS norms instead of the squared RKHS norm penalty. The penalty term $J(\eta)$ used in the Cosso is not a norm. However, it is a pseudo-norm in the following sense: for any $f, g \in \mathcal{H}$, $J(f) \geq 0$, $J(cf) = |c|J(f)$ and $J(f+g) \leq J(f) + J(g)$. The Lasso in linear cases can be seen as a special case of the Cosso. For the input space $[0,1]^d$, consider the linear function space $\{1\} \oplus \{x^{(1)} - 1/2\} \oplus ... \oplus \{x^{(d)} - 1/2\}$, with the usual $L^2$ inner product $(f,g) = \int fg$. The penalty in the Cosso becomes $J(\eta) = (12)^{-1/2}\sum_{j=1}^{d}|\beta_j|$ for a linear estimator $\eta(x) = \beta_0 + \sum_{j=1}^{d}\beta_j x^{(j)}$. This is equivalent to the $L_1$ penalty on the linear coefficients, which leads to the Lasso estimator.

The difference between the form of the Cosso and the usual smoothing spline mirrors the difference between the Lasso and the ridge regression. The Lasso estimate shrinks some of the components to be exactly zero, while the ridge regression shrinks every component but never produces zero coefficients. The

correspondence suggests that the Cosso is a possible variable selection procedure.

## 3.2 An Equivalent Formulation

Although the solution to (3.1.1) is posed in an infinite-dimensional space, the minimizer $\hat{\eta}$ is finite-dimensional, as shown in the following lemma.

**Lemma 3.2.1.** *Denote $\hat{\eta} = \hat{b} + \sum_{\alpha=1}^{p} \hat{\eta}_\alpha$ as the minimizer of (3.1.1) in (2.3.3), with $\hat{\eta}_\alpha \in \mathcal{W}^\alpha$. Then $\hat{\eta}_\alpha \in span\{K_\alpha(x_i, \cdot), i = 1, ..., n\}$, where $K_\alpha(\cdot, \cdot)$ is the reproducing kernel of the space $\mathcal{W}^\alpha$.*

*Proof.* For any $\eta \in \mathcal{H}$, write it as $\eta = b + \sum_{\alpha=1}^{p} \eta_\alpha$ with $\eta_\alpha \in \mathcal{W}^\alpha$. Denote the projection of $\eta_\alpha$ onto $\text{span}\{K_\alpha(x_i, \cdot), i = 1, ..., n\} \subset \mathcal{W}^\alpha$ as $\pi_\alpha$, and the orthogonal complement as $\omega_\alpha$. Then $\eta_\alpha = \pi_\alpha + \omega_\alpha$, and $\|\eta_\alpha\|^2 = \|\pi_\alpha\|^2 + \|\omega_\alpha\|^2$. By orthogonality, $\omega_\alpha(x_i) = (K_\alpha(x_i, \cdot), \omega_\alpha(\cdot)) = 0$. So

$$\eta(x_i) = (1 + \sum_{\alpha=1}^{p} K_\alpha(x_i, \cdot), b + \sum_{\alpha=1}^{p} (\pi_\alpha + \omega_\alpha)) = b + \sum_{\alpha=1}^{p} (K_\alpha(x_i, \cdot), \pi_\alpha).$$

Therefore, we can write (3.1.1) as

$$-\frac{1}{n} \sum_{i=1}^{N} \{b + \sum_{\alpha=1}^{p} (K_\alpha(x_{(i)}, \cdot), \pi_\alpha) - \log[\sum_{j \in R_i} \exp(b + \sum_{\alpha=1}^{p} (K_\alpha(x_j, \cdot), \pi_\alpha)]\} \\ + \tau(\|\pi_\alpha\|^2 + \|\omega_\alpha\|^2\|)^{1/2}. \tag{3.2.1}$$

We immdediately see that any minimizing $\eta$ satisfies $\omega_\alpha = 0,\ \alpha = 1, ..., p$. The conclusion of the lemma follows. □

The Cosso problem (3.1.1) is very hard to optimize due to the sum of the RKHS norms. Lin and Zhang (2002) gave an equivalent formulation which is to minimize

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \lambda_0\sum_{\alpha=1}^{p}\theta_\alpha^{-1}\|P^\alpha\eta\|^2\} + \lambda\sum_{\alpha=1}^{p}\theta_\alpha \qquad (3.2.2)$$
$$\text{subject to } \ \theta_\alpha \geq 0, \ \alpha = 1, ..., p,$$

where $\lambda_0$ is a fixed parameter and $\lambda$ is the smoothing parameter.

After introducing another set of parameters $\theta = (\theta_1, ..., \theta_p)^T$, the Cosso formulation has the same form as the usual smoothing spline ANOVA setup except that the sum of $\theta_\alpha$'s is penalized.

We remark that the additional parametrization on $\theta$ makes it possible to estimate some $\theta_\alpha$'s to be zeros, leading to zero components in the Cosso estimate. In addition, tuning is much easier since only one smoothing parameter $\lambda$ has to be chosen, compared to multiple smoothing parameters in the usual SS-ANOVA models.

An alternative formulation to (3.2.2) is through a contrained penalized partial likelihood problem, which minimizes

$$-\frac{1}{n}\sum_{i=1}^{N}\{\eta(x_{(i)}) - \log[\sum_{j\in R_i}\exp(\eta(x_j))]\} + \lambda_0\sum_{\alpha=1}^{p}\theta_\alpha^{-1}\|P^\alpha\eta\|^2\} \qquad (3.2.3)$$
$$\text{subject to } \ \sum_{\alpha=1}^{p}\theta_\alpha \leq M, \ \theta_\alpha \geq 0, \ \alpha = 1, ..., p.$$

Here $M$ assumes the role of the smoothing parameter $\lambda$ in (3.2.2).

## 3.3 The Form of the Solution

For any fixed $\theta$, the Cosso is equivalent to the usual smoothing spline. It is well known that the soluition has the form

$$\eta(x) = b + \sum_{i=1}^{n} K_\theta(x, x_i)c_i,$$

where $K_\theta = \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha$ and we set $b = 0$ for $\eta$ to be identifiable.

Let $\sum_{i=1}^{n} K_\alpha(x, x_i)c_i$ be $G_\alpha(x)$. A componentwise form of the estimate is simply

$$\eta(x) = \sum_{\alpha=1}^{p} \theta_\alpha G_\alpha(x).$$

$\theta_\alpha = 0$ implies that the corresponding component estimator is zero.

The exact solution to (3.1.1) has a form $\eta(x) = \sum_{i=1}^{n} \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_i)c_i$. In the case where $n$ is large, one way to reduce the computational load is to use the parsimonious approach as suggested by Xiang and Wahba (1996), Ruppert and Carroll (2000), Lin *et al.* (2000). In this approach, a proper subset $\{x_{1*}, ..., x_{m*}\}$ $(m \leq n)$ of $\{x_1, ..., x_n\}$ is used as the basis functions and the corresponding approximate solution is

$$\eta(x) = \sum_{i=1}^{m} \sum_{\alpha=1}^{p} \theta_\alpha K_\alpha(x, x_{i*})c_i.$$

It has been shown that there is only a little sacrifice in accuracy of the estimation by using a proper subset.

In our implementation, the simple random sampling scheme is used to choose the basis functions. Kim and Gu (2004) provides some empirical justification

of the efficacy of the random sampling technique in the Gaussian regression. Other sampling schemes, such as cluster sampling in Xiang and Wahba (1996), can be used for the purpose of sub-sampling. Better sampling techniques may improve the approximation accuracy of the estimate.

It is possible to simultaneously minimize the objective function in (3.2.3) with respect to both $\theta$ and $c$, however, the optimization is nonlinear and complex. It is noticed that it is much easier to optimize the objective function with respect to one set of variables ($\theta$ or $c$) when the other set ($c$ or $\theta$) is fixed. Existing algorithms for fitting the usual SS-ANOVA models can then be borrowed when the smoothing paramter is fixed. In our implementation, we iterate between estimating $\theta$ and $c$.

## 3.4 Algorithm for a Fixed Smoothing Parameter

Denote expression (3.2.3) as $A(c, \theta)$, where $c = (c_1, ..., c_m)^T$. Denote $Q$ as an $m \times m$ matrix with $(k, l)$ entry $K_\theta(x_{k*}, x_{l*})$ and $Q_\alpha$ as an $m \times m$ matrix with $(k, l)$ entry $K_\alpha(x_{k*}, x_{l*})$. Let $U$ be an $n \times m$ matrix with $(k, l)$ entry being $K_\theta(x_k, x_{l*})$ and $U_\alpha$ be an $n \times m$ matrix with $(k, l)$ entry $K_\alpha(x_k, x_{l*})$. Denoting $\delta$ as the vector of censoring indicators $\delta = (\delta_1, ..., \delta_n)^T$, we can write (3.2.3) in

the matrix form as

$$A(c,\theta) = -\frac{1}{n}\delta^T U c + \frac{1}{n}\sum_{i=1}^{N}\log(\sum_{j\in R_i} e^{U_j c}) + \lambda_0 c^T Q c, \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M,\ \theta_\alpha \ge 0, \tag{3.4.1}$$

where $U_j$ is the $j$th row of $U$.

The gradient vector and the hessian of $A$ with respect to $c$ are given by

$$\begin{aligned}
\frac{\partial A}{\partial c} &= -\frac{1}{n}U^T\delta + \frac{1}{n}\sum_{i=1}^{N}\frac{\sum_{j\in R_i} U_j^T e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}} + 2\lambda_0 Q c;\\
\frac{\partial^2 A}{\partial c \partial c^t} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i} U_j^T U_j e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}} - \frac{\sum_{j\in R_i} U_j^T e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}}\frac{\sum_{j\in R_i} U_j e^{U_j c}}{\sum_{j\in R_i} e^{U_j c}}\} + 2\lambda_0 Q.
\end{aligned} \tag{3.4.2}$$

When $\theta$ is fixed, the Newton-Rhaphson iteration is used to update $c$ as

$$c = c_0 - (\frac{\partial^2 A}{\partial c \partial c^T})^{-1}_{c_o}(\frac{\partial A}{\partial c})_{c_0}, \tag{3.4.3}$$

where $c_0$ is the current estimate of the coefficient vector, and the hessian and the gradient are evaluated at $c_0$.

Denote $G$ as an $m \times p$ matrix with $\alpha$th column being $Q_\alpha c$ and $S$ as an $n \times p$ matrix with $\alpha$th column being $U_\alpha c$, (3.2.3) can be written as a function of $\theta$

$$A(c,\theta) = -\frac{1}{n}\delta^T S\theta + \frac{1}{n}\sum_{i=1}^{N}\log(\sum_{j\in R_i} e^{S_j\theta}) + \lambda_0 c^T G\theta, \text{ s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M, \theta_\alpha \ge 0, \tag{3.4.4}$$

where $S_j$ is the $j$th row of $S$.

When $c$ is fixed, we can expand $A(c,\theta)$ around the current estimate $\theta_0$ via second order Taylor expansion as following

$$A(c,\theta) \approx A(c,\theta_0) + (\theta-\theta_0)^T(\frac{\partial A}{\partial \theta})_{\theta_0} + \frac{1}{2}(\theta-\theta_0)^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}(\theta-\theta_0),$$

where

$$\begin{aligned}\frac{\partial A}{\partial \theta} &= -\frac{1}{n}S^T\delta + \frac{1}{n}\sum_{i=1}^{N}\frac{\sum_{j\in R_i} S_j^T e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}} + \lambda_0 G^T c; \\ \frac{\partial^2 A}{\partial\theta\partial\theta^T} &= \frac{1}{n}\sum_{i=1}^{N}\{\frac{\sum_{j\in R_i} S_j^T S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}} - \frac{\sum_{j\in R_i} S_j^T e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\frac{\sum_{j\in R_i} S_j e^{S_j\theta}}{\sum_{j\in R_i} e^{S_j\theta}}\}.\end{aligned} \tag{3.4.5}$$

The iteration for updating $\theta$ is via the minimization of the following linearly constrained quadratic objective function

$$\frac{1}{2}\theta^T(\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta + [(\frac{\partial A}{\partial\theta})_{\theta_0} - (\frac{\partial^2 A}{\partial\theta\partial\theta^T})_{\theta_0}\theta_0]^T\theta, \quad \text{s.t. } \sum_{\alpha=1}^{p}\theta_\alpha \le M,\ \theta_\alpha \ge 0, \tag{3.4.6}$$

The precense of the linear constraint on the sum of $\theta_\alpha$'s makes it possible to estimate some of $\theta_\alpha$'s as exact zeros, which leads to zero fitted components.

For fixed $\lambda_0$ and $M$, we iterate between updating $c$ and $\theta$. Our experience shows that it can take a large number of iterations for the algorithm to converge. As argued in Fan and Li (2002), the one-step penalized partial likelihood estimator can be as efficient as the fully iterative one with a good initial starting estimate of $\eta$. The fact that our algorithm starts with the smoothing spline estimate indicates that we do not need an exact solution. We observe empirically that after the initial iteration, the update in the estimate changes fairly slowly. Thereofore, a one step quadratic programing update for estimating $\theta$ provides sufficient iteration.

The algorithm for a fixed smoothing paramter is the following one step update procedure:

1. Initialization: Fix $\theta_0 = (1, ..., 1)^T$;

2. Use Newton-Rhaphson iteration (3.4.3) to solve for $c$ until the change in $c$ is less than some threshold;

3. Expand $A(c, \theta)$ around $\theta_0$, and use (3.4.6) to solve for $\theta$. Denote the solution as $\theta_\tau$;

4. Use (3.4.3) to solve for $c$ with the new $\theta$ until the change in $c$ is less than some threshold. Denote the solution as $c_\tau$;

5. Output the estimate as $\eta_\tau = K_{\theta_\tau} c_\tau$.

# Chapter 4

# Choosing the Smoothing Parameter

The problem of choosing the smoothing parameter(s) is very important in non-linear estimation. The smoothing parameter $\tau$ in the Cosso formulation governs the fidelity to the data and the roughness of the estimate. Different values of $\tau$ give different estimates. When $\tau$ is small, the estimate has small bias but large variance. When $\tau$ increases, the solution goes to a parametric model in the null space $N_J = \{\eta : J(\eta) = 0\}$, which is zero in the Cosso case. The estimated hazard function corresponding to $\tau = +\infty$ is simply the Nelson-Aalen estimator (Nelson (1972), Aalen (1978)). By varying the smoothing parameter, features of the data that arise on different scales can be explored. In practice, a specific value of $\tau$ has to be chosen, which calls for effective methods for smoothing parameter selection.

An automatic method is desired whereby the smoothing parameter is adaptively chosen by the data such that the estimate is close to the true relative hazard function. The discrepancy between two probability distributions is often measured in terms of Kullback-Leibler (KL) loss. Suppose $g_\tau$ is an estimate

of the true density $g$, the KL loss between $g_\tau$ and $g$ is defined as

$$KL(g, g_\tau) = E_g \log \frac{g}{g_\tau},$$

where $E_g$ denotes the expectation under $g$. It is easy to see that $KL(g, g_\tau) \geq 0$ and the equality holds if and only if $g = g_\tau$. For the relative risk function, the KL loss is

$$KL(\eta, \eta_\tau) = E[l_n(\eta_\tau) - l_n(\eta)],$$

where $l_n(\eta_\tau)$ is the negative log likelihood of the data given an estimate $\eta_\tau$ and $l_n(\eta)$ is the negative log likelihood of the data at the true function.

Excluding a quantity not depending on $\eta_\tau$, we minimize the so called comparative Kullback-Leibler (CKL) loss, namely

$$CKL(\tau) = CKL(\eta, \eta_\tau) = KL(\eta, \eta_\tau) + E[l_n(\eta)] = E[l_n(\eta_\tau)].$$

Minimizing the $CKL$ score is equivalent to maximize the expected log-likelihood for future observations. Ideally, if enough data are available, we would set aside a validation set and use it to assess a sampled version of the CKL. Without a separate validation data set, a popular technique is the $K$ fold cross validation, usually $K = 5$ or $K = 10$. In the $K$ fold cross validation, the data is split into roughly equal size parts. For each part $k$, one fits the model using the other $K-1$ parts and calculates the CKL for the $k$th part which is not used for model fitting. The same is done for each part and one averages the $K$ estimates to get an estimate of CKL for each $\tau$. The smoothing parameter which gives the smallest average CKL loss corresponds to the final estimate. This technique

involves fitting the same model $K$ times and is not computationally efficient. The extreme case is the so called leave-out-one cross validation where $K = n$. It is the various approximations of the leave-out-one cross validation score to estimate the CKL loss that have been under extensive study. In Gaussian SS-ANOVA models, Craven and Wahba (1979) proposed the generalized cross validation (GCV) criterion to choose the smoothing parameter. Xiang and Wahba (1996) developed a generalized approximate cross validation (GACV) for SS-ANOVA models in exponential families. In the estimation of the relative risk function, O'Sullivan (1988a) proposed a GCV type criterion by using an iterative reweighted least square algorithm. An AIC type of criterion is derived in O'Sullivan (1988b) when covariates are absent.

In this chapter, we derive two approximations to the leave-out-one estimate of the CKL loss.

## 4.1 ACV Criterion

We first derive an approximate cross validation criterion (ACV) to estimate the CKL score. A similar criterion is obtained in Chapter 7.2 of Gu (2002) by estimating a leave-out-one cross validation score using the counting process approach.

### 4.1.1 Leave-Out-One Cross Validation

Let the observed minus fitted partial likelihood be

$$PL(\tau) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta_\tau(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta_\tau(x_j))].$$

It is well known that $PL(\tau)$ tends to underestimate $CKL(\tau)$ due to the fact that $(Z_i, \delta_i)$ is used to estimate $\eta_\tau(Z_i|x_i)$. To correct this bias, the leave-out-one cross validation criterion in the following is used to estimate $CKL$

$$CV(\tau) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\eta_\tau^{[-i]}(x_i) + \frac{1}{n}\sum_{i=1}^{N}\log[\sum_{j\in R_i}\exp(\eta_\tau(x_j))], \tag{4.1.1}$$

where $\eta^{[-i]}(x_i)$ stands for the fitted log relative risk at $x_i$ when (3.1.1) is fitted without the $i$th data point. The second term in (4.1.1) is not cross validated. Similar argument can be found in Xiang and Wahba (1996), Lin *et al.* (2000), Gao *et al.* (2001). Since $\eta_\tau^{[-i]}$ is independent of $(Z_i, \delta_i)$, it is expected

$$E\delta_i\eta_\tau^{[-i]}(x_i) \approx E\delta_i\eta_\tau(x_i).$$

Hence we can expect $CV(\tau)$ to be roughly unbiased for computing $CKL(\tau)$.

### 4.1.2 Approximate Estimate

Because of the partial likelihood formulation, the leave-out-one version of (3.1.1) is complicated and it is prohibitively expensive to compute. Here a one step Newton-Raphson expansion is used to approximate the leave-out-one estimate $\eta^{[-i]}(x_i)$. A modified leave-out-one cross validation is to minimize

$$-\frac{1}{n-1}\sum_{j\neq i}\delta_j\eta(x_j) + \frac{1}{n}\sum_{j=1}^{N}\log[\sum_{k\in R_j}\exp(\eta(x_k))] + \tau J(\eta), \tag{4.1.2}$$

where the $i$th data point is left out. The iteration starts at the solution of (3.1.1) and only one step Newton-Raphson update is used to estimate $\eta^{[-i]}(x_i)$. The invovlement of the $i$th observation in the second term makes computation much easier. The modified cross validation seems to work well in the simulations.

To estimate $\eta_\tau^{[-i]}$, we need to estimate $\theta_\tau^{[-i]}$ and $c_\tau^{[-i]}$ for (4.1.2). In our derivation, we ignore the variability in estimating $\theta$ and assume that the estimate $\theta_\tau^{[-i]}$ does not differ from $\theta_\tau$. The one step Newton-Raphson iteration is used to estimate $c_\tau^{[-i]}$.

Using the matrix form (3.4.3), the minimizer of (4.1.2) can be esimated from solution $c_\tau$ of (3.2.3) by one step iteration as following

$$c_\tau^{[-i]} \approx c_\tau - (H_{c_\tau} + 2\lambda_0 Q)^{-1}\{(\frac{\partial A}{\partial c})_{c_\tau} + \frac{U^T\delta}{n} - \frac{U^T\delta^{[-i]}}{n-1}\},$$

where $\delta^{[-i]} = (\delta_1, ..., \delta_{i-1}, 0, \delta_{i+1}, ..., \delta_n)^T$ and $H_{c_\tau}$ is the hessian matrix of the minus partial likelihood evaluated at $c_\tau$.

Follow simple algebra,

$$\begin{aligned} c_\tau^{[-i]} &\approx c_\tau - (H_{c_\tau} + 2\lambda_0 Q)^{-1}\{(\frac{\partial A}{\partial c})_{c_\tau} - \frac{U^T\delta}{n(n-1)} + \frac{\delta_i U_i^T}{n-1}\}, \\ &= c_\tau - (H_{c_\tau} + 2\lambda_0 Q)^{-1}(\frac{\partial A}{\partial c})_{c_\tau} - H_{c_\tau}^{-1}\{-\frac{U^T\delta}{n(n-1)} + \frac{\delta_i U_i^T}{n-1}\} \\ &\approx c_\tau - (H_{c_\tau} + 2\lambda_0 Q)^{-1}\{-\frac{U^T\delta}{n(n-1)} + \frac{\delta_i U_i^T}{n-1}\}, \end{aligned}$$

since $c_\tau \approx c_\tau - (H_{c_\tau} + 2\lambda_0 Q)^{-1}(\frac{\partial A}{\partial c})_{c_\tau}$ at convergence. We have

$$\begin{aligned} \eta_\tau^{[-i]}(x_i) &= U_i c_\tau^{[-i]} \\ &\approx \eta_\tau(x_i) - \frac{1}{n-1} U_i (H_{c_\tau} + 2\lambda_0 Q)^{-1}(\delta_i U_i^T - \frac{U^T\delta}{n}). \end{aligned}$$

It follows

$$\begin{aligned} CV(\tau) &= PL(\tau) + \frac{1}{n}\sum_{i=1}^{n}(\delta_i\eta_\tau(x_i) - \delta_i\eta_\tau^{[-i]}(x_i)) \\ &\approx PL(\tau) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\delta_i U_i(H_{c_\tau} + 2\lambda_0 Q)^{-1}(\delta_i U_i^T - \frac{U^T\delta}{n}) \\ &= PL(\tau) + \{\frac{\mathrm{tr}(\Delta U(H_{c_\tau} + 2\lambda_0 Q)^{-1}U^T\Delta)}{n(n-1)} - \frac{\delta^T U(H_{c_\tau} + 2\lambda_0 Q)^{-1}U^T\delta^T}{n^2(n-1)}\}, \end{aligned}$$

where $\Delta = \mathrm{diag}(\delta_1, ..., \delta_n)$.

A simple modification of the above expression, which is called approximate cross validation (**ACV**), appears as follows

$$\mathbf{ACV}(\tau) = PL(\tau) + \frac{N}{n}\{\frac{\mathrm{tr}(U^T(H_{c_\tau} + 2\lambda_0 Q)^{-1}U)}{n(n-1)} - \frac{\mathbf{1}^T U^T(H_{c_\tau} + 2\lambda_0 Q)^{-1}U\mathbf{1}}{n^2(n-1)}\}, \tag{4.1.3}$$

where $\mathbf{1} = (1, ..., 1)^T$ is a vector of ones. The ACV criterion averages the effect of the censoring. No extra computation is needed to compute ACV once an estimate is obtained.

## 4.2 Another Approximate Cross Validation Criterion

In deriving the leave-out-one estimate of the hazard function, the change of the second term due to the fact that the model is fitted without the $i$th observation in (4.1.2) is ignored. The corresponding second term in $PL(\tau)$ is not cross validated. In this section, we derive another cross validation criterion targeting

at the full leave-out-one version of the likelihood.

In the case that the baseline hazard is known, the cross validation scores are easy to define. We propose to use these scores with the baseline hazard replaced by the partial likelihood estimate.

Denote the baseline culmulative hazard as $\Lambda_0(t)$. The penalized partial likelihood can be replaced by the penalized full likelihood and (3.1.1) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\delta_i[\eta(x_i)+\log h_0(z_i)]+\frac{1}{n}\sum_{i=1}^{n}\Lambda_0(z_i)\exp(\eta(x_i))+\tau J(\eta). \qquad (4.2.1)$$

The full likelihood is denoted as $l_n$.

The variability in estimating $\theta$ is again ignored. A one step Newton-Raphson expansion is used to approximate the leave-out-one estimates

$$c_\tau^{[-i]} \approx c_\tau - (H+2\lambda_0 Q)^{-1} g^{[-i]},$$

where $g^{[-i]}$ is the gradient of the leave-out-one version of (4.2.1), and $H$ is the converged hessian of $l_n$ at $c_\tau$. Since the gradient is zero at $c_\tau$, one has

$$g^{[-i]} = \frac{1}{n}\delta_i U_i^T - \frac{1}{n}\Lambda_0(z_i)\exp(\eta_\tau(x_i))U_i^T.$$

The leave-out-one version of the CKL loss is simply

$$\begin{aligned} CV(\tau) &= -\frac{1}{n}\sum_{i=1}^{n}\delta_i[\eta_\tau^{[-i]}(x_i)+\log h_0(z_i)]+\frac{1}{n}\sum_{i=1}^{n}\Lambda_0(z_i)\exp(\eta_\tau^{[-i]}(x_i)) \\ &= LH(\tau)+\frac{1}{n}\sum_{i=1}^{n}\delta_i[\eta_\tau(x_i)-\eta_\tau^{[-i]}(x_i)] \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\Lambda_0(z_i)[\exp(\eta_\tau^{[-i]}(x_i))-\exp(\eta_\tau(x_i))] \\ &\approx LH(\tau)+\sum_{i=1}^{n}\{g^{[-i]}\}^T H^{-1} g^{[-i]} \\ &= LH(\tau)+\frac{1}{n}trace[(H+2\lambda_0 Q)^{-1}H^*] \end{aligned}$$

where $LH(\tau)$ stands for fitted minus log likelihood for $\eta_\tau$, and $H^* = n\sum_{i=1}^{n} g^{[-i]}\{g^{[-i]}\}^T$.

Let $l(t,\delta,c)$ be the contribution to the negative likelihood by a single observation, i.e.

$$l(z_i,\delta_i,c) = \Lambda_0(z_i)\exp(U_i c) - \delta_i U_i c - \delta_i \log h_0(z_i).$$

If the model is correct then with minimal conditions at the true value of $c$, there is a familiar relation between the first and the second derivatives of the negative log-likelihood function

$$E(\partial_c l \partial_c l) = E(\partial_c^2 l).$$

From this we develop the approximations

$$\begin{aligned} H^* &= n\sum_{i=1}^{n} g^{[-i]}\{g^{[-i]}\}^T \approx E(\partial_c l \partial_c l) \\ &= E(\partial_c^2 l) \approx H. \end{aligned}$$

Substituting the fitted likelihood by the fitted partial likelihood (up to a constant) and replacing the hessian by the hessian of the partial likelihood, one is

led to another cross validation criterion

$$ACV^* = PL(\tau) + \frac{1}{n} trace[(H + 2\lambda_0 Q)^{-1} H].$$

Similar to the argument in Chapter 6.3 of Gu (2002), the method may severely undersmooth up to about 10% of the replicates in simulation studies. A simple modification is to multiply the trace term by a constant $\gamma$. We denote the modified version as $ACV^*(\gamma)$. Simulation studies suggest that a $\gamma$ around 1.4 is most effective.

## 4.3 The Full Algorithm

Combined with the one step update procedure, the complete algorithm to fit the Cosso estimate is the following:

1 Neglect $M$ and fix $\theta = \theta_0 = (1, ..., 1)^T$, tune $\lambda_0$ according to $ACV$ (or $ACV^*$);

2 For $M$ in a reasonable range, use the one step update scheme to calculate $ACV$ (or $ACV^*$). Choose the $M$ which gives the minimum $ACV$ (or $ACV^*$). The estimate corresponding to this chosen $M$ is the Cosso estimate.

In our implementation, we use the reformulation (3.2.3) instead of (3.2.2). Our simulations show that once $\lambda_0$ is fixed, the estimated number of nonzero

components is roughly equal to $M$. This correspondence greatly facilitate the specification of a reasonable range for the tuning parameter.

# Chapter 5

# Simulation Results

We conduct some simulations in this chapter to study the efficacy of our estimates in terms of prediction accuracy and model selection.

To measure the prediction performance of an estimate $f_\tau$ of the true density function $f$, we use the Kullback-Leibler loss defined as

$$KL(f, f_\tau|X) = E_f \log \frac{f(T|X)}{f_\tau(T|X)}. \tag{5.0.1}$$

It is easy to see

$$\begin{aligned} E_f \log f_\tau(T|X) &= \int_0^\infty \{\log f_\tau(t|X)\} f(t|X) dt \\ &= \int_0^\infty \log\{h_0(t)\exp(\eta_\tau(X))\exp[-\int_0^t h_0(u)\exp(\eta_\tau(X))du]\} \\ &\quad \cdot h_0(t)\exp(\eta(X))\exp[-\int_0^t h_0(u)\exp(\eta(X))du]dt. \end{aligned} \tag{5.0.2}$$

In the simulations to follow, we use $h_0(t) = 1$ as the baseline hazard function. Plugging $h_0(t)$ into (5.0.2), one gets

$$\begin{aligned} E_f \log f_\tau(T|X) &= \int_0^\infty [\eta_\tau(X) - \exp(\eta_\tau(X))t] \\ &\quad \cdot \exp(\eta(X))\exp[-\exp(\eta(X))t]dt \\ &= \eta_\tau(X) - \exp\{\eta_\tau(X) - \eta(X)\}. \end{aligned} \tag{5.0.3}$$

Therefore the KL loss is simply

$$KL(f, f_\tau|X) = \eta(X) - \eta_\tau(X) + \exp[\eta_\tau(X) - \eta(X)] - 1.$$

We evaluate $KL(\eta, \eta_\tau) = E_X[KL(f, f_\tau|X)]$ by Monte Carlo integration using 10000 test points from the same distribution as the training points.

The following mechanism is used to generate d dimensional covariate $X = (X^{(1)}, ..., X^{(d)})$:

$$X^{(j)} = (U^{(j)} + tU)/(1 + t), j = 1, ..., d,$$

where $U^{(1)}, ..., U^{(d)}$ and $U$ are i.i.d. from uniform(0,1). Therefore, $corr(x^{(j)}, x^{(k)}) = t^2/(1+t^2)$ and the marginal distributions of $X^{(i)}$'s are uniform on $[0, 1]$. We use $t = 0$ and $t = 1$ to generate uncorrelated covariates and correlated covariates with a correlation $\rho = 0.5$.

## 5.1 Efficacy of the Corss Validation Criteria

We first illustrate the efficacy of the ACV as a computing proxy of the theoretical KL loss via a simple example.

Random samples consisting of $n = 100$ and $n = 400$ are drawn from the following hazard model

$$h(t|x) = \exp\{g_1(x^{(1)}) + g_2(x^{(2)}) - D\},$$

where $g_1(s) = sin(\pi s^2)$, $g_2(s) = 0$ and $D$ is a normalizing constant such that $D = \int_0^1 g_1(s)ds$. This example has one true component $x^{(1)}$ and one noisy

component $x^{(2)}$. The censoring time is exponentially distributed with mean $U \exp[-g_1(x^{(1)})]$, where $U$ is randomly generated from the uniform distribution on $[1, 3]$. The empirical censoring rate is about 34%. One hundred data sets are generated and the KL loss for each data set is calculated.

Additive models are fitted on the grid $M = (0.2)(0.1)(4)$. 50 basis functions are randomly chosen to approximate the solution. Figure 5.1.1 depicts the pairwise comparison for the theoretical best KL loss $KL(\eta, \eta_\tau)$ on the grid and the KL loss corresponding to the best estimate using ACV criterion. We see that the ACV serves as an excellent proxy for approximating the KL loss.

The same example with $n = 100$ and $\rho = 0$ is used to demonstrate the effectiveness of $ACV^*(\gamma)$. Figure 5.1.2 shows $KL(\eta, \eta_\tau)$ of the $ACV^*(\gamma)$ with (a) $\gamma = 1$ and (b) $\gamma = 1.4$ versus the minimum KL loss on the grid. Also plotted in (c) and (d) are the minimum KL loss obtained by versions of $ACV^*$ versus minimum KL loss obtained by $ACV$ which corresponds to Figure 5.1.1 (a). The plots suggest that the modified $ACV^*$ score ($\gamma = 1.4$) may gain significantly over unmodified one on some replicates but only lose minimally on some others. The plots also suggest that the performance of $ACV^*$ with $\gamma = 1.4$ is comparable to that of $ACV$.

We use $ACV$ score in the following simulations and data examples.

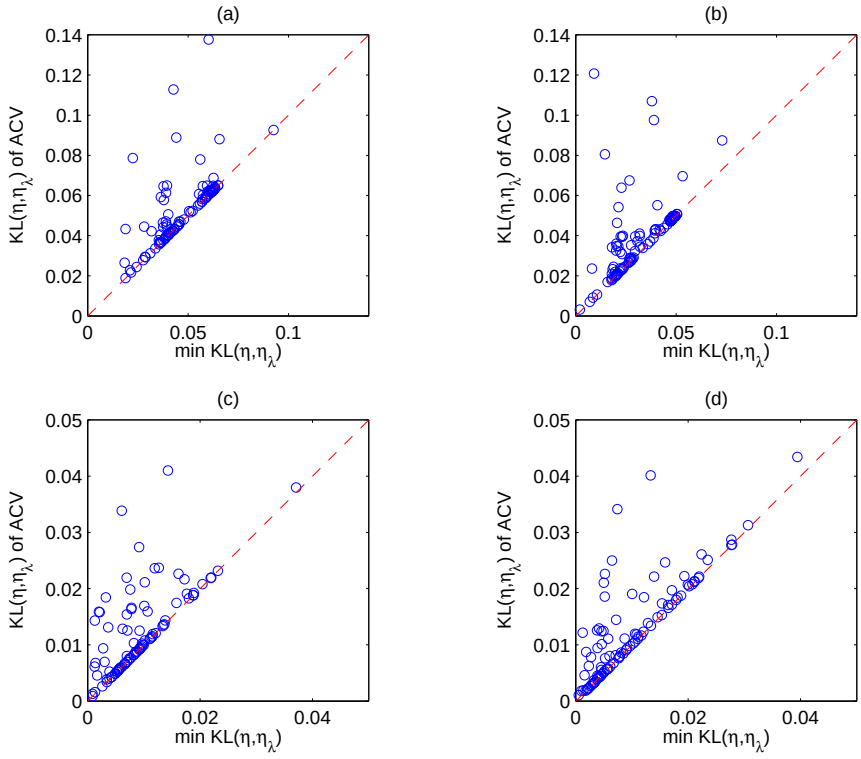


Figure 5.1.1: Performance of ACV for hazard estimation. (a) $n = 100$ and $\rho = 0$; (b) $n = 100$ and $\rho = 0.5$; (c) $n = 400$ and $\rho = 0$; (d) $n = 400$ and $\rho = 0.5$.

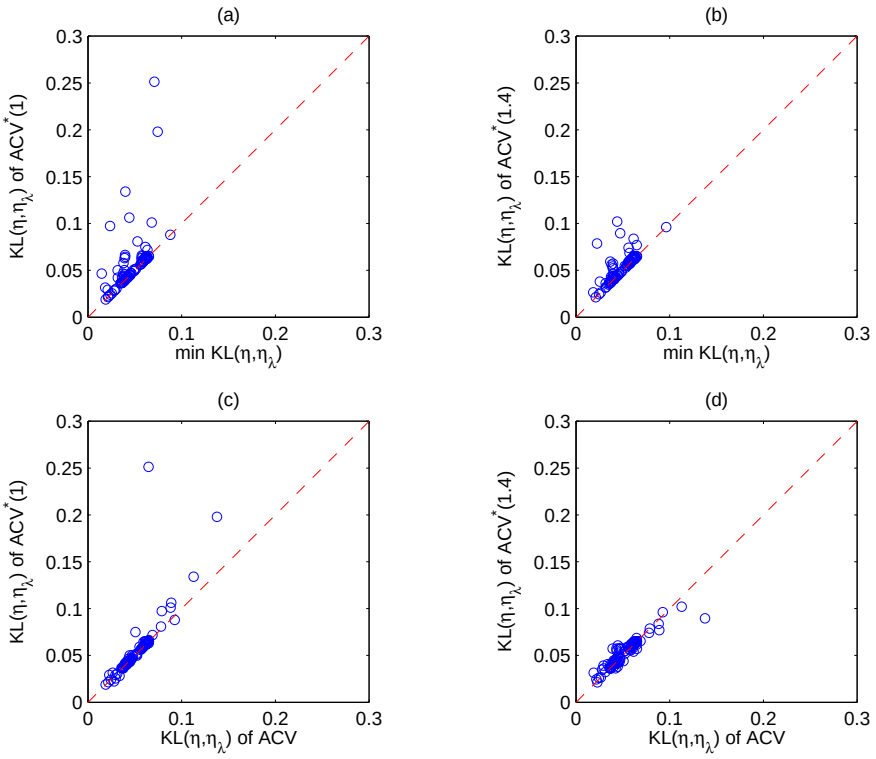


Figure 5.1.2: (a) Performance of $ACV^*(1)$: the minimum KL loss obtained by $ACV^*$ versus the theoretical minimum on the grid, (b) Performance of $ACV^*(1.4)$: the minimum KL loss obtained by $ACV^*$ versus the theoretical minimum on the grid, (c) the minimum KL loss obtained by $ACV^*(1)$ versus that by $ACV$, (d) the minimum KL loss obtained by $ACV^*(1.4)$ versus that by $ACV$,

## 5.2 More Complicated Simulations

To study the performance of model selection and component estimation, we conduct more complicated simulations. To assess the goodness of component estimation, the integrated square error is used instead of the KL loss,

$$ISE = E_X\{\eta(X) - \eta_\tau(X)\}^2.$$

For each replicate of the simulation, the ISE is estimated by Monte Carlo integration using 10000 test points from the same distribution as the training points. We run the simulation 100 times and average.

The following basic functions in Lin and Zhang (2002) are used for the building block for our examples,

$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{sin(2\pi t)}{2 - sin(2\pi t)};$$

$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t).$$

To include categorical variable, we add

$$g_5(t) = t, \quad t = 0, 1.$$

We consider additive model on $[0,1]^{10}$ with the true hazard function being

$$\eta(x) = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)} + 6g_4(x^{(4)}) + 3g_5(I_{(x^{(5)}>0.6)}).$$

Two more categorical variables are introduced as $I_{(x^{(6)}<0.8)}$ and $I_{(x^{(7)}>0.2)}$.

Sample sizes of $n = 100,\ 200,\ 400,\ 800$ are generated from the exponential hazard function

$$h(t|x) = \exp(\eta(x)).$$

The censoring time is exponentially distributed with mean $U\exp(-\eta(x))$, where $U$ is randomly generated from the uniform distribution on $[1, 3]$. The empirical censoring rate is about 35%. Notice the censoring is noninformative given $x$ since $\eta(x)$ is a known function. We use either all the observations up to 200 or randomly chosen $m = 50$ observations as basis functions. We fit additive models and tune $M$ in the range $(0.5)(0.5)(min\{p + 2, 35\})$.

The magnitudes of the functional components are measured by their empirical $L_1$ norms , defined as $1/n \sum_{i=1}^{n} |\eta_\alpha(x_i^{(j)})|$ for $\alpha = 1, ..., d$. Figure 5.2.1 depicts how the empirical $L_1$ norms of the estimated components change with the tuning parameter in one single run. The $ACV$ criterion chooses $M = 2.5$ in this simulation, giving a model of 5 components in the final estimate.

The model selection results are summarized in Table 5.2.1 for the independent case and Table 5.2.2 for the compound symmetry case. The column "No.Cor.Mod" refers to the number of models which are correctly identified. The average number of zero components is reported in the column "Aver.no. of 0 Comp", where "correct" presents the average restricted to the true nonzero components, and "incorrect" indicates the average number of components erroneously set to zero. For a sample size $n = 100$, about 50% of the estimates correctly identify the true model for independent covariates, while for correlated covariates with a correlation 0.5, this ratio is about 30%. For a sample size $n = 800$, approximately 90% of the estimates correctly identify the true model, for independent and correlated covariates.

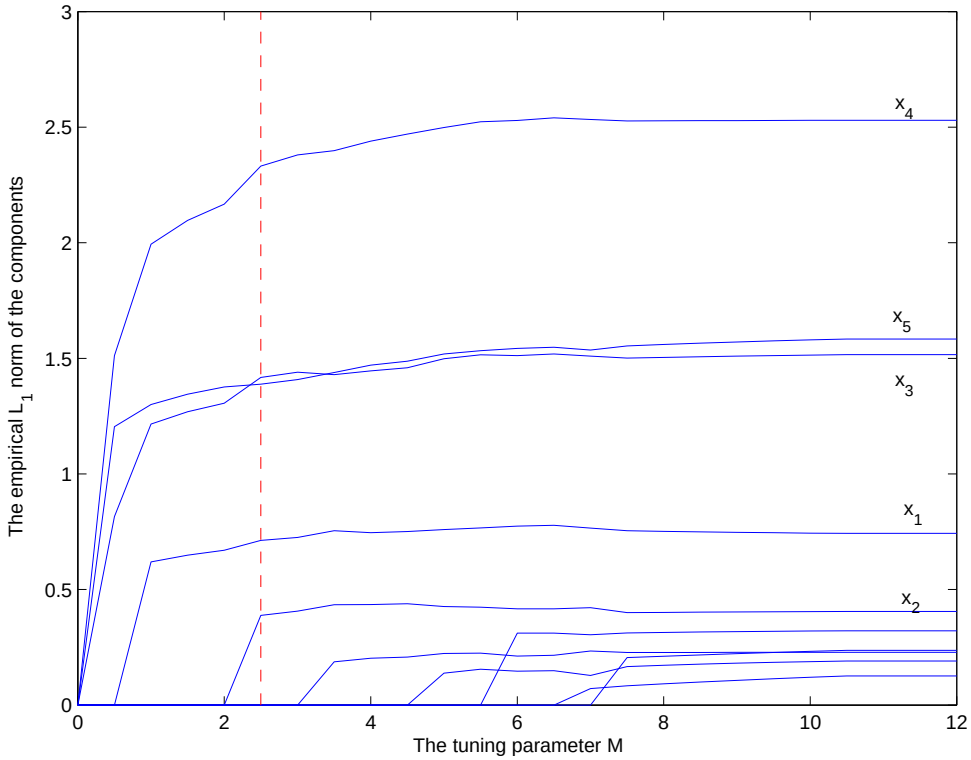


Figure 5.2.1: The empirical $L_1$ norm of the estimated components against the tuning parameter $M$ in one run when $n = 200$ and $\rho = 1$. The red dashed line indicates the $M$ chosen by $ACV$ criterion.

| $n(m)$ | No.Cor.Mod. | Aver.no.of 0 Comp. | |
|---|---|---|---|
| | | correct | incorrect |
| 100 (100) | 53 | 4.95 | 0.64 |
| 100 (50) | 55 | 4.94 | 0.55 |
| 200 (200) | 70 | 5.00 | 0.40 |
| 200 (50) | 69 | 5.00 | 0.40 |
| 400 (50) | 76 | 5.00 | 0.28 |
| 800 (50) | 89 | 5.00 | 0.11 |

Table 5.2.1: Simulation results in terms of model selection for $\rho = 0$.

| $n(m)$ | No.Cor.Mod. | Aver.no.of 0 Comp. | |
|---|---|---|---|
| | | correct | incorrect |
| 100 (100) | 32 | 4.65 | 0.71 |
| 100 (50) | 27 | 4.64 | 0.55 |
| 200 (200) | 69 | 5.00 | 0.43 |
| 200 (50) | 73 | 4.99 | 0.36 |
| 400 (50) | 85 | 5.00 | 0.20 |
| 800 (50) | 88 | 5.00 | 0.15 |

Table 5.2.2: Simulation results in terms of model selection for $\rho = 0.5$.

We summarize the performance in terms of the ISE in Table 5.2.3. We can see that using a subset of observations as the basis does not degrade the performance. Furthermore, the ISE clearly has a decreasing trend, while the sample size increases.

| n(m) | cov=0 | cov=0.5 |
|---|---|---|
| 100(100) | 3.91(0.11) | 4.08(0.22) |
| 100(50) | 3.86(0.13) | 4.12(0.20) |
| 200(200) | 1.17(0.05) | 1.02(0.05) |
| 200(50) | 1.10(0.05) | 0.89(0.05) |
| 400(50) | 0.36(0.02) | 0.32(0.02) |
| 800(50) | 0.14(0.01) | 0.16(0.01) |

Table 5.2.3: Estimated integrated square error for the simulation. In parethesis are the standard errors.

We plot the 5th, 50th, 95th best estimates according to the ISE in Figure 5.2.2, Figure 5.2.3, Figure 5.2.4, Figure 5.2.5, Figure 5.2.6, Figure 5.2.7 for different sample sizes when 50 randomly chosen observations are used as basis functions. We see that the estimates follow very well with the true component

functions.

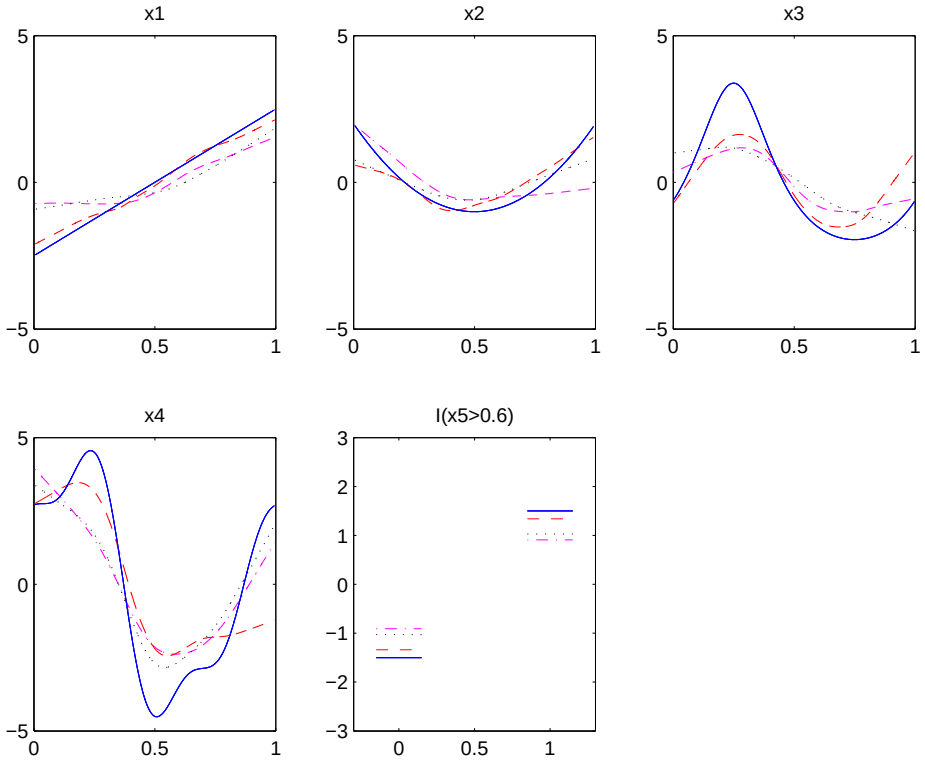


Figure 5.2.2: The estimated effects when $n = 100, m = 50$ and $\rho = 0$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

In summary, our proposal is very powerful at identifying the true subset of the important variables and estimating the components.

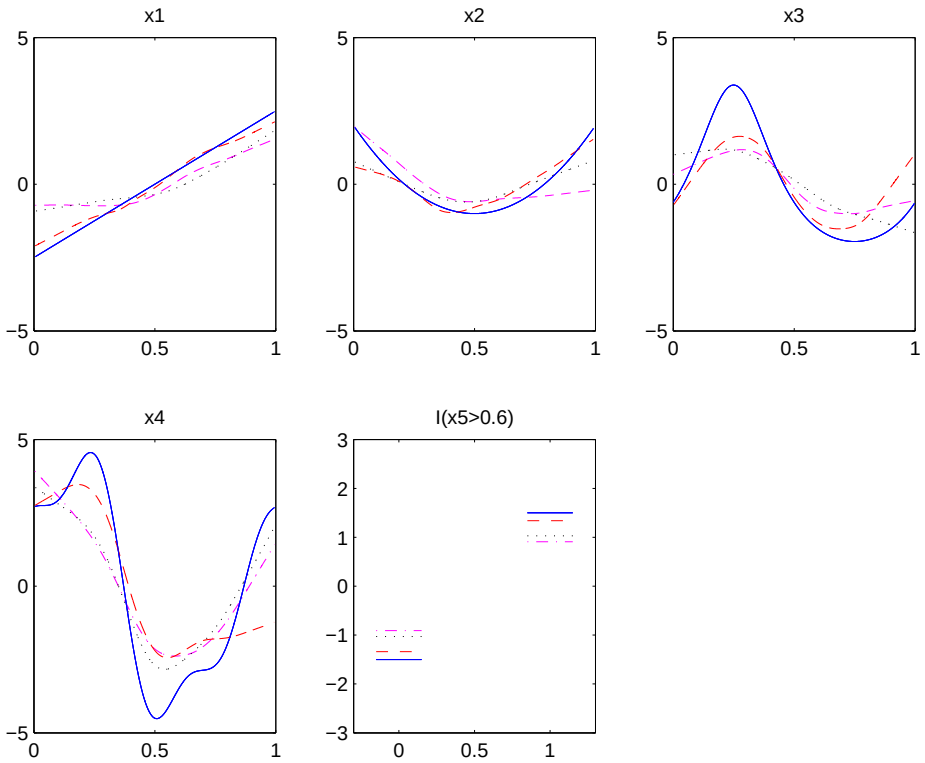


Figure 5.2.3: The estimated effects when $n = 100, m = 50$ and $\rho = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

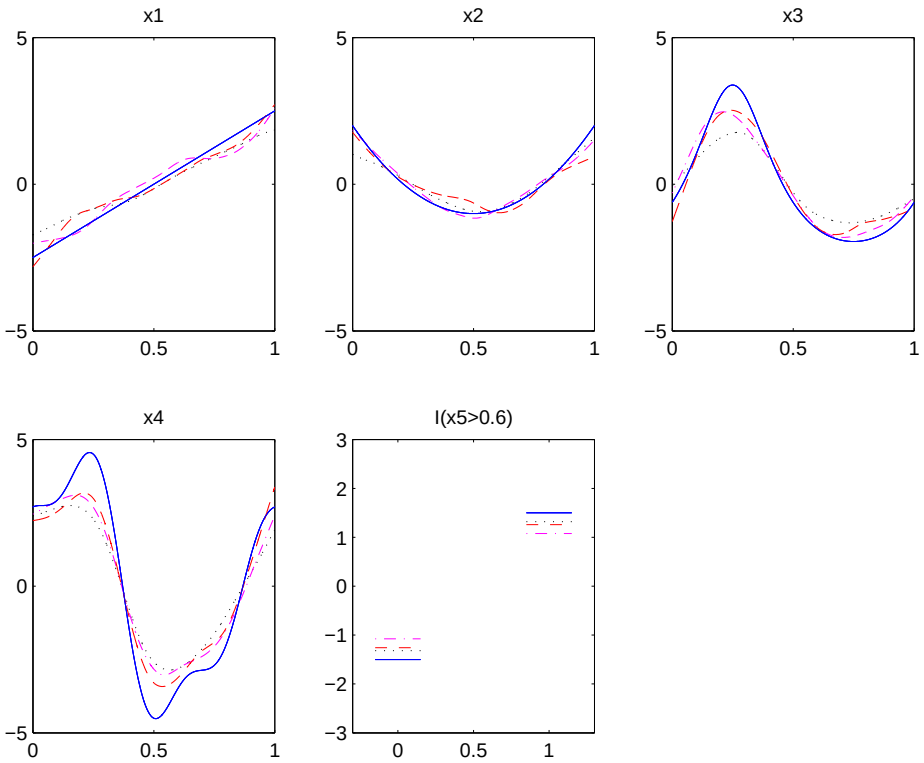


Figure 5.2.4: The estimated effects when $n = 200, m = 50$ and $\rho = 0$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

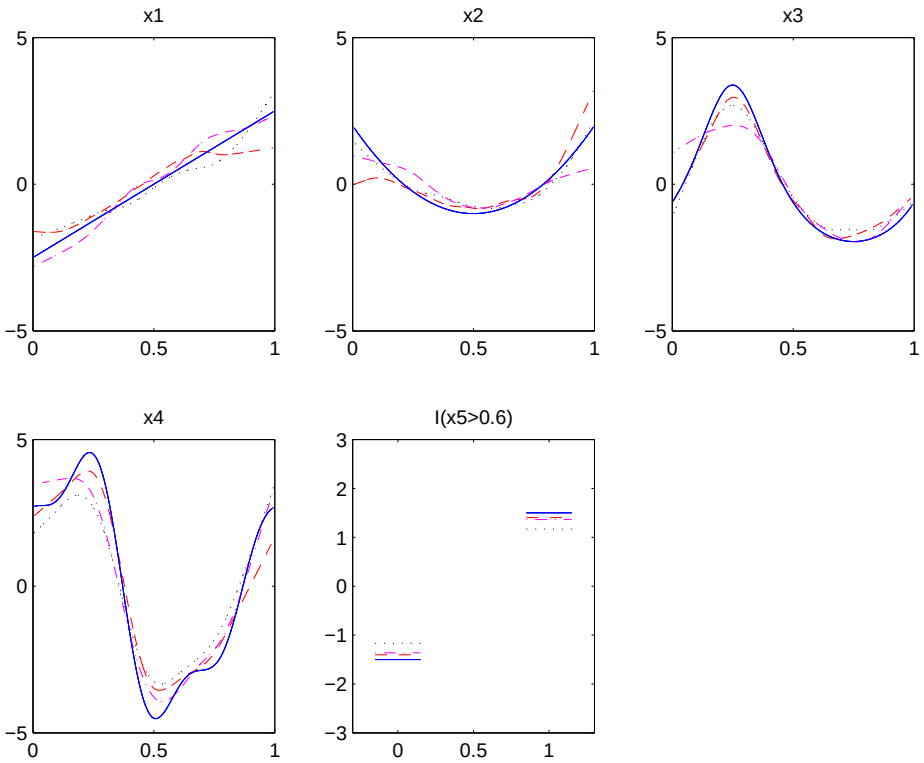


Figure 5.2.5: The estimated effects when $n = 200, m = 50$ and $\rho = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

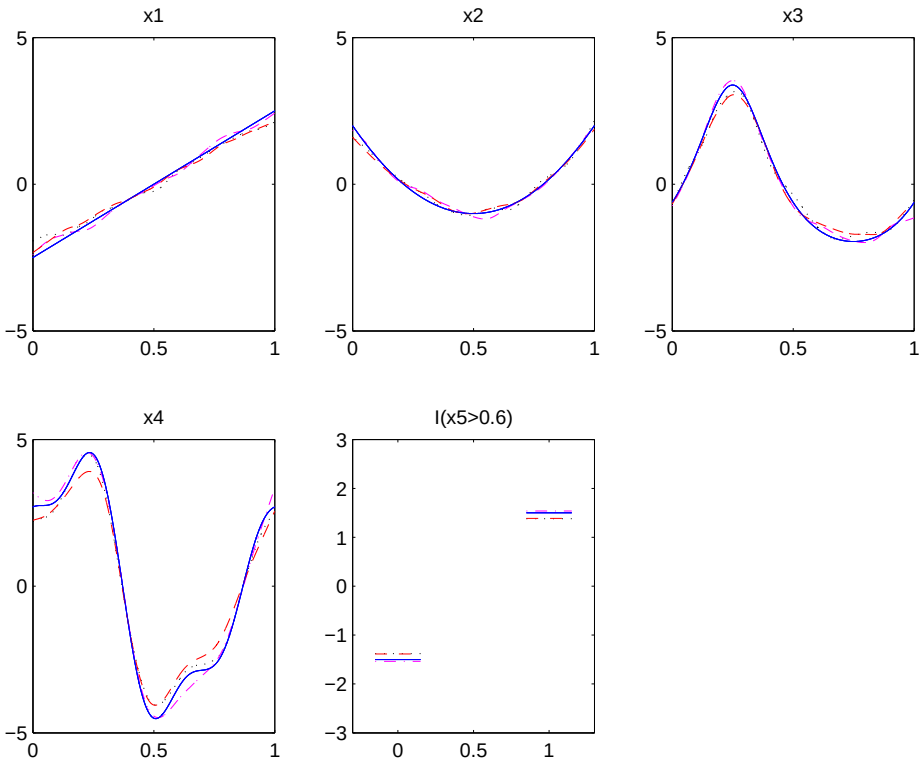


Figure 5.2.6: The estimated effects when $n = 800, m = 50$ and $\rho = 0$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

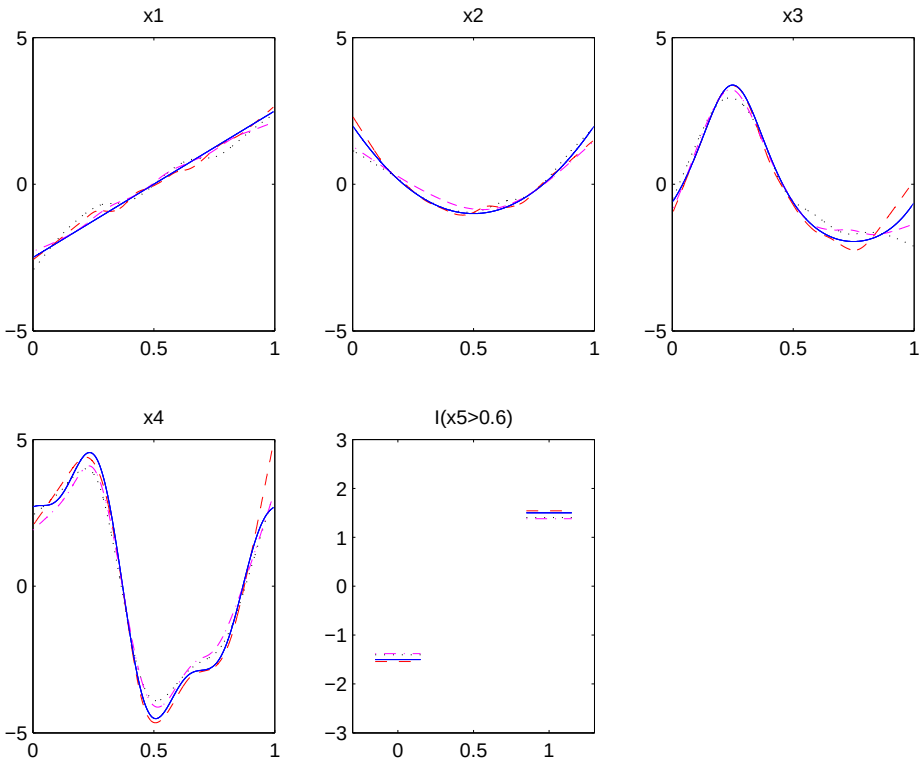


Figure 5.2.7: The estimated effects when $n = 800, m = 50$ and $\rho = 0.5$. The blue solid lines indicate the true components; the red dashed lines indicate the 5th best; the magenta dash-dot lines indicate the 50th best; the black dot lines are the 95th best.

# Chapter 6

# Real Data Examples

We apply our automatic model selection procedure to several well known data sets in this chapter.

## 6.1 Lung Cancer Data

The data is from the Veteran's Administration lung cancer trial, listed in Kalbfleisch and Prentice (2002), pp.378-379. There are 137 patients in the study and there are 9 censored observations among those. The main interest is to study the dependence of the survival time in days on the covariates listed in following:

1. treatment, 1=standard, 2=test.

2. celltype, 1=squamous, 2=smallcell, 3=adeno, 4=large.

3. Karnofsky performance score (10, 20, ..., 100=good).

4. months from diagnosis to randomization.

5. age in years.

6. prior therapy 0=no, 1=yes.

Tibshirani (1997) used the Lasso technique to choose important variables. Since the Lasso estimates a categorical variable as zero only when all the estimated levels are zeros, the cell type is treated as a continuous variable. It was found that only Karnofsky performance score is left in the final model using the Lasso or the stepwise regression.

Because of the nature of the SS-ANOVA models, a categocial component can be estimated as zero if its corresponding $\theta_\alpha$ is zero. Therefore, we treat cell type as a categorical variable. For parametric Cox proportional model, we use Mallow's Cp as the criterion in stepwise selection. The stepwise procedure chooses karnofsky performance score and cell type in the final model. The estimated effects for cell type are -0.550 for squamous, 0.166 for small cell, 0.608 for adeno and -0.224 for large cell. Clearly, cell type may not be treated as a continuous variable since the effect is clearly nonlinear. Our procedure yields the a similar estimate as plotted in Figure 6.1.1. The coefficients for different cell types are -0.545 for squamous, 0.198 for small cell, 0.592 for adeno and -0.244 for large cell. The estimates are quite close to those obtained by the stepwise variable selection. We also notice that the effect of Karnofsky performance score is linear, which suggests a linear model for karnofsky performance score may be sufficient for subsequent analysis.

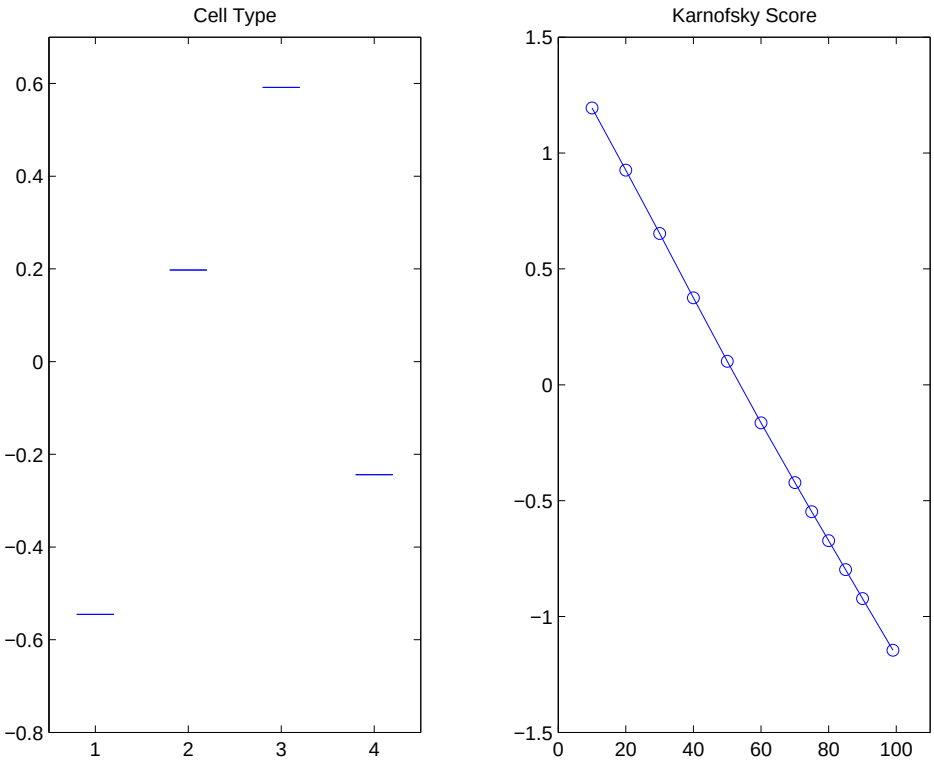


Figure 6.1.1: Fitted main effects for lung cancer data.

## 6.2 PBC Data

We analyze the primary biliary cirrhosis (PBC) data from Therneau and Grambsch (2000). The data are from the Mayo Clinic trial in primary biliary cirrhosis of liver conducted between 1974 and 1984. PBC is a progressive disease thought to be of an autoimmune origin. The subsequent inflammatory process eventually leads to cirrhosis and destruction of the liver's bileducts and death of the patient. The PBC database is a valuable resource to liver specialists because PBC is a rare but fatal disease. A more detailed discussion can be found in Dickson *et al.* (1989). In this study, 312 patients from a total of 424 patients who agreed to participate in the randomized trial are eligible for the analysis. For each of the 312 clinical trial patients, clinical, biochemical, serologic, and histologic parameters are collected. Of those, 125 patients died before the end of follow-up.

Two separate analyses are conducted to study the dependence of the survival time on the following covariates.

1 Continuous variables

age: age in years

alb: serum albumin in gm/dl

alk: alkaline phosphatase in U/liter

bil: serum bilirunbin in mg/dl

chol: serum cholesterol in mg/dl

cop: urine copper in $\mu$g/day

plat: platelets per cubic ml/1000

prot: standardized prothrombin time in seconds

sgot: liver enzyme (now called AST) in U/ml

trig: triglycerides in mg/dl

2 Categorical variables

asc: 0, absence of ascites; 1, presence of ascites

ede: 0 no edema; 0.5 untreated or successfully treated; 1 unsuccessfully treated edema

hep: 0, absence of hepatomegaly; 1, presence of hepatomegaly

sex: 0, male; 1, female

spid: 0, absence of spiders; 1, presence spiders

stage: histological stage of disease (needs biopsy), graded 1, 2, 3 or 4

trt: 1 for control, 2 for treatment

### 6.2.1 PBC Analysis A

We constrain our attention to the 11 covariates as shown in Table 4.4.2 of Fleming and Harrington (1991), including continuous variables age (age), albumin (alb), alkaline phos (alk), bilirubin (bili), platelet count (plat), pro time (prot) as well as categorical variables ascites (asc), edema (ede), hepatomegaly (hep), sex (sex) and spiders (spid). The four observations with missing data in those covariates are excluded.

We take log transformation to alk, bili and prot since they are skewed to

the right. After applying our procedure to the data, the ACV criterion selects $M = 4.5$ (Figure 6.2.1) as the final estimate of the smoothing parameter. The corresponding fitted main effects are shown in Figure 6.2.2. Comparing with Table 4.4.3 in Fleming and Harrington (1991), our model has one extra term asc. From the plot, the data could be fitted using proportional hazard models with parametric main effects.

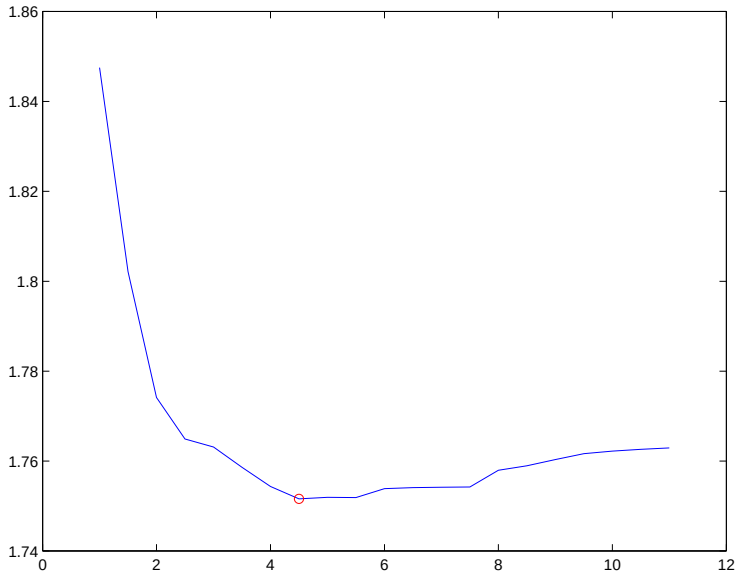


Figure 6.2.1: Cross validation curve for PBC analysis A. The minimum is obtained at $M = 4.5$.

### 6.2.2 PBC Analysis B

Tibshirani (1997) analyzed the PBC data via the Lasso using all the 17 covariates, which include 10 continuous variables and 7 categorical variables. We

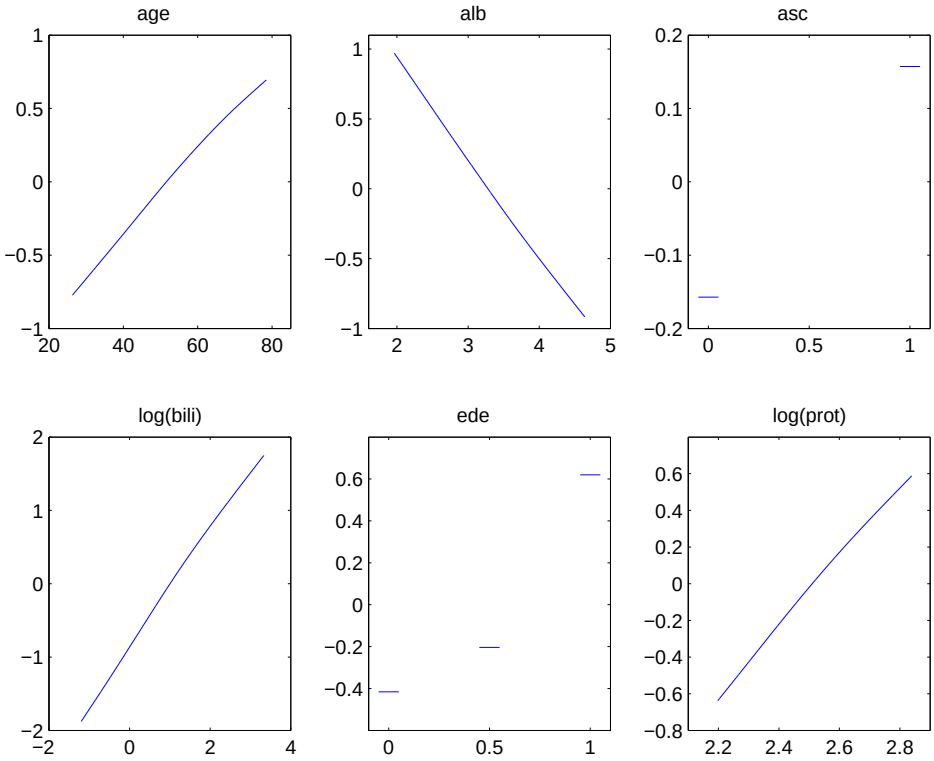


Figure 6.2.2: Fitted effects for PBC analysis A.

apply our procedure to 276 observations with no missing data in these covariates. As reported in Tibshirani (1997), the stepwise selection chooses age, ede, bili, alb, cop, sgot, prot and stage; totally 8 variables appear in the final selected model. The Lasso procedure selects three more variables, sex, asc and spid. Compared to the result of the stepwise selection, our procedure selects two more variables sex and chol. Quite interestingly, the stepwise model selects only those covariates with absolute Z-scores larger than 2.00, and our model selects only those covariates with absolute Z-scores larger than 1.00, where Z-scores refer to as the scores obtained via full parametric Cox proportional hazard model. The Lasso, instead, selects two covariates asc (Z-score 0.23) and spid (Z-score 0.42) with Z-scores less than 1 while leaving chol (Z-score 1.11) out of the model. It remains unclear that to what extent accounting for nonlinearity contributes to the discrepancy. The fitted effects of our model are shown in Figure 6.2.4. Except for cop, possible linear effects are quite obvious for the other covariates.

## 6.3 Mouse Leukemia Data

This data set is from Kalbfleisch and Prentice (2002), pp 390-395. The goal of this study is to examine genetic and viral factors which may influence the development of spontaneous leukemia in AKR mice. We consider continuous predictors antibody level (% gp 70 ppt), virus level (PFU/ml), and categorical predictors mhc phenotype (1 or 2), sex (1=male, 2=female) and coat color (1

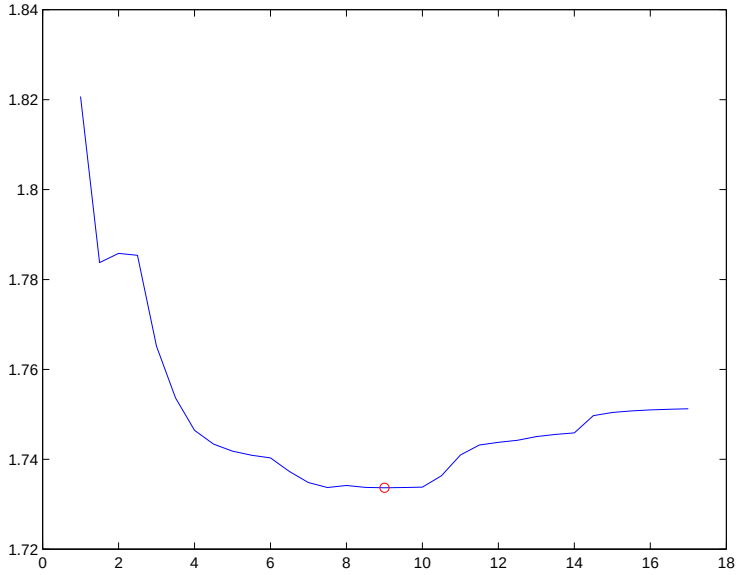


Figure 6.2.3: Cross validation curve for PBC analysis B. The minimum is obtained at $M = 9$.

or 2). The data set contains 175 mice after removing observations with missing covariates. We compare our analysis with parametric model selection, which is obtained by using the backward deletion option of the function stepAIC in the R library MASS and survival. This gives a final model with antibody as the only significant regressor. The linear model selection result is summarized in Table 6.3.1. We run our nonlinear procedure to the same data. Our procedure selects virus, mhc and coat as the final regressors. The fitted main effects are shown in Figure 6.3.1. This is to be compared with Analysis 6 in Kalbfleisch and Prentice (2002) Table 11.6. Kalbleisch and Prentice discretized virus and antibody and found when responses are allowed to vary among different level of

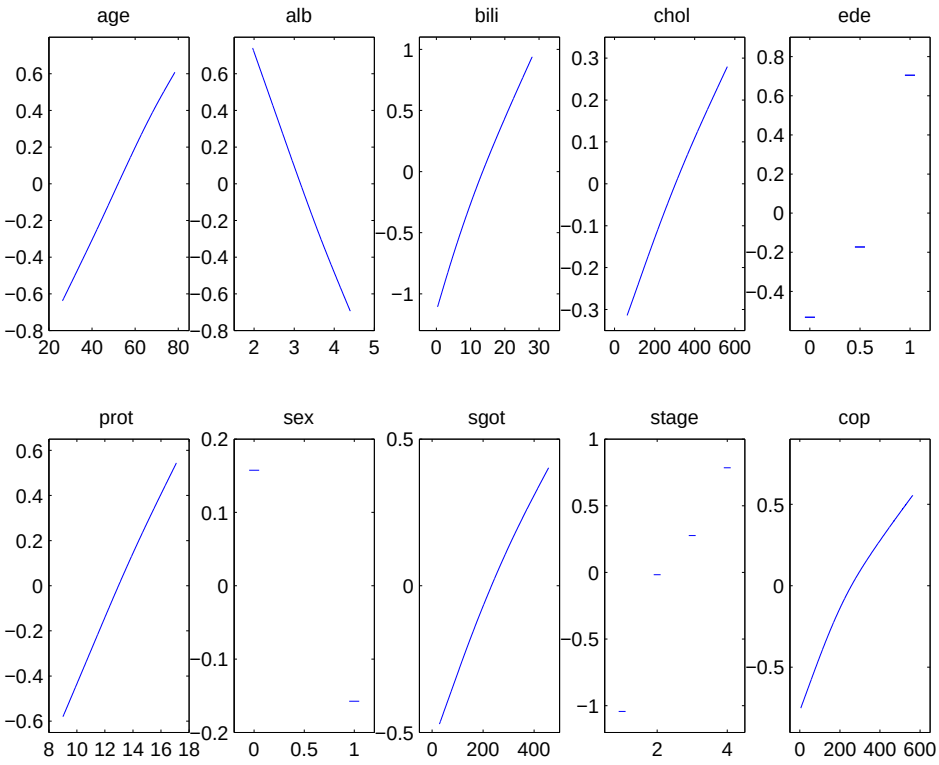


Figure 6.2.4: Fitted main effects while 17 covariates are considered.

virus, the effect of antibody is no longer significant. Our analysis treats virus and antibody as continuous variables and reaches the same conclusion. Hastie and Tibshirani (1990) analyzed the same data set using a backward stepwise procedure in generalized additive models, while fixing degrees of freedoms for antibody and virus level to be 4. They concluded that the final model contains virus level and coat color.

| | (a) First step, log likelihood = $-264.93$ | | |
|---|---|---|---|
| | Coef. | Std.Err | Z stat. |
| mhc | $-1.00e-02$ | $2.56e-01$ | $-0.0391$ |
| sex | $2.71e-01$ | $2.65e-01$ | 1.0249 |
| coat | $2.36e-01$ | $2.45e-01$ | 0.9651 |
| antibody | $-1.90e-02$ | $8.38e-03$ | $-2.2649$ |
| virus | $1.24e-05$ | $3.11e-05$ | 0.3981 |
| | (b) Last step, log likelihood = $-265.90$ | | |
| | Coef. | Std.Err | Z stat. |
| antibody | $-1.67e-2$ | $6.84e-3$ | $-2.44$ |

Table 6.3.1: Results of linear variable selection for mouse leukemia data.

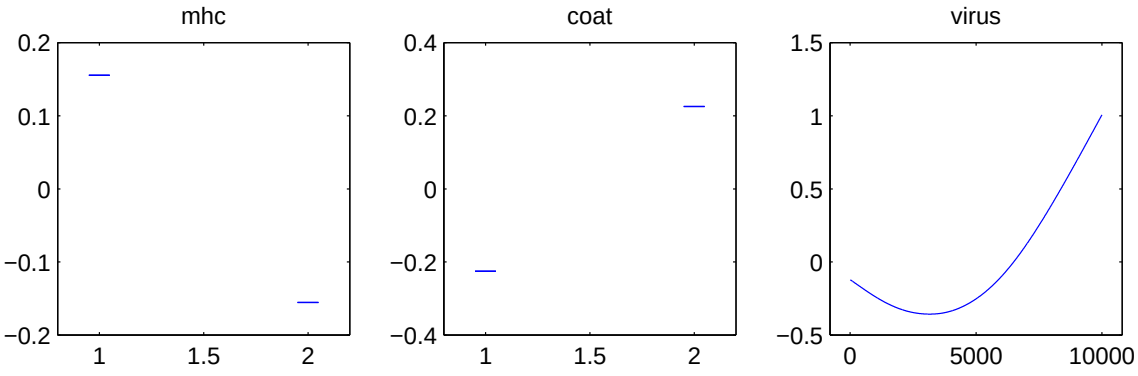


Figure 6.3.1: Fitted main effects for mouse leukemia data using our method.

# Chapter 7

# Conclusions and Future Research

The Cosso penalty for Gaussian SS-ANOVA models is used for model selection purpose in nonparamteric Cox's proportional hazard models. The connection of our proposal to the usual SS-ANOVA models is established. We further propose an approximate leave-out-one cross validation criterion $ACV$ for the selection of the smoothing parameter. We have shown that the method has attractive properties for model selection and component estimation.

In deriving the leave-out-one cross validation criterion, we did not account for the variability in estimating $\theta$. It remains a task to investigate how to incorporate the variability into the cross validation criterion. Unlike their counterparts in the linear cases, the confidence argument is not established. These are of great interest for future research.

Our procedure assumes the Cox's proportional hazard model, which is the most popular model for studying survival data. However, the proportionality assumption may not hold for every time-to-event data set. To check the Cox

assumption, the procedures in Lin, Wei and Ying (1993) can be used. Alternatively, we may consider accelerated failure time models analogous to the classical linear regression approach. In this approach, the log of the survival time is a regression function of the covariates as follows

$$\log(T) = \mu + \eta(X) + \sigma W,$$

where $\mu$ is a constant and $W$ is the error distribution. Another direction of research is to consider the hazard function as

$$h(t|x) = \eta(t, x),$$

of which the Cox model is a special case. In this model, smooth estimate of the baseline hazard is incoporated into the model, and time covariate interaction can be explored in the functional decomposition. Chapter 7 of Gu (2002) contains discussion of this estimate in the usual SS-ANOVA models. The extension of our proposal to multivariate survival data where multiple survival times are correlated is under way.

To summarize, the nonparametric model selection for time-to-event data can be very useful for selecting important risk factors.

# Chapter 8

# Part II - Penalized Log Likelihood Density Estimation

This part has appeared in Wahba, Lin and Leng (2002) . Gu and Wang (2003) studied the same problem via SS-ANOVA models.

In this part, we will examine a penalized likelihood method for the (log) density estimation problem. It is based on solving a variational problem in an infinite dimensional (Hilbert) space, where the problem has a Bayesian flavor, and where the solution to the variational problem is (essentially) known to lie in a particular $n$ dimensional subspace. Then the smoothing parameter(s) are chosen by a predictive loss criteria. If the penalty functional is square integral second derivative, the $n$-dimensional subspace is spanned by a basis of cubic splines with knots at the observation points.

We will discuss the extension to several dimensions via a smoothing spline ANOVA (SS-ANOVA) model. We briefly demonstrate a three dimensional result. The conceptual extension of the penalized likelihood method to higher dimensions is fairly straightforward, and the real thrust of the work is to be able to estimate densities in higher dimensions. One of the rationales behind

the use of the SS-ANOVA model for density estimation in several dimensions is that the pattern of main effects and interactions has an interesting interpretation in terms of conditional dependencies, and can thus be used to fit graphical models (Darroch and Lauritzen and Speed (1980), Whittaker (1990), Jordan (1998)) nonparametrically.

## 8.1 Introduction

Our density estimate is based on the penalized log likelihood estimate of Silverman (1982). When going to higher dimensions we will use the basic ANOVA decomposition idea in Gu (1993) . Our density estimate will have compact support $\Omega$, which will be scaled to the unit interval or the unit cube in $E^d$ and then rescaled back after fitting. Let the density $p = e^g$ with $g$ in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with square seminorm $J(g)$ , where the null space of $J$ contains the constant function and is low dimensional. Letting $x_i \in \Omega$, Silverman showed that the penalized log likelihood minimization problem: min $g \in \mathcal{H}$

$$-\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \lambda J(g) \tag{8.1.1}$$

subject to the condition

$$\int_{\Omega} e^g = 1 \tag{8.1.2}$$

is the same as the minimizer of

$$\mathcal{I}_\lambda(g) = -\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int_{\Omega} e^g + \lambda J(g). \tag{8.1.3}$$

We will describe the estimate in general form so that its extension from the univariate to the multivariate case is clear. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is the null space of $J$, and let the reproducing kernel for $\mathcal{H}_1$ be $K(x, x')$. If the term $\int_\Omega e^g$ were not in (8.1.3), then (it is well known that) the minimizer of (8.1.3) would be in $\mathcal{H}^n \equiv \mathcal{H}_0 \oplus span\{\xi_i, i = 1, \cdots, n\}$, where $\xi_i(x) = K(x, x_i)$. ($\xi_i$ is known as a representer.) We will therefore feel confident that the minimizer of (8.1.3) in $\mathcal{H}^n$ is a good approximation to the minimizer of (8.1.3) in $\mathcal{H}$. In fact, we will seek a minimizer in $\mathcal{H}^N = \mathcal{H}_0 \oplus span\{\xi_{i_r}, r = 1, \cdots N\}$ where the $i_r$ is a representative subset chosen sufficiently large that the minimizer in $\mathcal{H}^N$ is a good approximation to the minimizer in $\mathcal{H}^n$.

## 8.2 Choosing the Smoothing Parameter

In order to carry out penalized log likelihood estimation a method for choosing $\lambda$ is required. We have obtained a randomized Generalized Approximate Cross Validation ($ranGACV$) estimate for $\lambda$, for density estimation. We briefly describe it here, details will be given elsewhere. Let $f_\lambda$ be the estimate of the log density, and let $f_\lambda^{[-i]}(x_i)$ be the estimate with the $i$th observation left out. Define the ordinary leaving-out-one function as

$$V_0(\lambda) = OBS(\lambda) + D(\lambda) \tag{8.2.1}$$

where

$$OBS(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} f_\lambda(x_i) \tag{8.2.2}$$

and

$$D(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[f_\lambda(x_i) - f_\lambda^{[-i]}(x_i)]. \tag{8.2.3}$$

Elsewhere (to appear) we show that $nD(\lambda)$ can be approximated by the trace of the inverse Hessian of $\mathcal{I}_\lambda$ with respect to $f_\lambda(x_i), i = 1, \cdots, n$ and that it can be estimated by a randomization technique as follows. Let $\mathcal{I}_\lambda(g, y)$ be

$$\mathcal{I}_\lambda(g, y) = -\frac{1}{n}\sum_{i=1}^{n} y_i g(x_i) + \int_\Omega e^g + \lambda J(g). \tag{8.2.4}$$

When $y = (1, \cdots, 1)'$ then (8.2.4) becomes (8.1.3).

Letting $f_\lambda^y$ be the minimizer of (8.2.4), $D(\lambda)$ is estimated as

$$\hat{D}(\lambda) = \frac{1}{n\sigma_\epsilon^2}\epsilon'(f_\lambda^{y+\epsilon} - f_\lambda^y) \tag{8.2.5}$$

where $y = (1, \cdots, 1)'$, $\epsilon$ is a random vector with mean 0 and covariance $\sigma_\epsilon^2 I$, and, with some abuse of notation $f_\lambda^z = (f_\lambda^z(x_1), \cdots, f_\lambda^z(x_n))'$. Several replicates in $\epsilon$ may be used for greater accuracy. Then

$$ranGACV(\lambda) = OBS(\lambda) + \hat{D}(\lambda). \tag{8.2.6}$$

Our numerical results (to appear) show that $ranGACV$ is a good proxy for the comparative Kullback Liebler distance between the density determined by $f_\lambda$ and the true density.

## 8.3 Algorithm

The procedure is to start with $N$ representers. In the one-dimensional case we choose roughly equally spaced order statistics. Fix $\lambda$ large. Use a Newton

Raphson iteration to estimate the coefficients of $f_\lambda$ in the basis functions spanning $\mathcal{H}^N$. Evaluate $ranGACV(\lambda)$. Decrease $\lambda$ and repeat, until the minimizer over $\lambda$ is found. Double $N$ and repeat. Compare the resulting estimates with $N$ and $2N$, if they agree within a specified tolerance, stop, otherwise double $N$ again. We tried this penalized log likelihood estimate on the examples in HK, using $\mathcal{H} = W_2^2 \equiv \{g : g, g' abs.cont., g'' \in \mathcal{L}_2\}$ and $J(g) = \int_0^1 (g''(x))^2$. In this case $\mathcal{H}_0$ is spanned by linear functions and $K(x, x') = k_2(x)k_2(x') - k_4([x - x'])$, $x \in [0, 1]$ where $[\tau]$ is the fractional part of $\tau$ and $k_m(x) = B_m(x)/x!$ where $B_m$ is the $m$th Bernoulli polynomial. The estimate is a cubic spline (Wahba (1990)) with knots at the $x_{i_r}$.

## 8.4 Multivariate Smoothing Spline ANOVA Density Estimation

The univariate penalized log likelihood density estimation procedure we have described can be generalized to the multivariate case in various ways. Here we describe the smoothing spline ANOVA (SS-ANOVA) model. The use of SS-ANOVA in a density estimate was suggested by Gu (1993), who also gave a method for choosing the smoothing parameter(s). It can be shown that (for the same smoothing parameters) the estimates of Gu and Silverman are mathematically equivalent, however we found the variational problem in Silverman

easier to compute. The problem in $d$ dimensions is transformed to the $d$-dimensional unit cube, and $x_i = (x_{i1}, \cdots, x_{id})$. $\mathcal{H}$ will be an RKHS on the $d$ dimensional cube which is formed as the direct sum of subspaces of the tensor product of $d$ one dimensional RKHS's. Details of SS-ANOVA models may be found in Wahba (1990), Wahba, and *el at* (1995) Lin, and *el at* (2000). Letting $u = (u_1, \cdots, u_d) \in [0,1]^d$, we have

$$g(u) = \mu + \sum_{\alpha=1}^{d} g_\alpha(u_\alpha) + \sum_{\alpha \neq \beta} g_{\alpha\beta}(u_\alpha, u_\beta) + ... \tag{8.4.1}$$

where the terms satisfy averaging conditions analogous to those in ordinary ANOVA that insure identifiability, and the series may be truncated somewhere. The interesting feature of this representation of a log density is the fact that the presence or absence of interaction terms determines the conditional dependencies, that is, a graphical model, see Whittaker (1990). For example the main effects model represents independent component random variables, and if, for example $d = 3$ and the $g_{23}$ and $g_{123}$ terms are missing then the second and third component random variables are conditionally independent, given the first.

Let $\tilde{\mathcal{H}}$ be the $d$-fold tensor product of $W_2^2$ and let $\mathcal{H}$ be the subspace of $\tilde{\mathcal{H}}$ consisting of the direct sum of subspaces containing the terms retained in the expansion. (They are orthogonal in $\tilde{\mathcal{H}}$) We have $\int_0^1 g_\alpha(u_\alpha) du_\alpha = 0$, and so forth. The penalty functional $J(g)$ of (8.1.3) becomes $J_\theta(g)$ where the $\theta$ represents a vector of (relative) weights on separate penalty terms for each of the components of (8.4.1). As before $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0$ is the (low dimensional) null space of $J_\theta$. Let $K_\theta(x, x'), x, x' \in [0,1]^d$ be the reproducing

kernel for $\mathcal{H}_1$ where $\theta$ has been incorporated into the norm on $\mathcal{H}_1$. (See Wahba (1990) Chapter 10.) Let $\xi_i(x) = \xi_{i\theta}(x) = K_\theta(x, x_i)$. The same arguments hold as in the one dimensional case, and we seek a minimizer of (8.1.3) (with $J = J_\theta$) in $\mathcal{H}^N = \mathcal{H}_0 \oplus span\{\xi_{i_r\theta}, r = 1, \cdots, N\}$, and $\lambda$ and $\theta$ are chosen using the $ranGACV$ of (8.2.6).

## 8.5 A 3-dimenional Example

We will give a three dimensional example, essentially to demonstrate that the calculations are possible and the $ranGACV$ reasonable in higher dimensions. The SS-ANOVA model for this example contained only the main effects and two factor interactions, and we had altogether 6 smoothing parameters, parameterized in a convenient manner. For fixed smoothing parameters $\lambda, \theta$ the coefficients in the expansion in $\mathcal{H}^N$ are obtained via a Newton-Raphson iteration. In this case integrations over $[0, 1]^3$ are required, and we used quadrature formulae based on the hyperbolic cross points, see Novak and Ritter (1996), Wahba (1978). These quadrature formulae seem particularly appropriate for SS-ANOVA models and make high dimensional quadrature feasible. Then the $ranGACV$ was minimized over smoothing parameters via a 6-dimensional downhill simplex calculation.

The underlying true density used in the example is $p(x) = 0.5N(\mu_1, \Sigma) +$

$0.5N(\mu_2, \Sigma)$, where $\mu_1 = (0.25, 0.25, 0.25)$, $\mu_2 = (0.75, 0.75, 0.75)$,

$$\Sigma = \begin{pmatrix} 10 & 0 & 10 \\ 0 & 20 & 30 \\ 10 & 30 & 80 \end{pmatrix}^{-1} = \begin{pmatrix} 0.14 & 0.06 & -0.04 \\ 0.06 & 0.14 & -0.06 \\ -0.04 & -0.06 & 0.04 \end{pmatrix}.$$

(This density has a non zero three factor interaction which is not in our two factor model.) In this example the sample size was $n = 1000$. $N = 40$ and the 40 representers were randomly chosen from among the $n$ possibilities. The $N = 80$ estimate was essentially indistinguishable from the $N = 40$ case. (Note that the smoothing parameters will not generally be the same in the two cases.) Figure 8.5.1 gives cross sections of the true density, and Figure 8.5.2 gives the SS-ANOVA penalized log likelihood estimate. Figure 8.5.3 compares the $ranGACV$ and the $CKL$ ($CKL(\lambda) = -\int_\Omega f_{\lambda,\theta}(u)p(u)du$) as a function of iteration number in a downhill simplex minimization of the $ranGACV$.

## 8.6 Closing Remarks

We have proposed a penalized likelihood density estimate with $ranGACV$ to choose the smoothing parameter(s). We have shown that these penalized likelihood estimates can be extended to the multivariate case (work in progress). It remains to develop tests to allow the construction of graphical models from the SS-ANOVA estimates in higher dimensions. Alternatively, we can use the Cosso penalty to develop component selection procedures for the log density estimation.

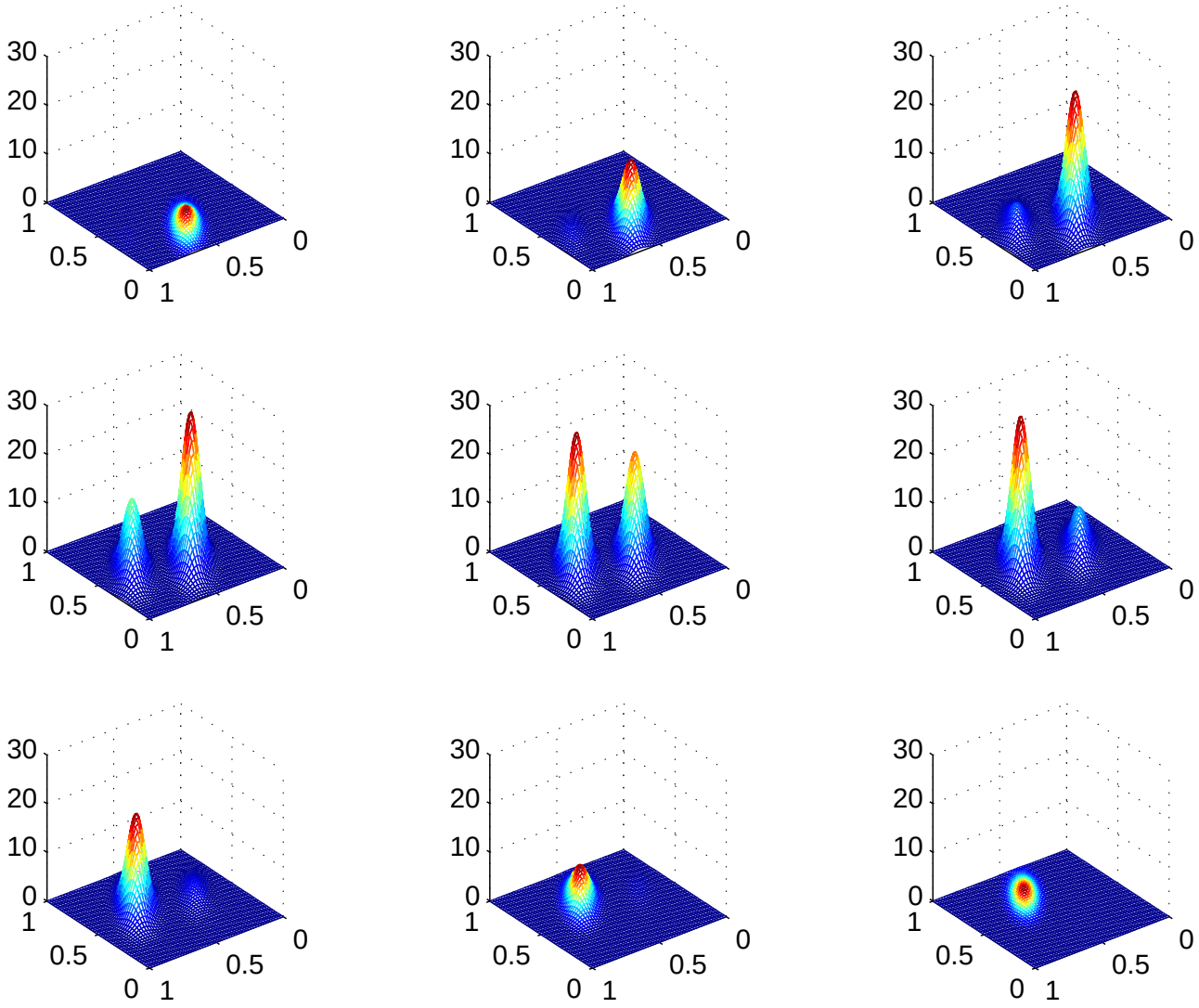


Figure 8.5.1: The true density. $x_1 = .1, ..., .9.$ is fixed in the plots, left to right, then top to bottom.

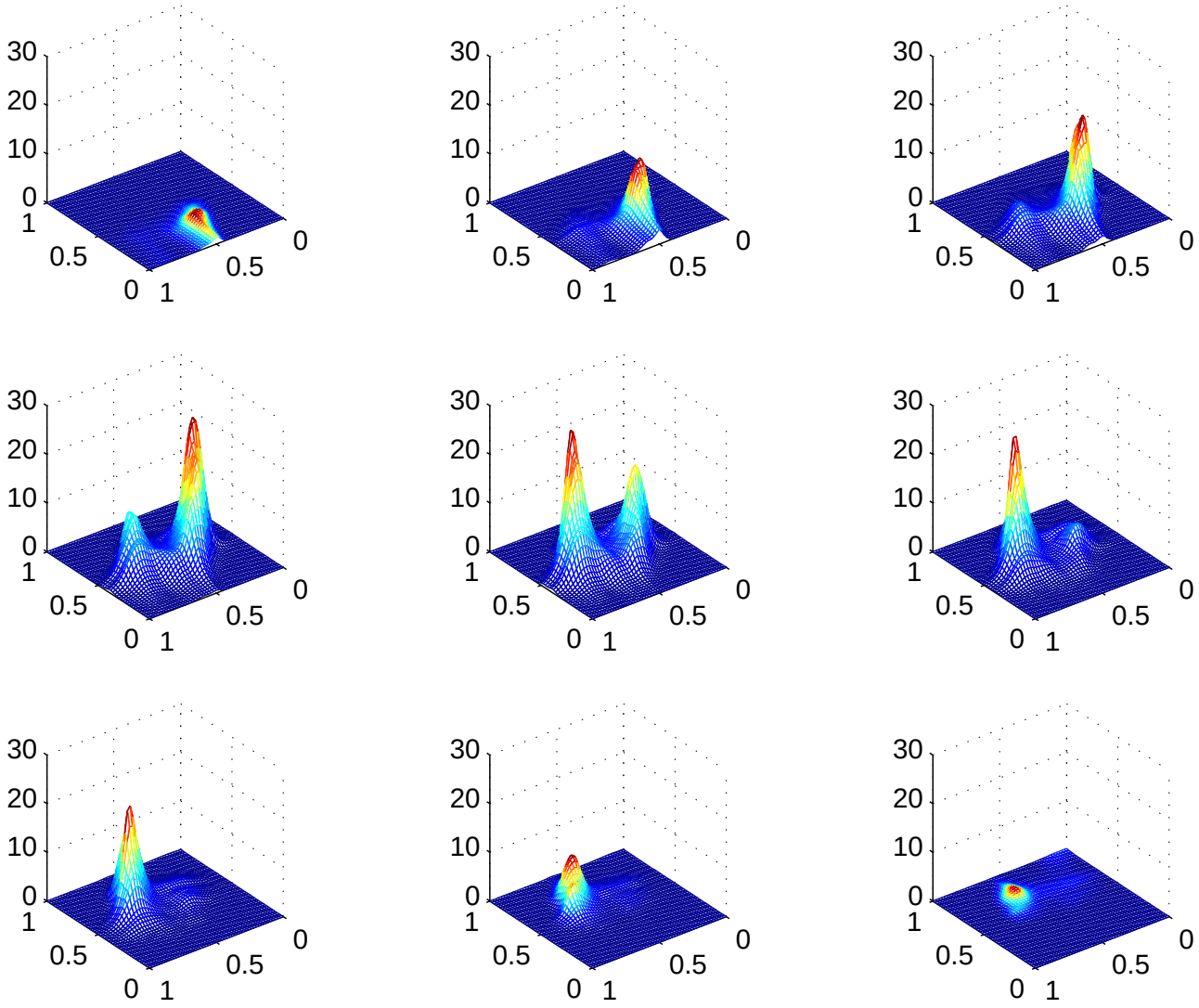


Figure 8.5.2: The estimated density. $x_1 = .1, ..., .9$ is fixed in the plots, left to right, then top to bottom.

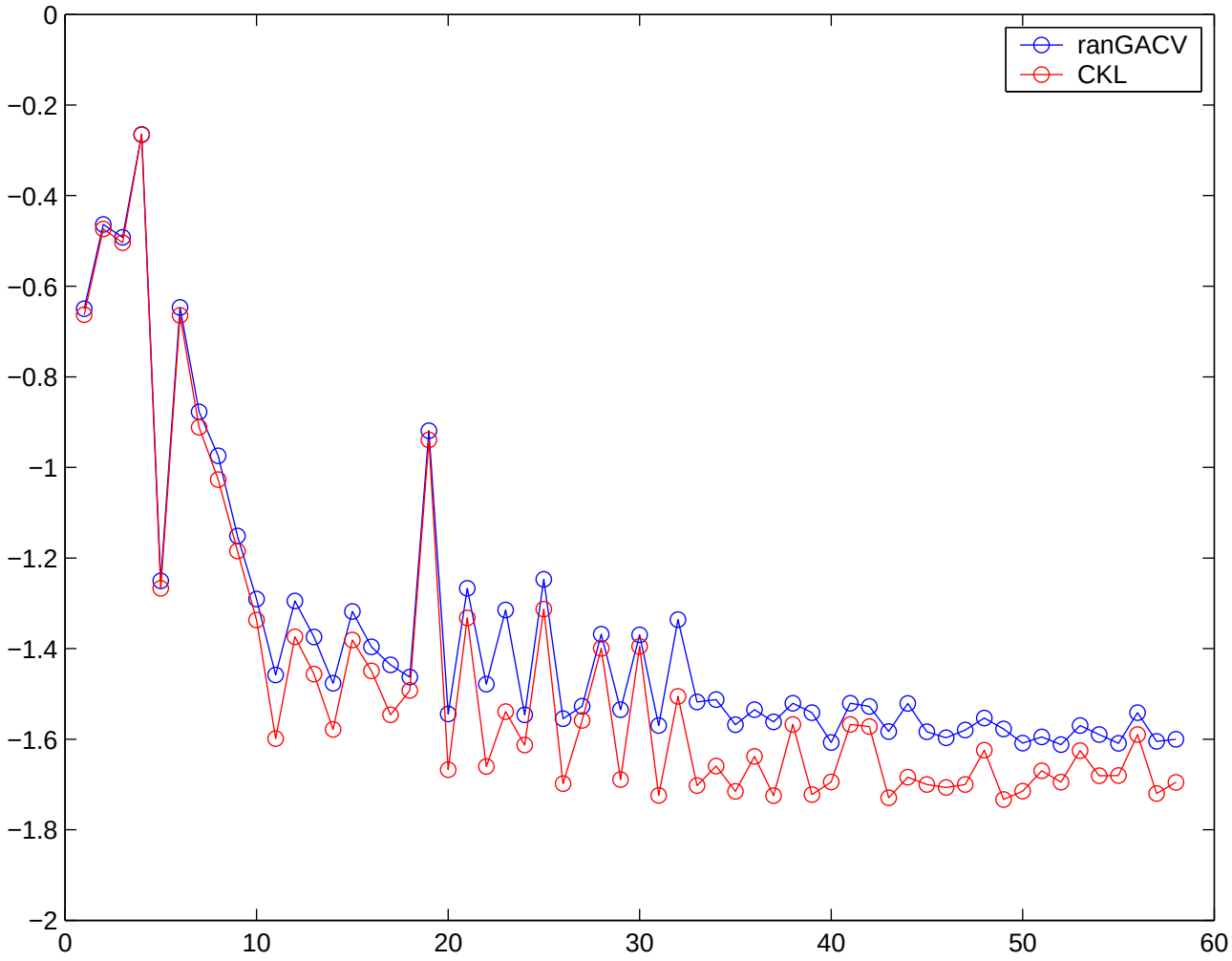


Figure 8.5.3: The *ranGACV* and the *CKL* compared. The horizontal axis is iteration number, using the downhill simplex method. The *ranGACV* is minimized and the *ranGACV* and *CKL* are computed at the minimizer at each step.

# Chapter 9

# Part III - Consistency of Selected Linear Model Selection Techniques

In this part, we study the consistency of several popular model selection techniques in linear model selection, including the highly celebrated Lasso, and two related procedures, the forward stagewise selection and a newly proposed algorithm, the Lars. All these methods are shrinkage type of methods which can give a sequence of models. We show that these methods are not consistent if a prediction based tuning criterion is used.

## 9.1 Introduction

The Least Absolute Shrinkage and Selection Operator (the Lasso) proposed by Tibshirani (1996) is a popular technique for model selection and estimation in linear regression models. It employs an $L_1$ type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable

selection tool as in Tibshirani (1997), Osborne, Presnell and Turlach (2000). Knight and Fu (2000) studied the asymptotic properties of Lasso-type estimators. They showed that under appropriate conditions, the Lasso estimators are consistent for estimating the regression coefficients, and the limit distribution of the Lasso estimators can have positive probability mass at 0 when the true value of the parameter is 0. It has been demonstrated in Tibshirani (1996) that the Lasso is more stable and accurate than the traditional variable selection methods such as the best subset selection. Efron, Hastie, Johnstone and Tibshirani (2004) proposed the Least Angle Regression (the Lars), and showed that there is a close connection between the Lars, the Lasso, and another model selection procedure called the Forward Stagewise regression. Each of these procedures involves a tuning parameter that is chosen to minimize the prediction error. This paper is concerned with the properties of the resulting estimators in terms of variable selection.

Consider the common Gaussian linear regression model

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

where $\mathbf{y} = (y_1, ..., y_n)^T$ are the responses, $\beta = (\beta_1, ..., \beta_d)^T$ are the regression coefficients, $X = (\mathbf{x}_1, ..., \mathbf{x}_d)$ is the covariate matrix, and $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim N(0, \sigma^2 I_n)$ are the normal noises. Without loss of generality, throughout this paper we assume that the covariates have been standardized to mean 0 and variance 1, and the response has mean 0. That is,

$$\mathbf{1}^T\mathbf{y} = 0,\ \mathbf{1}^T\mathbf{x}_j = 0,\ \text{and } \mathbf{x}_j^T\mathbf{x}_j = 1 \text{ for } j = 1, ..., d.$$

In many practical situations, some covariates are superfluous. That is, only a proper subset of the regression coefficients are nonzero. The problem of variable selection is to identify this set of important covariates. A variable selection procedure is said to be consistent, if the probability that the procedure correctly identifies the set of important covariates approaches one when the sample size $n$ goes to infinity. See, for example, Rao and Wu (1989) and Shao (1997) for some earlier studies on the consistent variable selection problem.

It is of interest to investigate whether the Lasso and related methods are consistent in terms of variable selection as they are often used as variable selectors. Tibshirani (1996) noted in one of the simulation examples, that in the majority of the runs the Lasso chose models that contain the true model, but only in a small fraction of runs did the Lasso pick the correct model. Fan and Li (2001) studied the penalized likelihood methods in linear regression, of which the Lasso is a special case. They proposed a nonconcave penalized likelihood method that enjoys the oracle property when the tuning parameter is appropriately chosen. The nonconcave penalized likelihood method is consistent in terms of variable selection, and it estimates the nonzero regression coefficients as well as when the correct submodel is known. They conjectured that the Lasso does not enjoy the oracle property. In this paper we show that when the tuning parameter is chosen to minimize the prediction error, as is commonly done in practice, in general the Lasso and related procedures are not consistent variable selectors. In particular, we show that when there are superfluous variables in

the linear regression model and the design matrix is orthogonal, the probability of the procedures correctly identifying the true set of important variables is less than a constant (smaller than one) not depending on $n$. This implies the inconsistency for model selection but is actually much stronger. It is a finite sample result, since it is true for any sample size $n$.

The remaining part of this article is organized as follows. In section 9.2, we review the Lasso, the Lars and the Forward Stagewise regression. In section 9.3, we give a simple example to illustrate the ideas and demonstrate that the three methods fail to find the right model with certain probability. The general results are given in section 9.4. We present some simulation results in section 9.5 and a summary is given in section 9.6.

## 9.2 The Lasso, the Lars and the Forward Stagewise Regression

The Lasso estimate is the solution to

$$\min_{\beta} (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta), \quad \text{s.t. } \sum_{j=1}^{d} |\beta_j| \leq t.$$

Here $t \geq 0$ is a tuning parameter. Let $\hat{\beta}^0$ be the ordinary least square (OLS) estimate and $t_0 = \sum |\hat{\beta}_j^0|$. Values of $t < t_0$ will shrink the solutions toward 0. As shown in Tibshirani (1996), the Lasso gives sparse interpretable models and has excellent estimation accuracy. Equivalently, the Lasso estimate can be

obtained as the solution to the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) + \lambda \sum_{j=1}^{d} |\beta_j|, \tag{9.2.1}$$

where there is a one to one correspondence between $t$ and $\lambda$.

The Forward Stagewise regression, which will be called the FSW hereafter, is an iterative procedure, where successive estimates are built via a series of small steps. Letting $\eta = X\beta$, and beginning with $\hat{\eta}_0 = 0$, if $\hat{\eta}$ is the current estimate, the next step is taken in the direction of the greatest correlation between covariate $\mathbf{x}_j$ and the current residual. That is, writing $\hat{\mathbf{c}} = X^T(\mathbf{y} - \hat{\eta})$ and $\hat{j} = \text{argmax}|\hat{c}_j|$, the update is

$$\hat{\eta} \leftarrow \hat{\eta} + \epsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}},$$

where $\epsilon > 0$ is some small constant. It is readily seen that $\epsilon = |\hat{c}_{\hat{j}}|$ yields the familiar standard forward selection. Smaller $\epsilon$ yields less greedy algorithm for the FSW and is recommended.

The Lars is a newly proposed model selection tool. We briefly describe the procedure in the following. For a detailed account of the procedure, the readers are referred to Efron, Hastie, Johnstone & Tibshirani (2004). The algorithm begins at $\hat{\eta}_0 = 0$. Suppose $\hat{\eta}$ is the current estimate and write $\hat{\mathbf{c}} = X^T(\mathbf{y} - \hat{\eta})$. Define the active set $\mathcal{A}$ as the set of the indices corresponding to the covariates with the largest absolute correlations,

$$\hat{C} = \max_j\{|\hat{c}_j|\} \text{ and } \mathcal{A} = \{j : |\hat{c}_j| = |\hat{C}|\}.$$

Define the active matrix corresponding to $\mathcal{A}$ as

$$X_{\mathcal{A}} = (s_j \mathbf{x}_j)_{j \in \mathcal{A}}, \text{ where } s_j = \text{sign}(\hat{c}_j).$$

Let

$$G_{\mathcal{A}} = X_{\mathcal{A}}^T X_{\mathcal{A}} \text{ and } A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2},$$

where $\mathbf{1}_{\mathcal{A}}$ is a vector of ones of length being $|\mathcal{A}|$, the size of $\mathcal{A}$. A unit equiangular vector with columns of the active set matrix $X_{\mathcal{A}}$ can be defined as

$$u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}, \text{ where } w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}},$$

so that

$$X_{\mathcal{A}}^T u_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} \text{ and } ||u_{\mathcal{A}}||^2 = 1.$$

The next step of the Lars estimate gives the update

$$\hat{\eta} \leftarrow \hat{\eta} + \hat{\gamma} u_{\mathcal{A}},$$

where $\hat{\gamma}$ is the smallest positive number such that one and only one new index joins the active set $\mathcal{A}$. It can be shown that

$$\hat{\gamma} = \min^+_{j \in \mathcal{A}^C} \{ \frac{\hat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathcal{A}} + a_j} \},$$

where $\min^+$ means the minimum is taken over only positive components and $a_j$ is the $j$th component of the vector $\mathbf{a} = X_{\mathcal{A}} u_{\mathcal{A}}$.

The Lasso, the FSW and the Lars all build a sequence of candidate models, from which the final model is chosen. In the Lasso, the sequence is controlled by $t$ and in the FSW, it is controlled by the number of steps (the step size in the

procedure is taken to be a small constant arbitrarily close to zero). The Lars builds $(d+1)$ models with the number of variables ranging from 0 to $d$. Efron, Hastie, Johnstone & Tibshirani (2004) showed that there is a close relationship among these procedures in that they give almost identical solution paths. That is, if the candidate models are connected in each of these procedures, the resulting graphs are very similar. The solution path of the Lars is formed by connecting the $(d+1)$ models with linear segments. They noted that in the special case of orthogonal design matrix, the solution paths of the procedures are identical. Therefore we concentrate on the Lasso in the following, and all the results apply to the Lars and the FSW as well. In the orthogonal design matrix case, the Lasso solution has the form

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \ j = 1, ..., d, \tag{9.2.2}$$

where $\gamma = \lambda/2$ for the $\lambda$ in (9.2.1); and $(\pi)^+ = \pi, \ \pi > 0; \ 0, \ \pi \leq 0$. It coincides with the soft thresholding solution of Donoho and Johnstone (1994) , where it is applied to wavelet coefficients.

In the implementation of the Lars, it is often the case that only the $(d+1)$ models at the end of the steps are considered as candidate models. The final model is chosen among the $(d+1)$ models, not the whole solution path. In this case the Lars is slightly different from the Lasso or the FSW, even in the orthogonal design matrix case. We will treat this case separately in this article.

The implementation of the Lasso, the Lars and the FSW attempts to find a model with the smallest estimation error among the sequence of candidate

models built by these procedures. The estimation error is typically in terms of the squared loss ($SL$). For an estimate $\hat{\eta} = X\hat{\beta}$, the squared loss is

$$SL(\hat{\eta}) = (\hat{\eta} - \eta)^T(\hat{\eta} - \eta) = (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta).$$

In practice, since $\beta$ is unknown, several methods, such as generalized cross validation (Craven & Wahba 1979), k-fold cross validation or Stein's unbiased estimate of risk (Stein 1981), can be used for the purpose of minimizing the squared error.

## 9.3 A Simple Example

In this section we give a simple example to demonstrate that the Lasso, the FSW and the Lars when tuned to minimize the squared error (as people usually attempt to do), miss the right model with a certain probability.

Consider a linear regression model with two predictors. Suppose the true coefficient vector is $\beta^0 = (\beta_1^0, 0)^T$ with $\beta_1^0 > 0$, and the design matrix $X$ is orthonormal. Therefore the model has one true component $\mathbf{x}_1$ and one noisy component $\mathbf{x}_2$, and $X^T X = I_2$. Denote the ordinary least squares solution by $\hat{\beta}^0$. In this case the solution to the Lasso problem (9.2.1) is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \ j = 1, 2. \tag{9.3.1}$$

Figure 9.3.1 shows the Lasso estimate versus the OLS estimate. The Lasso estimate is shifted towards zero by some constant. For completeness, the estimate by subset selection versus the OLS estimate is included in the figure.

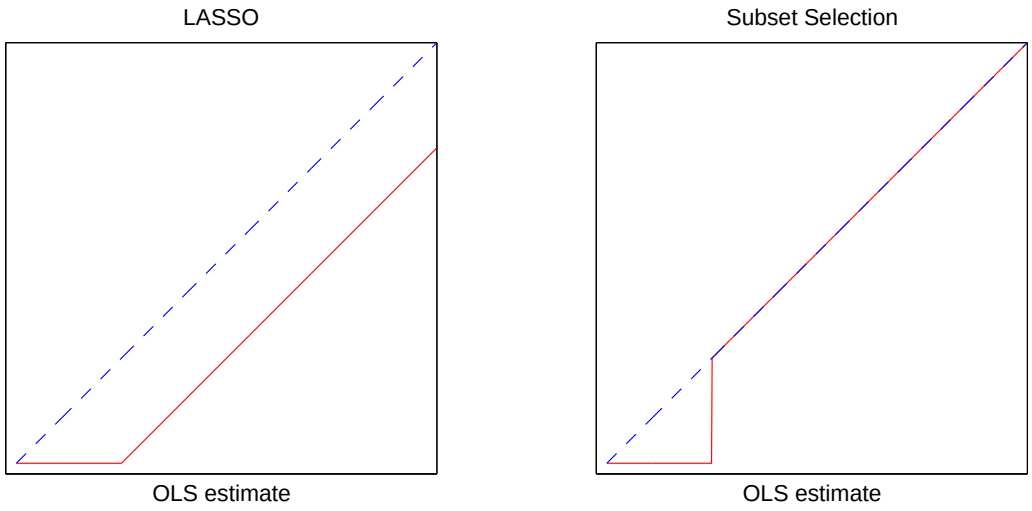


Figure 9.3.1: In the left plot, the red solid line indicates the Lasso estimate versus the OLS estimate; the right plot shows the subset estimate against the OLS estimate. For comparison, the 45 degree lines are drawn.

Let $\hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2)^T = \hat{\beta}^0 - \beta^0$. Since $\epsilon$ is a normal variate and $X^T X = I_2$, we have

$$\hat{\delta} \sim N(0, \sigma^2 I_2), \tag{9.3.2}$$

where $\sigma^2$ is the noise variance.

Define

$$\mathcal{R}_1 = \{(\delta_1, \delta_2)^T : \ \delta_1 < 0\},$$

$$\mathcal{R}_2 = \{(\delta_1, \delta_2)^T : \ \delta_1 > 0, \ \delta_1 < |\delta_2|\},$$

$$\mathcal{R}_3 = \{(\delta_1, \delta_2)^T : \ \delta_1 > 0, \ \delta_1 > |\delta_2|\}.$$

We will show that when $\hat{\delta} \in \mathcal{R}_1$ or $\hat{\delta} \in \mathcal{R}_2$, the Lasso does not select the right model. The only region where the Lasso selects the right model is $\mathcal{R}_3$. Thus

from (9.3.2), the probability of the Lasso selecting the correct model is 1/4.

It is clear that when $|\hat{\beta}_1^0| \leq |\hat{\beta}_2^0|$, the Lasso can not select the correct variables by (9.3.1). As a result, we only need to consider the situation where $|\hat{\beta}_1^0| > |\hat{\beta}_2^0|$ in the following.

For $\hat{\delta} \in \mathcal{R}_1$, we consider the situations $\hat{\beta}_1^0 \leq 0$ and $\hat{\beta}_1^0 > 0$ separately. For $\hat{\beta}_1^0 \leq 0$, when $|\hat{\beta}_1^0| > |\hat{\beta}_2^0|$, it is easy to see that a naive estimate $\hat{\eta} = 0$ with $\hat{\gamma} = |\hat{\beta}_1^0|$ yields the Lasso estimate with the smallest squared loss.

For $\hat{\beta}_1^0 > 0$, the Lasso solution minimizes

$$SL(\gamma) = (\hat{\beta}_1 - \beta_1^0)^2 + (\hat{\beta}_2 - \beta_2^0)^2 = [(\beta_1^0 + \hat{\delta}_1 - \gamma)^+ - \beta_1^0]^2 + [(|\hat{\delta}_2| - \gamma)^+]^2.$$

For $\gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, $SL(\gamma) = (\hat{\delta}_1 - \gamma)^2$. Since $\hat{\delta}_1 < 0$, we have

$$SL(\gamma) = (\hat{\delta}_1 - \gamma)^2 > \hat{\delta}_1^2 + \hat{\delta}_2^2 = SL(0),$$

where $SL(0)$ is the $SL$ of the OLS estimate. Therefore, the optimal $\gamma$ that minimizes $SL(\gamma)$ is not in the interval $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$. We then see from (9.3.1) that the optimal $\gamma$ does not yield the correct model. The claim is proved for $\hat{\delta} \in \mathcal{R}_1$.

For $\hat{\delta} \in \mathcal{R}_2$ and $\gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, $SL(\gamma) = (\hat{\delta}_1 - \gamma)^2$. Since $\hat{\delta}_1 < |\hat{\delta}_2|$, the minimum is obtained at $\gamma_1 = |\hat{\delta}_2|$ on this interval and $SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2$. However, when $\gamma_2 = (\hat{\delta}_1 + |\hat{\delta}_2|)/2 < |\hat{\delta}_2|$,

$$SL(\gamma_2) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2/2 < SL(\gamma_1).$$

The estimated coefficients corresponding to $\gamma_2$ are

$$\hat{\beta}_1 = \hat{\beta}_1^0 + (\hat{\delta}_1 - |\hat{\delta}_2|)/2 \ \text{ and } \ \hat{\beta}_2 = \text{sign}(\hat{\delta}_2)(|\hat{\delta}_2| - \hat{\delta}_1)/2 \neq 0.$$

Again, the optimal $\gamma$ that minimizes $SL(\gamma)$ is not in the interval $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$. Therefore, the Lasso does not select the right model for $\hat{\delta} \in \mathcal{R}_2$ either.

The Lasso, however, selects the right model when $\hat{\delta} \in \mathcal{R}_3$. It is easy to see that $SL(\hat{\gamma}) = 0$ and $\eta(\hat{\gamma}) = \beta_1^0 \mathbf{x}_1$ for the Lasso solution $\hat{\gamma} = \hat{\delta}_1$. Therefore, we have shown that the Lasso selects the right model only when $\hat{\delta} \in \mathcal{R}_3$. The probability associated with $\mathcal{R}_3$ is $1/4$.

The argument above is valid for any finite sample size, and shows that with probability $3/4$, the Lasso with the tuning parameter selected to minimize the estimation error does not select the correct model. The argument can be generalized to the following lemma.

**Lemma 9.3.1.** *When $\beta^0 = (\beta_1^0, 0, \cdots, 0)^T$ with $(d-1) > 0$ zero components and $X^T X = I_d$, the Lasso selects the right model only when $\hat{\delta} = \hat{\beta}^0 - \beta^0 \in \mathcal{R}$, where*

$$\mathcal{R} = \big\{\delta : \ \delta_1 \beta_1^0 > 0, \ |\delta_1| \geq max\{|\delta_2|, \cdots, |\delta_d|\}\big\},$$

*that is, the probability of the Lasso selecting the right model is $1/(2d)$.*

*Proof.* The Lasso solution has the form (9.2.2) when $X^T X = I$. Without loss of generality, assume $\beta_1^0 > 0$ and $|\hat{\beta}_2^0| > |\hat{\beta}_3^0| > \cdots > |\hat{\beta}_d^0|$. We will show for $\hat{\delta}$ not in $\mathcal{R}$, the Lasso does not select the right model. It is clear that when $|\hat{\beta}_1^0| \leq |\hat{\beta}_2^0|$, the Lasso can not select the correct variables by (9.2.2). Therefore, we concentrate on the situation where $|\hat{\beta}_1^0| > |\hat{\beta}_2^0|$ in the following.

1. For $\hat{\delta}_1 \leq 0$ and $\hat{\beta}_1^0 \leq 0$, a naive estimate $\hat{\eta} = 0$ yields the Lasso estimate.

2. For $\hat{\delta}_1 \leq 0$ and $\hat{\beta}_1^0 > 0$, for the Lasso to select the right model, $\gamma$ must satisfy $\gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$ and thus $SL(\gamma) = (\hat{\delta}_1 - \gamma)^2$. It is easy to see that the minimum is obtained at $\gamma_1 = |\hat{\delta}_2|$ and $SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2$. But when $\gamma_2 = |\hat{\delta}_3|$,

$$\begin{aligned} SL(\gamma_1) =& (\hat{\delta}_1 - |\hat{\delta}_2|)^2 = (\hat{\delta}_1 - |\hat{\delta}_3|] + |\hat{\delta}_3| - |\hat{\delta}_2|)^2 \\ =& (\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 + 2(\hat{\delta}_1 - |\hat{\delta}_3|)(|\hat{\delta}_3| - |\hat{\delta}_2|) \\ >& (\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 = SL(\gamma_2). \end{aligned}$$

   The estimated model corresponding to $\gamma_2$ is

$$\eta(\gamma_2) = (\hat{\beta}_1^0 - |\hat{\beta}_3^0|)\mathbf{x}_1 + \text{sign}(\hat{\beta}_2^0)(|\hat{\beta}_2^0| - |\hat{\beta}_3^0|)\mathbf{x}_2,$$

   which is not the right model.

3. For $0 < \hat{\delta}_1 < |\hat{\delta}_2|$, the $\gamma$ which minimizes $SL(\gamma)$ on the interval $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$ is obtained at $\gamma_1 = |\hat{\delta}_2|$ and $SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2$. However, when $|\hat{\delta}_3| < (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, if we let $\gamma_2 = (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, we have $SL(\gamma_2) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2/2 < SL(\gamma_1)$. When $|\hat{\delta}_3| \geq (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, if we let $\gamma_3 = |\hat{\delta}_3|$, we have $SL(\gamma_3) < SL(\gamma_1)$. The estimated models corresponding to $\gamma_2$ and $\gamma_3$ both include $\mathbf{x}_1$ and $\mathbf{x}_2$.

Therefore, when $\hat{\delta} \in \mathcal{R}^C$, the Lasso selects a wrong model. For $\hat{\delta} \in \mathcal{R}$, the Lasso solution $\hat{\gamma} = \hat{\delta}_1$ yields the correct model $\eta(\hat{\gamma}) = \beta_1^0 \mathbf{x}_1$ with $SL(\hat{\gamma}) = 0$. Since $\hat{\delta} \sim N(0, \sigma^2 I_d)$, we have $Pr(\hat{\delta} \in \mathcal{R}) = 1/(2d)$. This completes the proof. □

Now we return to our two dimensional example considered at the beginning of this section. Figure 9.3.2 is a schematic sketch of the Lars algorithm in this

situation. The ordinary least squares estimate is shown in the figure as point B. The initial Lars estimate is simply $\hat{\eta}_0 = 0$, corresponding to point A. The Lars estimate after step one is shown in the figure as point D, which has the property that the angle between the line DB and the axes is 45 degrees. The step two estimate is simply the ordinary least squares estimate corresponding to point B in the figure. If we consider the whole path of solutions (line segments AD and DB), and choose the estimate along the path with the smallest squared error, in our example with orthonormal design matrix, the Lars is exactly equivalent to the Lasso and the above results for the Lasso applies to the Lars directly. In practical implementation of the Lars, however, the final solution is often chosen only among the models after each complete step, that is, points A, D, and B in the figure, while in subset selection, the final solution is chosen among points A, C and B, where point C is the projection of point B to $\mathbf{x}_1$. Thus we consider this situation in the following, and study the probability of choosing the correct model (point D in this example) when the squared error is used as the criterion. In the following we show that when $\hat{\delta} \in \mathcal{R}_1$, which has probability 1/2, the Lars does not select the correct model. It is clear from the Lars algorithm that the Lars does not yield the correct model when $\hat{\beta}_1^0 \leq 0$ or $|\hat{\beta}_1^0| \leq |\hat{\beta}_2^0|$. We only need to consider the situation when $\hat{\beta}_1^0 > 0$ and $\hat{\beta}_1^0 > |\hat{\beta}_2^0|$. In this case, the Lars estimate can be written as $\hat{\eta}_0 = 0,\ \hat{\eta}_1 = (\beta_1^0 + \hat{\delta}_1 - |\hat{\delta}_2|)\mathbf{x}_1$ and $\hat{\eta}_2 = (\beta_1^0 + \hat{\delta}_1)\mathbf{x}_1 + \hat{\delta}_2\mathbf{x}_2$. It follows

$$SL(\hat{\eta}_0) = (\beta_1^0)^2,\ SL(\hat{\eta}_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2, \text{ and } SL(\hat{\eta}_2) = \hat{\delta}_1^2 + \hat{\delta}_2^2.$$

We immediately see $SL(\hat{\eta}_1) > SL(\hat{\eta}_2)$ when $\hat{\delta}_1 < 0$. Therefore, we have shown that the Lars does not select the right model when the OLS estimate satisfies $\hat{\beta}_1^0 < \beta_1^0$. The probability of this region is $Pr(\mathcal{R}) = 1/2$ since $\hat{\delta}$ is a normal variate. The overall probability that the Lars selects the right model is no larger than $1/2$.

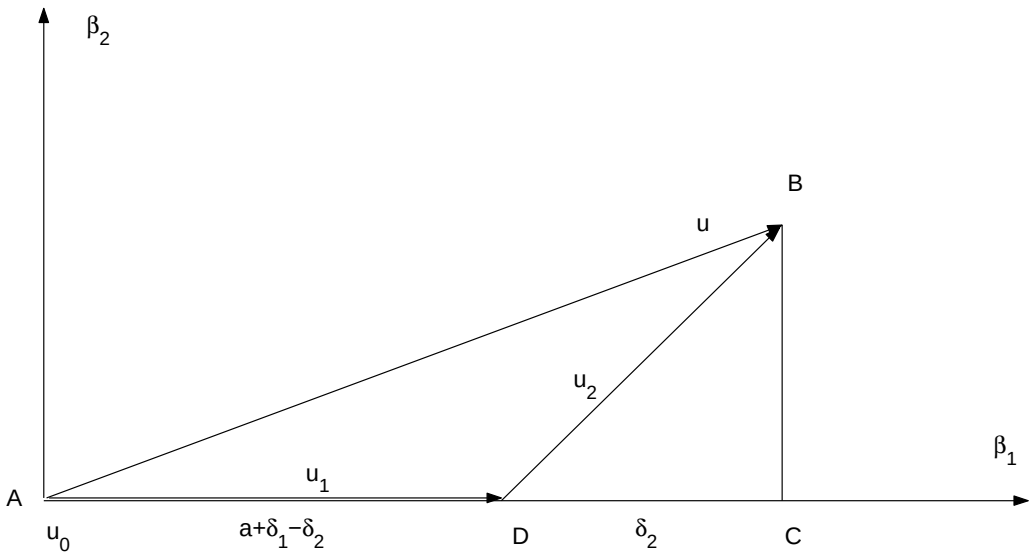


Figure 9.3.2: Lars algorithm when $d = 2$.

We prove a more general theorem in the following section.

## 9.4 More General Situations

**Theorem 9.4.1.** *When the true coefficient vector is* $\beta^0 = (\alpha_1, ..., \alpha_{d_1}, 0, ..., 0)^T$ *with* $d_2 = (d - d_1) > 0$ *zero coefficients and* $X^T X = I_d$*, we have*

$$Pr(\text{the Lasso selects the right model}) \leq C,$$

*with respect to any sample size, where* $C < 1$ *is a constant depending only on* $\sigma^2$ *and* $d_1$*.*

*Proof.* Let the OLS estimate be $\hat{\beta}^0$ and denote $\hat{\beta}^0 - \beta^0 = (\hat{\delta}_1, ..., \hat{\delta}_d)^T$. Without loss of generality we assume $|\hat{\delta}_{d_1+1}| > |\hat{\delta}_{d_1+2}| > ... > |\hat{\delta}_d|$ and $\alpha_i > 0,\ i = 1, ..., d_1$. We will show for the region

$$\mathcal{R} = \{(\delta_1, ..., \delta_d)^T :\ \delta_j > -\alpha_j,\ j = 1, ..., d_1 \text{ and } \sum_{i=1}^{d_1} \delta_i < 0\},$$

the Lasso does not select the right model.

If $\hat{\beta}^0$ does not satisfy

$$\left\{ |\hat{\beta}^0_j| > |\hat{\beta}^0_k|,\ \text{for } j \in \{1, ..., d_1\} \text{ and } k \in \{d_1 + 1, .., d\} \right\}, \tag{9.4.1}$$

obviously the Lasso does not select the right model. So we can concentrate on the situation where (9.4.1) is satisfied. For the Lasso to select the right model, the solution must satisfy

$$\min\{|\hat{\beta}^0_1|, ..., |\hat{\beta}^0_{d_1}|\} > \gamma \geq |\hat{\beta}^0_{d_1+1}|. \tag{9.4.2}$$

Since $\hat{\delta} \in \mathcal{R}$, we have $\hat{\beta}^0_j > 0,\ j = 1, ..., d_1$. The estimate corresponding to

any $\gamma$ satisfying (9.4.2) is

$$\begin{aligned}\eta(\gamma) &= (\hat{\beta}_1^0 - \gamma)\mathbf{x}_1 + ... + (\hat{\beta}_{d_1}^0 - \gamma)\mathbf{x}_{d_1} \\ &= (\alpha_1 + \hat{\delta}_1 - \gamma)\mathbf{x}_1 + ... + (\alpha_{d_1} + \hat{\delta}_{d_1} - \gamma)\mathbf{x}_{d_1}.\end{aligned}$$

On the other hand, the estimate with $\gamma_1 = |\hat{\beta}_{d_1+2}^0|$ has the form

$$\begin{aligned}\eta(\gamma_1) &= (\hat{\beta}_1^0 - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_1 + \cdots + (\hat{\beta}_{d_1}^0 - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_{d_1} \\ &\quad + \text{sign}(\hat{\beta}_{d_1+1}^0)(|\hat{\beta}_{d_1+1}^0| - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_{d_1+1} \\ &= (\alpha_1 + \hat{\delta}_1 - |\hat{\delta}_{d_1+2}|)\mathbf{x}_1 + \cdots + (\alpha_{d_1} + \hat{\delta}_{d_1} - |\hat{\delta}_{d_1+2}|)\mathbf{x}_{d_1} \\ &\quad + \text{sign}(\hat{\delta}_{d_1+1})(|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)\mathbf{x}_{d_1+1}.\end{aligned}$$

It is easy to see the squared losses for the two estimates are

$$\begin{aligned}SL(\gamma) &= \sum_{i=1}^{d_1}(\hat{\delta}_i - \gamma)^2; \\ SL(\gamma_1) &= \sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}|)^2 + (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2.\end{aligned}$$

We show for any $\gamma$ satisfying (9.4.2), $SL(\gamma) > SL(\gamma_1)$. Simple algebra yields

$$\begin{aligned}SL(\gamma) &= \sum_{i=1}^{d_1}(\hat{\delta}_i - \gamma)^2 = \sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}| + |\hat{\delta}_{d_1+2}| - \gamma)^2 \\ &= \sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}|)^2 + d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|)\sum_{i=1}^{d_1}(|\hat{\delta}_{d_1+2}| - \hat{\delta}_i) \\ &= SL(\gamma_1) - (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2 \\ &\quad + d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|)\sum_{i=1}^{d_1}(|\hat{\delta}_{d_1+2}| - \hat{\delta}_i).\end{aligned}$$

Since $\gamma \geq |\hat{\delta}_{d_1+1}|$, we have

$$d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 - (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2 \geq (d_1 - 1)(\gamma - |\hat{\delta}_{d_1+2}|)^2.$$

It follows

$$SL(\gamma) \geq SL(\gamma_1) + (d_1 - 1)(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|) \sum_{i=1}^{d_1} (|\hat{\delta}_{d_1+2}| - \hat{\delta}_i).$$

It is easy to see when $\sum_{i=1}^{d_1} \hat{\delta}_i < 0$, the following satisfies

$$(d_1-1)(\gamma-|\hat{\delta}_{d_1+2}|)+2\sum_{i=1}^{d_1}(|\hat{\delta}_{d_1+2}|-\hat{\delta}_i) = (d_1+1)|\hat{\delta}_{d_1+2}|+(d_1-1)\gamma-2\sum_{i=1}^{d_1}\hat{\delta}_i > 0.$$

Therefore, we have $SL(\gamma) > SL(\gamma_1)$ when $\hat{\delta} \in \mathcal{R}$. The optimal $\gamma$ that minimizes $SL(\gamma)$ does not satisfy (9.4.2), that is, the optimal $\gamma$ does not yield the correct model. Since $(\hat{\delta}_1, \cdots, \hat{\delta}_d)^T$ follows a multivariate normal distribution $N(0, I_d)$, it is readily seen

$$Pr(\mathcal{R}) > Pr(\{(\delta_1, ..., \delta_d) : 0 > \delta_j > -\alpha_j, j = 1, ..., d_1\}) = C,$$

where $C$ is a constant strictly less than 1 depending on $\sigma^2$ and $d_1$ but not on the sample size $n$. We have proved that with a positive probability not depending on $n$, the Lasso algorithm does not select the right model. □

The conclusion holds for the Lars and the FSW due to the equivalence of the three procedures, if the whole solution path of the Lars is considered. When only $(d + 1)$ candidate models in the Lars are considered, the conclusion follows by replacing $\gamma$ by $|\hat{\delta}_{d_1+1}|$ in the preceding proof. When the design matrix satisfies $X^T X = nI_d$, following the same argument in theorem 9.4.1, we can prove that the probability of the Lasso selecting the wrong model is larger than a strictly positive constant not depending on $n$.

Although the conclusion of the theorem is proved with the design matrix being orthonormal, it is expected to hold for general design matrix cases, as shown in the simulation in the next section.

## 9.5 Simulations

We conduct some simple simulations in general design matrix cases to demonstrate that the Lasso is not consistent in terms of model selection, when the prediction error is to be minimized. All simulations were conducted using MATLAB code. We used the algorithm as suggested in Tibshirani (1996). Each $\beta_j$ is rewritten as $\beta_j^+ - \beta_j^-$, where $\beta_j^+$ and $\beta_j^-$ are nonnegative. We then used the quadratic programming module quadprog in MATLAB to find the Lasso solution.

We generate data from two models which have the form

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

where

$$\text{Model 1}: \ \beta = (1, 0)^T,$$

$$\text{Model 2}: \ \beta = (3, 1.5, 0, 0)^T;$$

$\epsilon$ follows standard normal distribution and $\mathbf{x}_j$ has marginal distribution $N(0, 1)$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$, is $\rho$ with $\rho = 0,\ 0.5,\ 0.9$. We simulate data with sample size $n = 40,\ 400,\ 4000$. For each $\rho$ and each sample

size, we simulate 100 data sets and apply the Lasso method. We summarize the result for various sample sizes and correlations in Table 9.5.1. The percentage of correctly selected models is summarized in the PCM column. We see that the Lasso misses the right model a large fraction of the time, and this is independent of the sample size and the correlation. The results of the experiment are consistent with the previous conclusion.

| n | $\rho$ | Model 1 PCM (%) | Model 2 PCM (%) |
|---|---|---|---|
| | 0 | 26 | 15 |
| 40 | 0.5 | 16 | 20 |
| | 0.9 | 22 | 16 |
| | 0 | 27 | 9 |
| 400 | 0.5 | 23 | 15 |
| | 0.9 | 25 | 13 |
| | 0 | 22 | 15 |
| 4000 | 0.5 | 21 | 18 |
| | 0.9 | 24 | 20 |

Table 9.5.1: Simulation results for the Lasso.

## 9.6 Conclusion

We have showed in this paper that the Lasso, the Lars and the FSW are not consistent in terms of model selection when a prediction based criterion is used to select the tuning parameters, and there are superfluous variables in the model.

We remark that our results should not be taken to imply that the Lasso

and related methods can not be used as variable selection tools. What our results imply is that the dual goal of accurate estimation and consistent variable selection can not be achieved simultaneously by these methods. The common practice in applying these methods is to choose the tuning parameter to minimize the prediction error, our results state that in this case the procedures are not consistent in terms of variable selection. It is possible that some other criteria of choosing the tuning parameter can yield consistent variable selection for these methods.

# Bibliography

Aalen, O. (1978), 'Nonparametric inference for a family of counting processes', *The Annals of Statistics* **6**, 701–726.

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Trans. Amer. Math. Soc.* **68**, 337–404.

Breiman, L. (1995), 'Better subset regression using the nonnegative garrote', *Technometrics* **37**, 373–384.

Breslow, N. (1974), 'Covariance analysis of censored survival data', *Biometrics* **30**, 89–99.

Cox, D. R. (1972), 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B, Methodological* **34**, 187–220.

Cox, D. R. (1975), 'Partial likelihood', *Biometrika* **62**, 269–276.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **31**, 377–403.

Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980), 'Markov fields and log-linear interaction models for contingency tables', *The Annals of Statistics* **8**, 522–539.

Dickson, E., Grambsch, P., Fleming, T., Fisher, L. & Langworthy, A. (1989), 'Prognosis in primary biliary cirrhosis: model for decision making', *Hepatology* **10**, 1–7.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *The Annals of Statistics* **32**, 000–???

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. & Li, R. (2002), 'Variable selection for Cox's proportional hazards model and frailty model', *The Annals of Statistics* **30**(1), 74–99.

Fleming, T. R. & Harrington, D. P. (1991), *Counting processes and survival analysis*, John Wiley and Sons.

Gao, F., Wahba, G., Klein, R. & Klein, B. (2001), 'Smoothing spline Anova for multivariate Bernoulli observations with application to ophthalmology data', *Journal of the American Statistical Association* **96**(453), 127–160.

Gray, R. J. (1992), 'Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis', *Journal of the American Statistical Association* **87**, 942–951.

Gray, R. J. (1994), 'Spline-based tests in survival analysis', *Biometrics* **50**, 640–652.

Gu, C. (1992), 'Diagnostics for nonparametric regression models with additive term', *Journal of the American Statistical Association* **87**, 169–179.

Gu, C. (1993), 'Smoothing spline density estimation: A dimensionless automatic algorithm', *Journal of the American Statistical Association* **88**, 495–504.

Gu, C. (1996), 'Penalized likelihood hazard estimation: A general procedure', *Statistica Sinica* **6**, 861–876.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer-Verlag.

Gu, C. & Wahba, G. (1991), 'Minimizing Gcv/gml scores with multiple smoothing parameters via the Newton method', *SIAM Journal on Scientific and Statistical Computing* **12**, 383–398.

Gu, C. & Wang, J. (2003), 'Penalized likelihood density estimation: Direct cross-validation and scalable approximation', *Statistica Sinica* **13**, 811–826.

Gunn, S. & Kandola, J. (2002), 'Structural modelling with sparse kernels', *Machine Learning* **48**, 137–163.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman & Hall Ltd.

Jordan, M. I. E. (1998), *Learning in Graphical Models*, Kluwer Academic.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The statistical analysis of failure time data*, John Wiley and Sons.

Kim, Y.-J. & Gu, C. (2004), 'Smoothing spline gaussian regression: more scalable computation via efficient approximation', *Journal of the Royal Statistical Society Series B* **66**(2), 337–356.

Klein, J. P. & Moeschberger, M. L. (1997), *Survival analysis: techniques for censored and truncated data*, Springer-Verlag Inc.

Knight, K. & Fu, W. (2000), 'Asymptotics for lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378.

Kooperberg, C., Stone, C. J. & Truong, Y. K. (1995), 'Hazard regression', *Journal of the American Statistical Association* **90**, 78–94.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993), 'Checking the Cox model with cumulative sums of martingale-based residuals', *Biometrika* **80**, 557–572.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (2000), 'Smoothing spline Anova models for large data sets with Bernoulli observations and the randomized Gacv', *The Annals of Statistics* **28**(6), 1570–1600.

Lin, Y. (2000), 'Tensor product space ANOVA models', *The Annals of Statistics* **28**(3), 734–755.

Lin, Y. & Zhang, H. (2002), Component selection and smoothing in smoothing

spline analysis of variance model, Technical Report 1072, University of Wisconsin, Madison.

Nelson, W. (1972), 'Theory and applications of hazard plotting for censored failure data', *Technometrics* **14**, 945–966.

Novak, E. & Ritter, K. (1996), 'High dimensional integration of smooth functions over cubes', *Numerical Mathematics* **75**, 79–97.

Osborne, M. R., Presnell, B. & Turlach, B. A. (2000), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

O'Sullivan, F. (1988*a*), 'Fast computation of fully automated log-density and log-hazard estimators', *SIAM Journal on Scientific and Statistical Computing [Formerly: SIAM Journal on Scientific Computing]* **9**(2), 363–379.

O'Sullivan, F. (1988*b*), 'Nonparametric estimation of relative risk using splines and cross-validation', *SIAM Journal on Scientific and Statistical Computing [Formerly: SIAM Journal on Scientific Computing]* **9**, 531–542.

O'Sullivan, F. (1993), 'Nonparametric estimation in the Cox model', *The Annals of Statistics* **21**, 124–145.

Rao, C. R. & Wu, Y. (1989), 'A strongly consistent procedure for model selection in a regression problem', *Biometrika* **76**, 369–374.

Ruppert, D. & Carroll, R. J. (2000), 'Spatially-adaptive penalties for spline fitting', *The Australian and New Zealand Journal of Statistics* **42**(2), 205–223.

Shao, J. (1997), 'An asymptotic theory for linear model selection (Disc: P243-264)', *Statistica Sinica* **7**, 221–242.

Silverman, B. W. (1982), 'On the estimation of a probability density function by the maximum penalized likelihood method', *The Annals of Statistics* **10**, 795–810.

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

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling survival data: extending the Cox model*, Springer-Verlag Inc.

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

Tibshirani, R. (1997), 'The lasso method for variable selection in the Cox model', *Statistics in Medicine* **16**, 385–395.

Wahba, G. (1978), Interpolating surfaces: High order convergence rates and their associated designs, with applications to x-ray image reconstruction, Technical Report 523, Dept. of Statistics, University of Wisconsin, Madison, WI.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics.

Wahba, G., Lin, Y. & Leng, C. (2002), 'Comment on "Spline adaptation in extended linear models"', *Statistical Science* **17**(1), 33–37.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline Anova for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *The Annals of Statistics* **23**, 1865–1895.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley.

Wood, S., Kohn, R., Shively, T. & Jiang, W. (2002), 'Model selection in spline nonparametric regression', *Journal of the Royal Statistical Society, Series B, Methodological* **64**(1), 119–139.

Xiang, D. & Wahba, G. (1996), 'A generalized approximate cross validation for smoothing splines with non-Gaussian data', *Statistica Sinica* **6**, 675–692.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2002), Variable selection and model building via likelihood basis pursuit, Technical Report 1059, University of Wisconsin, Madison.

Zucker, D. M. & Karr, A. F. (1990), 'Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach', *The Annals of Statistics* **18**, 329–353.