

To the Graduate Council:

I am submitting herewith a dissertation written by Artin Armagan entitled “Bayesian Shrinkage Estimation and Model Selection”. I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Business Administration.

Russell L. Zaretzki, Chair

We have read this dissertation
and recommend its acceptance:

Robert W. Mee

Hairong Qi

Ohannes Karakashian

Accepted for the Council:

Carolyn R. Hodges, Vice Provost
and Dean of the Graduate Studies

(Original signatures are on file with official student records.)

Bayesian Shrinkage Estimation and Model Selection

A Dissertation

Presented for the

Doctor of Philosophy Degree

The University of Tennessee, Knoxville

Artin Armagan

August 2008

Copyright © 2008 by Artin Armagan.
All rights reserved.

Dedication

This dissertation is dedicated to

Aris, Gulfidan, Kevork, Alis, Zabel, Hena and Lidya.

Acknowledgments

First and foremost I would like to mention how indebted and grateful I am to my family for their continuous and endless support throughout my studies, or more generally speaking, throughout all my life. I could never make it through without them. I have always considered myself very lucky to have been born into such a loving and caring family.

I would like to thank Dr. Zaretzki very much for being such a great advisor and a great friend throughout this nerve-wracking process.

I would like to thank Dr. Bozdogan for giving me the idea of pursuing a degree in Statistics, which after all worked rather well. I would like to thank my doctoral committee members Dr. Mee, Dr. Qi and Dr. Karakashian for their time and great advices. I would like to thank Dr. Gilbert for being a great department head who was always very understanding and supportive of students. I would like to thank the rest of the faculty and staff for all their contributions to this department.

I would also like to thank many great friends I met during these years in Knoxville, Tennessee.

Abstract

We introduce a new shrinkage variable selection operator which we term Adaptive Ridge Selector (ARiS). This approach is inspired by the Relevance Vector Machine (RVM) of Tipping (2001), which uses a Bayesian hierarchical linear model to do sparse estimation. RVM was originally introduced to obtain sparse solutions in the case of kernel regression where one has many highly correlated bases (features).

Extending the RVM algorithm, we include a proper prior distribution for the precisions of the regression coefficients along with a hyper-parameter to be chosen. Based upon this model, we derive the full set of conditional posterior distributions for parameters as would typically be done when applying Gibbs sampling. However, instead of simulating samples from the posterior distribution in order to estimate posterior means of quantities, we apply the Lindley-Smith mechanism (Lindley and Smith, 1972). This approach sequentially maximizes the conditional distributions, in order to find the joint maximum of the posterior distribution given the value of the hyper-parameter. An empirical Bayes method is proposed for choosing this hyper-parameter leading to ARiS-eB. Having moved from a Bayes argument, we also look at the problem from a penalized least squares estimation angle.

From the conventional viewpoint, the proposed method eliminates the need for combinatorial search techniques over a discrete model space, converting the model

selection problem into the maximization of the marginal likelihood over a one dimensional continuous space.

Close similarities exist between this estimator obtained and the lasso-type shrinkage estimators. The lasso(Tibshirani, 1996) and its variants, as will be thoroughly discussed, use 1-norm for regularization leading to sparse solutions. The estimator proposed here is contrasted with various other shrinkage estimators along with simulation studies and real data examples.

Inference is also possible using a very straight forward Gibbs sampling procedure after the active variables are determined in the model. The model is also extended to handle departures from normality in the likelihood.

Contents

1	Introduction	1
1.1	Background	3
1.1.1	Subset Selection	3
1.1.2	Regularization and Shrinkage	4
1.2	Contributions	8
1.3	Document Organization	10
2	Literature Review	12
2.1	Bayesian Analysis of Linear Models	12
2.2	Bayesian Treatment of the Independent Student- t Linear Model . . .	17
2.3	Relevance Kernel Machines	19
2.4	Regularization and Variable Selection	23
3	Adaptive Ridge Selector	29
3.1	Extending the Normal Linear Regression Model	29
3.1.1	Computing Posterior Modes	35
3.1.2	Relation to Regularized LS	36
3.1.3	Approximations for the marginal likelihood.	39
3.1.4	The Marginal Posterior Mode of β via EM	43

3.2	Developing ARiS with a Scale Mixtures of Normals	48
3.2.1	Computing Posterior Modes	51
3.2.2	The Joint Posterior Mode of β and ω via EM	51
4	Experimental Results	54
4.1	Simulated Data Sets	54
4.2	Real Data Sets	63
4.2.1	Diabetes Data	63
4.2.2	Prostate Cancer Data	68
5	Conclusion	70
	Bibliography	74
	Appendix	78
A	Some Useful Probability Density Functions	79
B	Conditional Maximization	82
C	Proofs of Theorems	83
	Vita	84

List of Tables

3.1	Estimators for each η	39
4.1	Results for model 0.	55
4.2	Results for model 1.	56
4.3	Results for model 2.	57
4.4	Results for model 3.	58
4.5	Model selection accuracy of ARiS.	61
4.6	Results for $\sigma_c = 10$	64
4.7	Results for $\sigma_c = 20$	65
4.8	Models for ARiS and lasso solution paths.	66
4.9	Results for Diabetes data.	67
4.10	Models for ARiS and lasso solution paths.	68
4.11	Results for Prostate Cancer data.	69

List of Figures

2.1	Laplace (dashed) vs. Normal (solid) density.	25
2.2	Contours of ℓ_1 (dashed), ℓ_2 (dot-dash) and elastic net (solid) penalties.	27
3.1	Contours of the penalty imposed by the independent (log) t -priors for $\mu = 1$ and $\mu = 0.001$	31
3.2	Progression of the ARiS algorithm with η values of $-0.25, 0, 2, 5, 10, 20$. Solid and dashed lines indicate the level sets of the squares residuals and the adapted constrained region at each iteration respectively.	38
3.3	The solution path for the toy example with respect to η	38
4.1	$\mathcal{MSE}$ values for ARiS($\eta = 0$) (solid lines) vs. OLS (dashed lines) estimator for model 1,2,3,4. Sample size increases in the direction $\downarrow$ as 20, 100 and 500 while residual standard deviation increases in the direction $\rightarrow$ as 1, 3 and 6.	62
4.2	Solution path for the diabetes data with respect to η	67
4.3	Solution path for the prostate cancer data with respect to η	69
5.1	Log-marginal likelihood plotted against η values. Solid, dashed, dotted and dot-dash lines stand respectively for $k = 3, 10, 100, 1000$	72

Chapter 1

Introduction

One of the most heavily studied problems in statistics is that of sparse estimation and model selection. Model selection typically refers to the class of problems known as regression where one tries to find a linear combination of basis functions which can be used to predict the outcome of a certain response variable. In choosing a model, a basic goal is the principle of parsimony or sparsity also known as Occam's Razor*. By eliminating redundant variables, the model tends to predict more accurately as may be judged by applying the model to data outside the sample used to fit. Many approaches have been taken to penalize for the inclusion of redundant variables.

With the increasing computational power and various MCMC algorithms, the Bayesian approach to model selection has become increasingly popular. Although this approach appears to provide a comprehensive solution to the variable selection problem, the difficulties of prior specification are significant. Specification of the prior has always been an important question in Bayesian framework. The methods

*Occam's razor (sometimes spelled Ockham's razor) is a principle attributed to the 14th-century English logician and Franciscan friar William of Ockham. The principle states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory. The principle is often expressed in Latin as the *lex parsimoniae* ("law of parsimony" or "law of succinctness").

we consider here automatically select the parameters of the prior distribution from a parametric family using the observed data. Such a method is usually referred to as an empirical Bayesian method.

One distinguishing characteristic of the variable selection problem is the increasing computational cost as a function of the number of variables considered, p . Even with a moderate value of p , computing a goodness-of-fit type characteristic for all the 2^p possible subsets will inevitably be expensive. A brief review, George (2000), provides a good overview of recent developments in the variable selection problem.

Another effective and computationally very attractive approach to model selection has been via shrinkage. The nonnegative garrote of Breiman (1995) and the lasso of Tibshirani (1996) are two recent examples which take advantage of this phenomenon. Though ridge regression proposed by Hoerl and Kennard (1970) may lead to smaller MSE s than the unbiased maximum likelihood estimator, it does not lead to sparse models. This is due to the structure of the constraint introduced on the least-squares problem which will be discussed in detail.

Our approach to the problem of model selection will focus on shrinkage. Creating a Bayes argument, we will obtain automatic shrinkage which will lead us to a specific type of regularization of the usual sum of squares problem. Here we will see the close connections and distinctions between our method and the other regularized estimators.

1.1 Background

1.1.1 Subset Selection

Consider the following familiar linear regression equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1.1)$$

where $\mathbf{y}$ is an n -dimensional vector, $\mathbf{X}$ is the $n \times p$ dimensional model matrix and $\boldsymbol{\epsilon}$ is an n -dimensional vector observed from $N_p(\mathbf{0}, \sigma^2 \mathbf{I}_p)$. Some of the components of $\boldsymbol{\beta}$ may be zero corresponding to variables that have no value in the prediction of y . We would like to determine the subset of variables that best explains the measured responses $\mathbf{y}$. As noted by George and Foster (2000) a common strategy for this variable selection problem has been to select a model that maximizes a penalized sum of squares criterion. Enumerating different subsets by an index $\gamma = 1, \dots, 2^p$, let

$$SS_\gamma = \hat{\boldsymbol{\beta}}' \mathbf{X}_\gamma' \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}, \quad (1.2)$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimate for the γ th subset. Following this, the model selection can be realized by maximizing

$$SS_\gamma / \hat{\sigma}^2 - Fq_\gamma, \quad (1.3)$$

or equivalently minimizing $SSE_\gamma / \hat{\sigma}^2 + Fq_\gamma$ where q_γ is the number of variables in the γ th subset, F is a fixed constant, $\hat{\sigma}^2$ is an estimate of σ^2 and $SSE_\gamma = \mathbf{y}'\mathbf{y} - SS_\gamma$.

Familiar model selection criteria can be obtained with different choices of F . For example, $F = 2$ corresponds to C_p (Mallows, 1973), and, approximately, AIC (Akaike, 1974), $F = \log n$ corresponds to BIC (Schwarz, 1978). These three approaches had

distinct motivations. C_p was derived as an unbiased estimate of predictive risk, AIC as an expected information distance between the chosen model and the unknown true model, while BIC is an asymptotic approximation to the Bayes factor.

The Bayesian alternative to the above selection criteria can be obtained through the maximization of the evidence (marginal likelihood) conditioned over the model current subset, $p(\mathbf{y}|\gamma)$. If a prior distribution on the discrete model space γ is present, a function of the form $g(\gamma) \propto p(\gamma|\mathbf{y})$ may be maximized instead. Such a structure is usually obtained through the use of conjugate priors which provide analytical tractability to the computation of the posterior distribution of the parameters. George and Foster (2000) reveal the connection between the criteria of the form (1.3) and this Bayesian approach under a specific conjugate prior setup. Chipman et al. (2001) discuss stochastic search techniques based on Metropolis-Hastings for model selection. The approach taken in this paper is quite distinct from the methods just cited and contrasts heavily with the usual requirement to employ a greedy-search algorithm of some sort to search through a relatively large discrete set of possible models. In particular, although quite effective and useful, search techniques such as those discussed in Chipman et al. (2001) will require a huge number of Markov Chain steps as the number of variables in the model increases.

1.1.2 Regularization and Shrinkage

An important property of the solution that we derive for the variable selection problem is referred to as the shrinkage effect. A good illustration of this effect occurs in ridge regression problem where by intentionally biasing the estimator we hope for a reduced MSE (mean squared error). From a Bayesian perspective, when the prior mean is a

null vector, and the prior covariance is $\sigma^2\lambda^{-1}\mathbf{I}$, the ridge estimator is obtained as

$$\bar{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad (1.4)$$

where λ is to be set or estimated and σ^2 is the error variance. As $\lambda \rightarrow \infty$, the regression coefficients are shrunk toward zero, or more generally, to the mean vector specified by the prior distribution. From a Bayesian perspective, if $\lambda \rightarrow 0$ indicating a vague prior opinion about the value of the regression coefficients, then there is little shrinkage and the ridge estimator approaches the ML estimator. In models with extreme collinearity, where $\mathbf{X}'\mathbf{X}$ is nearly singular, the corresponding posterior distribution will have very large variances as the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1} \rightarrow \infty$. One basic consequence of such a situation is that it masks the effects of the regressors, as is commonly illustrated in introductory texts on regression modeling. Using the ridge estimator can be a remedy at the cost of obtaining a biased estimator.

Within a decision theoretic approach, assuming a squared error loss function, the average loss or risk is simply the mean squared error $\mathcal{MSE}(\theta)$ which can be decomposed as:

$$\begin{aligned} \mathcal{MSE}(\theta) &= \mathbb{E}(\hat{\theta} - \theta)^2 \\ &= \mathbb{V}(\hat{\theta}) + Bias^2(\hat{\theta}). \end{aligned} \quad (1.5)$$

Using a ridge estimator reduces the variance of the estimator of the parameter as the ill-conditioned $\mathbf{X}'\mathbf{X}$ is corrected, at the expense of adding bias. This is a trade-off situation and we hope that the decrease in $\mathbb{V}(\hat{\theta})$ is large enough to overcome the increase due to the $Bias^2(\hat{\theta})$ term. This result extends beyond situations where $\mathbf{X}'\mathbf{X}$

is nearly singular. In general, one may obtain a reduced- $\mathcal{MSE}$ estimator using the ridge estimator.

In ridge regression, one can estimate the optimal value of λ with various methods. A frequentist may choose to minimize the $\mathcal{MSE}$ under a decision theoretic framework, while a Bayesian may approach the problem within the empirical Bayes framework and try to estimate λ using marginal likelihood maximization. We propose to estimate λ using the marginal likelihood (evidence) derived by integrating the product of the density times prior over all model parameters. The result is a one parameter likelihood function depending on λ which may be maximized over values of λ , an empirical Bayes approach. We may write

$$p(\mathbf{y}|\lambda) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}, \quad (1.6)$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood of the observed values and $p(\boldsymbol{\theta}|\lambda)$ is the prior density assumed over the model parameters.

Tipping (2001) created a hierarchical Bayesian normal linear model utilizing standard conjugate priors[†] for the model parameters with the exception of the covariance matrix for the regression coefficients $\mathbf{V}$. By adopting $\mathbf{V}$, a diagonal matrix, in lieu of $\lambda\mathbf{I}$ used in ridge regression, Tipping (2001) has introduced the flexibility of a non-spherical prior distribution over the regression coefficients. As will be discussed, this allows individual diagonal elements to be estimated and shrinks the coefficients of the irrelevant variables/bases to zero. The idea was introduced in the context of kernel regression but is applicable to the ordinary linear case as well. The goal is to rid the model of “irrelevant” basis functions and create a parsimonious prediction model.

[†]The familiar conjugate priors under the normal model; inverted gamma density on the error variance and a multivariate normal density on the regression coefficients.

Remember that the ridge estimator may be written as a solution to a regularized least squares problem,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 \quad s.t. \quad \sum_{j=1}^p \beta_j^2 < s, \quad (1.7)$$

which can also be written as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \beta_j^2, \quad (1.8)$$

which, combining the squares, easily leads to the ridge estimator in 1.4.

Tipping's approach can analogously be written as a regularization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \sum_{j=1}^p v_j \beta_j^2, \quad (1.9)$$

which imposes an elliptical rather than spherical constraint on the least squares. He determines these v_j s using an empirical Bayes approach as will be discussed in detail.

We can write the regularization problem in 1.8 in a more general fashion as

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|^q, \quad (1.10)$$

where $q \geq 0$. One may choose q to be any desired value. For example $q = 0$ leads to the subset selection or as some may call hard-thresholding. $q = 2$, as shown earlier, will lead to the ridge estimator which will never lead to sparse solutions. As q departs downward from the value 2, some coefficients will be zeroed. An important advantage of $q = 1$ is that it leads to the sparsest possible solution while maintaining the convexity of the optimization problem. Values of q smaller than 1 will lead to multimodal objective functions and thus would not be reasonable choices. This is one

of major reasons the Lasso of Tibshirani (1996) has become quite popular. Another more recent reason is the LARS algorithm (Efron et al., 2004), which provides very efficient calculation.

1.2 Contributions

This dissertation proposes a new variable selection and estimation technique having been inspired by the hierarchical Bayesian linear model discussed in Tipping (2001). The simultaneous estimation and model selection is realized by only a simple, continuous maximization procedure. This provides tremendous advantages in terms of computational time since a search technique over the discrete subset space is not required. This is a feature shared with the other lasso-type estimators. As will be seen through numerical examples, our model will lead to significantly better results when only a small number of considered effects are active.

As will be discussed in later chapters, Tipping (2001) estimates the hyper-parameters of hierarchical linear model via marginal likelihood maximization. The marginal likelihood, $p(\mathbf{y}|\sigma^2, \mathbf{v})$, is maximized with respect to the error variance σ^2 and the prior precisions of the regression coefficients v_j^{-1} , $j = 1, \dots, p$.

Instead, we proceed by constructing a Bayesian hierarchical structure and obtaining the joint posterior of $\boldsymbol{\beta}$, σ^2 and v_j^{-1} up to a normalizing constant. This joint posterior density is maximized with respect to these parameters via conditional maximization (Lindley and Smith, 1972). Although the joint posterior distribution is not of known form, we are able to obtain the fully conditional posterior densities of $\boldsymbol{\beta}$, σ^2 and v_j^{-1} which are known distributions with closed form expressions for the modes. Sequential maximization of these conditional distributions effectively finds the joint posterior mode. This joint posterior mode is useful as a point estimator analogous to

the maximum likelihood estimator. It is not a Bayes estimator under squared error loss; however it is Bayes under a 0 – 1 type of loss. One of the intuitive justifications we can provide for the use of modal values instead of the expected values is that, under the support of the data, the mode is the value that is most likely to be observed. Since it is not sensible to talk about probabilities of individual points in a continuous parameter space, one may intuitively also think that in most cases a higher posterior density region will be obtained in the vicinity of the posterior mode. With such fully conditional distributions of the parameters in hand, one may also implement a Gibbs sampling algorithm to sample from the joint posterior distribution for inference.

We form the joint posterior distribution assuming a proper prior over the precision of the regression parameters $v_j^{-1} \sim p(v_j^{-1}|\eta)$ where η is the shape parameter of the distribution. This additional hyper-prior extends the RVM (*relevance vector machines*) model developed in Tipping (2001). When η is set equal to 0 a special case arises which will be discussed later. Having defined this proper density for v_j^{-1} , solutions will differ with the value of η . As will be discussed in more detail, for increasing values of η a larger proportion of β_j will be forced to 0. Since η prescribes the amount of shrinkage imposed on the model, one would need to choose the value for η which yields the “correct” amount of shrinkage.

As a further contribution, we introduce maximization of the marginal likelihood as a function of η . The marginal likelihood of the observations is obtained by integration over all the model parameters which are treated as random quantities. We proceed to maximize this marginal density $p(\mathbf{y}|\eta)$ with respect to the parameter η . Since the integration over all the model parameters is not analytically tractable, we use either Laplace’s method or numerical integration to approximate the integral. The modal value obtained earlier is utilized for a second-order Taylor expansion for the Laplace approximation.

We also re-formulate the problem in the context of regularized least squares estimation and show that the proposed method is a sequence of ridge regression problems by re-weighting the predictor variables at each iteration by a certain quantity.

The setup introduced above assumes a normal distribution for the random noise, which may not be realistic in certain situations. Utilizing a Student- t model for the random noise, we may be better able to handle possible outliers and data with more kurtosis. Geweke (1993) exploited the equivalence of the Student- t distribution and an appropriate scale mixture of normals for robust Bayesian regression. Once such a hierarchical setup is obtained under a linear model, a similar conditional maximization step follows.

We also derive the marginal mode of the regression coefficients (the joint mode of the regression coefficients and the sample weights in the robust case) using an expectation-maximization (EM) procedure. Our derivations indicate that the marginal mode of the regression coefficients is identical in form to the β component of the joint posterior mode. There are only slight differences in the marginal case due to the use of expectations instead of the modes.

1.3 Document Organization

The remainder of the thesis is organized in the following manner:

Chapter 2 consists of a literature review and background information on the methods we will be using for either comparison purposes or as building blocks of the methods proposed. Section 2.1 introduces a Bayesian linear model with a particular choice of conjugate priors to guide us all throughout the dissertation. Section 2.3 discusses the Bayesian learning procedure introduced in Tipping (2001) which yields a shrinkage estimator creating sparse models. This methodology is applied in the context of

relevance vector machines (RVM). Section 2.4 goes over numerous regularized least squares estimation methods and their relations to Bayesian arguments.

Chapter 3 introduces the method we propose, *adaptive ridge selector* (ARiS), under the normal and Student- t setups.

Chapter 4 presents experimental results on both simulated and real data sets in variety of situations.

Chapter 5 discusses conclusions including advantages and disadvantages of the proposed method.

Chapter 2

Literature Review

2.1 Bayesian Analysis of Linear Models

Linear regression analysis is arguably one of the most widely used statistical methods. In this section we present the Bayesian treatment of such models. We also establish notation that will be used throughout this study.

The normal linear regression model is typically denoted

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (2.1)$$

where the error term is $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, and σ^2 is an unknown dispersion parameter.

The likelihood function then can be written up to a constant as

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right]. \quad (2.2)$$

Let us decompose the numerator of the exponent as

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = S_e + S_\beta, \quad (2.3)$$

where

$$S_e = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (2.4)$$

$$S_\beta = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \quad (2.5)$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}. \quad (2.6)$$

This decomposition will lead us to the joint, conditional and marginal posterior distributions of interest. Here we will consider inference under conjugate priors.

Priors

Suppose that an inverse-gamma prior is adopted for σ^2 and a multivariate normal prior is assigned to $\boldsymbol{\beta}$. We denote these as

$$p(\boldsymbol{\beta}|\sigma^2, \mathbf{V}) \sim \mathcal{N}_p(\mathbf{m}_\beta, \sigma^2 \mathbf{V}) \quad (2.7)$$

$$p(\sigma^2|\nu, \lambda) \sim \text{inv-gamma}(\nu/2, \lambda/2). \quad (2.8)$$

Given these priors, the joint posterior density of $\boldsymbol{\beta}$ and σ^2 can be written up to a constant as

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, H) \propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} | \sigma^2, H_\beta) p(\sigma^2 | H_{\sigma^2}) \quad (2.9)$$

$$\begin{aligned} &\propto (\sigma^2)^{-\left(\frac{n+\nu+p}{2}+1\right)} \\ &\quad \times \exp \left[-\frac{S_\beta + (\boldsymbol{\beta} - \mathbf{m}_\beta)' \mathbf{V}^{-1} (\boldsymbol{\beta} - \mathbf{m}_\beta) + S_e + \lambda}{2\sigma^2} \right], \end{aligned} \quad (2.10)$$

where $H = (H_\beta, H_{\sigma^2})$ represent the hyper-parameters. Two quadratic forms can be combined using the identity

$$S_\beta + (\beta - \mathbf{m}_\beta)' \mathbf{V}^{-1} (\beta - \mathbf{m}_\beta) = (\beta - \bar{\beta})' (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} (\beta - \bar{\beta}) + S_{\hat{\beta}},$$

where

$$\begin{aligned}\bar{\beta} &= (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} (\mathbf{X}'\mathbf{y} + \mathbf{V}^{-1}\mathbf{m}_\beta), \\ S_{\hat{\beta}} &= (\hat{\beta} - \mathbf{m}_\beta)' \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{V}^{-1} (\hat{\beta} - \mathbf{m}_\beta).\end{aligned}$$

This decomposition makes it rather simple to extract the conditional and marginal posterior distributions of β and σ^2 . For additional details see Sorensen and Gianola (2002).

Conditional Posterior Distribution of the Regression Coefficients

The conditional posterior distribution of the regression coefficients are easily attainable via examining the joint posterior distribution and fixing the parameter(s) on which we condition. Thus, it becomes easy to see that the conditional posterior distribution of the regression coefficients, given σ^2 , is a multivariate normal process

$$\beta | \sigma^2, \mathbf{y}, H \sim \mathcal{N}(\bar{\beta}, (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \sigma^2). \quad (2.11)$$

Notice that when $\mathbf{V}^{-1} \rightarrow \mathbf{0}$, meaning the prior information becomes increasingly diffuse, the posterior covariance tends to $(\mathbf{X}'\mathbf{X})^{-1} \sigma^2$, which is identical to the variance-covariance matrix of the ML estimator. The same occurs when the information in the data is very large relative to that contributed by the prior distribution. Hence, in the limit, our conditional distribution is centered at the ML estimator, and the posterior

covariance matrix is identical in form to the asymptotic covariance of the ML estimator. This is a particular case of a more general result on asymptotic approximations to posterior distributions under regularity conditions.

In the case of linear regression under the Bayesian paradigm, the observed information matrix is equivalent to the expected information under the repeated sampling paradigm. This is due to the fact that the matrix of negative second derivatives of the log-posterior with respect to $\boldsymbol{\beta}$ does not involve the observations. Thus, the observed information is a constant and, therefore, is equal to its expectation taken over the distribution of the data. It is important not to forget that this would take the form of the asymptotic variance-covariance matrix of the ML estimator only when $\mathbf{X}'\mathbf{X}$ overwhelms $\mathbf{V}^{-1}$.

Conditional Posterior Distribution of the Error Variance

Treating the regression coefficients as fixed quantities, the conditional posterior distribution of the variance is an inverse-gamma density with a shape parameter $(n + p + \nu)/2$, and a scale parameter of $[S_\beta + (\boldsymbol{\beta} - \mathbf{m}_\beta)' \mathbf{V}^{-1} (\boldsymbol{\beta} - \mathbf{m}_\beta) + S_e + \lambda] / 2$.

Marginal Posterior Distribution of the Regression Coefficients

Using the properties of the inverse gamma distribution, the joint density in (2.10) can be integrated over σ^2 to obtain the closed-form marginal density of the regression coefficients

$$p(\boldsymbol{\beta}|\mathbf{y}, H) \propto \int (\sigma^2)^{-\left(\frac{n+\nu+p}{2}+1\right)} \times \exp \left[-\frac{(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + S_{\hat{\boldsymbol{\beta}}} + S_e + \lambda}{2\sigma^2} \right] d\sigma^2$$

$$\begin{aligned}
&\propto \left[\frac{(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + S_{\hat{\boldsymbol{\beta}}} + S_e + \lambda}{2} \right]^{-\frac{n+\nu+p}{2}} \\
&\propto \left[\frac{1 + (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})}{(n + \nu) \left(\frac{S_{\hat{\boldsymbol{\beta}}} + S_e + \lambda}{n + \nu} \right)} \right]^{-\frac{n+\nu+p}{2}}. \tag{2.12}
\end{aligned}$$

This is the kernel of a p -dimensional multivariate- t density. The corresponding distribution has a mean of $\bar{\boldsymbol{\beta}}$, a covariance matrix of $\frac{S_{\hat{\boldsymbol{\beta}}} + S_e + \lambda}{n + \nu - 1} (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1}$ and degrees of freedom of $n + \nu$. Thus, all marginal distributions of individual or of subsets of regression coefficients are either univariate or multivariate- t .

Note that at first sight there does not seem to be a “loss of degrees of freedom”, relative to n , in the process of accounting for uncertainty about the variance. In fact, however, there is a “hidden loss” of p degrees of freedom, which is canceled by the contribution made by the conditional prior of the regression coefficients, which involves σ^2 . If the prior for $\boldsymbol{\beta}$ had not involved σ^2 , then the degrees of freedom of the marginal posterior distribution of the regression coefficients would have been $n + \nu - p$. If, in addition, the “degree of belief” parameter of the prior distribution for σ^2 had been taken to be $\nu = 0$, the degrees of freedom would have been $n - p$, exactly the number of degrees of freedom arising in a standard classical analysis of linear regression. Note, however, that the Bayesian assignment $\nu = 0$ produces the improper prior distribution $p(\sigma^2) \propto 1/\sigma^2$. In this case, the posterior distribution of the regression coefficients would be proper if $n - p > 0$.

Marginal Posterior Distribution of the Error Variance

Given the form of (2.10) it is rather easy to integrate over the regression coefficients to obtain the marginal posterior density of σ^2 .

$$\begin{aligned}
 p(\sigma^2 | \mathbf{y}, H) &\propto (\sigma^2)^{-\left(\frac{n+\nu+p}{2}+1\right)} \exp\left[-\frac{S_{\hat{\beta}} + S_e + \lambda}{2\sigma^2}\right] \\
 &\quad \times \int \exp\left[-\frac{(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})}{2\sigma^2}\right] d\boldsymbol{\beta} \\
 &\propto (\sigma^2)^{-\left(\frac{n+\nu}{2}+1\right)} \exp\left[-\frac{S_{\hat{\beta}} + S_e + \lambda}{2\sigma^2}\right].
 \end{aligned} \tag{2.13}$$

Hence, the marginal posterior of σ^2 is an inverse-gamma process with a shape parameter of $(n + \nu)/2$ and a scale parameter of $(S_{\hat{\beta}} + S_e + \lambda)/2$.

2.2 Bayesian Treatment of the Independent Student- t Linear Model

Geweke (1993) adopted the methods for Bayesian inference in a linear model in which the disturbances are independent and have identical Student- t distributions. His approach exploits the equivalence of the Student- t distribution and an appropriate scale mixture of normals, and uses a Gibbs sampler to perform the computations.

We first demonstrate the equivalence of Student- t distribution and a specific scale mixture of normals. Let us assume

$$y_i | \mathbf{X} \sim t(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2; v), \tag{2.14}$$

where $t(\cdot)$ denotes a univariate Student- t distribution with mean $\mathbf{x}_i \boldsymbol{\beta}$, variance σ^2 and degrees-of-freedom parameter v . The probability density function of a single

observation can be written as

$$f(y_i) = \Gamma\left(\frac{v+1}{2}\right) \left[\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{v}{2}\right) \sigma \right]^{-1} v^{-1/2} [1 + (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 / v \sigma^2]^{-(v+1)/2}. \quad (2.15)$$

Here, v , the degrees-of-freedom parameter is part of the specification of the model.

Given the improper prior $p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$, the posterior density will have the kernel

$$\sigma^{-(n+1)} \prod_{i=1}^n [1 + (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 / v \sigma^2]^{-(v+1)/2}. \quad (2.16)$$

We have obtained a posterior density under the assumptions that the observations are independent and generated from a Student- t distribution. Now, consider the following specifications:

$$y_i | \mathbf{X} \sim \mathcal{N}(\mathbf{x}_i \boldsymbol{\beta}, \sigma^2 \omega_i) \quad (2.17)$$

where ω_i are additional unknown model parameters. The likelihood in this case can be written as

$$p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \boldsymbol{\omega}) = \sigma^{-n} \prod_i^n \omega_i^{-1/2} \exp \left[- \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 / 2 \sigma^2 \omega_i \right], \quad (2.18)$$

where $\boldsymbol{\omega}$ denotes the vector of ω_i s. The prior density is of the form $p(\boldsymbol{\beta}, \sigma, \boldsymbol{\omega}) = p(\boldsymbol{\beta}) p(\sigma^2) p(\boldsymbol{\omega})$ where the priors on $\boldsymbol{\beta}$ and σ are as given before. In the prior distribution of $\boldsymbol{\omega}$, ω_i are independent, with $v/\omega_i \sim \chi^2(v)$ which leads to

$$p(\boldsymbol{\omega}) = (v/2)^{nv/2} \left[\Gamma\left(\frac{v}{2}\right) \right]^{-n} \prod_{i=1}^n \omega_i^{-(v+2)/2} \exp(-v/2\omega_i). \quad (2.19)$$

This prior distribution for ω_i was suggested by Lindley (1971) for cell variances in the analysis of variance with multiple observations per cell. The product of the likelihood

and the prior densities yields the posterior density kernel

$$(v/2)^{nv/2} \left[\Gamma\left(\frac{v}{2}\right) \right]^{-n} \sigma^{-(n+1)} \prod_i^n \omega_i^{-(v+3)/2} \exp \left\{ - \sum_{i=1}^n [\sigma^{-2} (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + v] / 2\omega_i \right\}. \quad (2.20)$$

Using the result that

$$\int_0^\infty x^{-a/2} \exp(-b/2x) dx = (2/b)^{(a-2)/2} \Gamma\left(\frac{a-2}{2}\right), \quad (2.21)$$

and integrating the above posterior density in (2.20) with respect to $\boldsymbol{\omega}$, we obtain

$$(v/2)^{nv/2} \left[\Gamma\left(\frac{v+1}{2}\right) \right]^n \left[\Gamma\left(\frac{v}{2}\right) \right]^{-n} 2^{n(v+2)/2} \sigma^{-(n+1)} \prod_i^n [\sigma^{-2} (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + v]^{-(v+1)/2}, \quad (2.22)$$

which, as a function of $\boldsymbol{\beta}$ and σ , is proportional to (2.16). Therefore, the normal mixture model with the independent priors $v/\omega_i \sim \chi^2(v)$ is exactly same as the independent Student- t linear model.

The statements on the existence of the posterior density and the moments can be found in Geweke (1993). As Geweke (1993) states as well, this posterior density kernel is not analytically tractable. However, using the fully conditional posterior densities, a very straight-forward Gibbs sampler can be implemented for inference. We will derive similar conditional posterior densities to those of Geweke (1993) in section 3.2.

2.3 Relevance Kernel Machines

The relevance vector machine (RVM) (Tipping, 2001) is a Bayesian sparse kernel technique for regression and classification that shares many characteristics of the support

vector machine (SVM) while avoiding its principle limitations. Additionally, it typically leads to much sparser models resulting in correspondingly faster performance on test data whilst maintaining comparable generalization error. Here we introduce only the regression form of the RVM.

The relevance vector machine for regression is a linear model shown in section 2.1 with a little twist to the priors leading to sparse solutions. Here we will be working in the context of kernel transformations replacing $\mathbf{X}$ by the Gram matrix Φ which is an $n \times n$ matrix representing the n -dimensional feature space after the transformation from the original space. As the kernel transformations are not the main focus of this thesis, we will not spend any time on this. An excellent reference for this method is Bishop (2006).

Priors

Tipping introduces a prior for the regression coefficients of the form

$$\boldsymbol{\beta}|\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}), \quad (2.23)$$

where

$$\mathbf{V} = \begin{pmatrix} v_1 & & 0 \\ & \ddots & \\ 0 & & v_p \end{pmatrix}. \quad (2.24)$$

Further, v_j^{-1} s are gamma distributed with shape parameter a and scale parameter of b . Tipping later proceeds to set $a = b = 0$ to obtain a non-informative improper prior of the form $p(v_j^{-1}) \propto 1/v_j^{-1}$. Henceforth, we will assume a and b are set to zero.

Likewise, he proceeds to put a non-informative improper prior on the residual error, $p(\sigma^2) \propto 1/\sigma^2$.

Inference

The joint posterior distribution of the unobservables can be obtained with Bayes theorem as

$$p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1} | \mathbf{y}) = \frac{p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}) p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1})}{p(\mathbf{y})}, \quad (2.25)$$

where $\mathbf{v}^{-1} = \text{diag}(\mathbf{V}^{-1})$. Given a new test point, $\mathbf{x}^*$, predictions are made for a corresponding target $\mathbf{y}^*$, in terms of the predictive distribution:

$$p(\mathbf{y}^* | \mathbf{y}) = \int p(\mathbf{y}^* | \boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}) p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1} | \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 d\mathbf{v}^{-1}. \quad (2.26)$$

The posterior density in (2.26) is analytically unattainable as one cannot perform the integral analytically to obtain the normalizing constant.

Instead, he proceeds to decompose the joint posterior density as

$$p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1} | \mathbf{y}) = p(\boldsymbol{\beta} | \mathbf{y}, \sigma^2, \mathbf{v}^{-1}) p(\sigma^2, \mathbf{v}^{-1} | \mathbf{y}). \quad (2.27)$$

The integral in (2.26) can be written as

$$\begin{aligned} p(\mathbf{y}^* | \mathbf{y}) &= \int p(\mathbf{y}^* | \boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}) p(\boldsymbol{\beta} | \mathbf{y}, \sigma^2, \mathbf{v}^{-1}) p(\sigma^2, \mathbf{v}^{-1} | \mathbf{y}) d\boldsymbol{\beta} d\sigma^2 d\mathbf{v}^{-1} \\ &= \int p(\mathbf{y}^* | \sigma^2, \mathbf{v}^{-1}) p(\sigma^2, \mathbf{v}^{-1} | \mathbf{y}) d\sigma^2 d\mathbf{v}^{-1}, \end{aligned} \quad (2.28)$$

since the first two terms in the integral (the likelihood of a future observation and the conditional posterior density of the regression coefficients) are both normal densities.

After this, Tipping proceeds to get an approximation to the above integral by approximating $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y})$ by a delta-function* at its mode; $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}) \approx \delta(\tilde{\sigma}^2, \tilde{\mathbf{v}}^{-1})$, where, $\tilde{\sigma}^2$ and $\tilde{\mathbf{v}}^{-1}$ stand for the modal values of $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y})$.

Relevance vector ‘learning’ thus becomes a search for the hyper-parameter posterior mode, i.e. the maximization of $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}) \propto p(\mathbf{y}|\sigma^2, \mathbf{v}^{-1})p(\sigma^2)p(\mathbf{y})$ with respect to σ^2 and $\mathbf{v}^{-1}$. Equivalently, and more straightforwardly, the log of this quantity is maximized. In addition, the maximization is realized with respect to $\log v_j^{-1}$ and $\log \sigma^{-2}$. This is convenient since in practice uniform hyper-priors over a logarithmic scale are assumed, and the derivatives of the prior terms vanish in the space.

The objective function then to be maximized is

$$\log p(\mathbf{y}|\log v_j^{-1}, \log \sigma^{-2}) = -\frac{1}{2} \left[\log |\sigma^2 \mathbf{I} + \Phi \mathbf{V} \Phi'| + \mathbf{y}' (\sigma^2 \mathbf{I} + \Phi \mathbf{V} \Phi')^{-1} \mathbf{y} \right]. \quad (2.29)$$

Taking the derivatives and setting them equal to zero leads to the following re-estimation equations:

$$v_j^{new} = \frac{m_j^2}{\gamma_j}, \quad (2.30)$$

$$\sigma^{2new} = \frac{\|\mathbf{y} - \Phi \mathbf{m}\|^2}{n - \sum_j \gamma_j}, \quad (2.31)$$

where m_j is the j th component of the posterior mean of $\boldsymbol{\beta}$, $\mathbf{m} = \sigma^{-2} \Sigma \Phi' \mathbf{y}$, $\gamma_j = 1 - \Sigma_{jj}/v_j$, and Σ_{jj} is the j th diagonal of the conditional posterior covariance of $\boldsymbol{\beta}$, $\Sigma = (\mathbf{V}^{-1} + \sigma^{-2} \Phi' \Phi)^{-1}$.

*The Dirac delta function can be loosely thought of as a function on the real line which is zero everywhere except at x_0 , where it is infinite, $\delta(x) = \begin{cases} \infty & x = x_0 \\ 0 & x \neq x_0 \end{cases}$. It is also constrained to satisfy the identity $\int_{-\infty}^{\infty} \delta(x) dx = 1$.

2.4 Regularization and Variable Selection

In this section we will briefly introduce some regularized least squares type estimators with which we contrast the proposed estimator.

Ridge Regression

In cases where $\mathbf{X}'\mathbf{X}$ is a nearly singular matrix, the OLS estimator may be a poor estimator of $\boldsymbol{\beta}$. Some diagonals of $(\mathbf{X}'\mathbf{X})^{-1}$ will be inflated due to this ill-conditioning, leading to very unstable, high-variance estimators. Hoerl and Kennard (1970) investigated the option of perturbing the eigenvalues of the ill-conditioned matrix $\mathbf{X}'\mathbf{X}$ by a diagonal matrix of the form $\lambda\mathbf{I}$. This adjusted estimator takes the form

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{y}, \quad (2.32)$$

where $k \geq 0$.

It is well known that the OLS estimator is an unbiased estimator in the case of normal likelihood. This estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ gives the minimum sum of squared residuals $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Let us define two quantities as provided in Hoerl and Kennard (1970):

- i. $\mathbb{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.
- ii. $\mathbb{E}(\mathcal{L}^2) = \sigma^2 \text{Trace}(\mathbf{X}'\mathbf{X})^{-1}$, where $\mathcal{L} := \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2$.

In the case of ill-conditioning of $\mathbf{X}'\mathbf{X}$ both of these quantities may be severely inflated indicating instability and departures from $\boldsymbol{\beta}$.

The ridge estimator leads to an increase in bias and subsequent decrease in the variance of $\hat{\boldsymbol{\beta}}$. Given $\mathcal{MSE} = \mathbb{V}(\hat{\boldsymbol{\beta}}) + \text{Bias}^2(\hat{\boldsymbol{\beta}})$, there is a possibility that there are values of λ for which the $\mathcal{MSE}$ is less for $\hat{\boldsymbol{\beta}}^*$ than it is for the usual solution $\hat{\boldsymbol{\beta}}$.

The Nonnegative Garrote

The ridge estimator has its own drawbacks. Its solutions are no more parsimonious than those of ordinary regression. If the underlying model has a subset of coefficients equal to zero, the ridge estimator will not be able to capture this. Since the ridge estimator results from an optimization problem with a spherically constrained region, the solution will never lie on the axes unless the OLS estimator lies on them.

Breiman (1995) proposed a new estimator that results from the following minimization problem:

$$\min_{c_j} \sum_i \left(y_i - \sum_j c_j \hat{\beta}_j x_{ij} \right)^2 \quad s.t. \quad c_j \geq 0, \quad \sum_j c_j \leq s. \quad (2.33)$$

Here $\hat{\beta}$ is the OLS estimator. As s decreases, more of the c_j become zero and the remaining nonzero components are shrunk. The garrote eliminates some variables and shrinks others. Breiman (1995) shows that it is both relatively stable and scale invariant.

Lasso

Tibshirani (1996) introduced a new estimator named the lasso (least absolute shrinkage and selection operator) which used an ℓ_1 polytope as the constrained region instead of ℓ_2 . An obvious consequence of this penalty is its ability to drive the coefficients to zero which the ℓ_2 -penalty can never achieve. The lasso estimator is given as a solution to the following quadratic program:

$$\min_{\beta_j} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \quad s.t. \quad \sum_j |\beta_j| \leq t \quad (2.34)$$

Here $t \geq 0$ is a tuning parameter and controls the amount of shrinkage that is applied to the estimates. The lasso constraint $\sum_j |\beta_j| \leq t$ is equivalent to the addition of a penalty term $\lambda \sum |\beta_j|$ to the residual sum of squares. This easily follows looking at λ as a Lagrange multiplier. In fact the same argument may easily be derived from a Bayesian perspective. Using a Laplace (double-exponential) distribution as a prior on the regression coefficients will lead to an estimator of this form where λ is the scale parameter of this distribution. Considering both the ridge and the lasso estimates as Bayes estimators, the obvious distinction between the two is the form of the priors. The ridge estimator results from a normal prior on the regression coefficients while the lasso results from a Laplace prior. A Laplace distribution, relative to a normal distribution, assigns more density near 0 and in the tails which explains lasso's tendency to produce estimates that are either large or 0 (see figure 2.1).

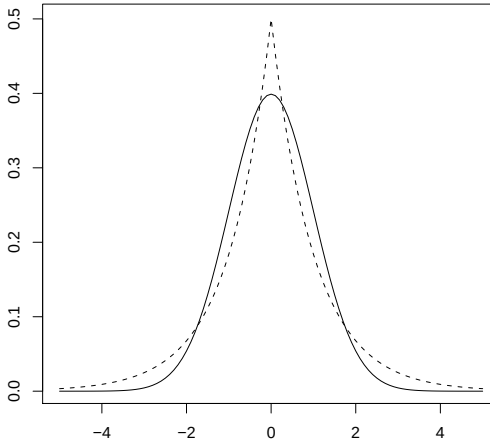


Figure 2.1: Laplace (dashed) vs. Normal (solid) density.

Tibshirani (1996) outlines the performance of lasso under three different scenarios:

- i. *Small number of large effects:* The lasso does not perform as well as subset selection. In fact we will see in the experimental results that our proposed method outperforms any other method it's contrasted with in this case.
- ii. *Small to moderate number of moderate-sized effect:* The lasso performs the best. As we'll see, our proposed method is at least competitive under such circumstances for smaller sample sizes, and much superior for larger sample sizes.
- iii. *Large number of small effects:* The lasso does not perform as well as the ridge estimator. As we'll see in the experiments, none of the methods tested can surpass the performance of the ridge estimator under such a setup.

The Elastic Net

Zou and Hastie (2005) introduced a new regularization and variable selection method, the elastic net, with the goal of combining the attractive features of the lasso and the ridge estimators.

For any nonnegative λ_1 and λ_2 , the naïve elastic net estimator is obtained as the solution to the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda_2 \sum_j \beta_j^2 + \lambda_1 \sum_j |\beta_j| \quad (2.35)$$

If we let $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$, the above problem can be re-written as

$$\min_{\boldsymbol{\beta}} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \quad s.t. \quad \alpha \sum_j \beta_j^2 + (1 - \alpha) \sum_j |\beta_j| \leq t \quad (2.36)$$

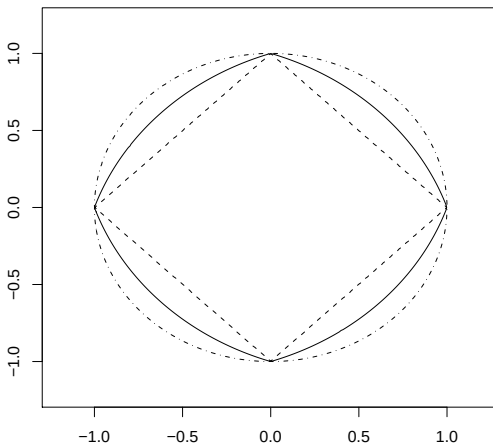


Figure 2.2: Contours of ℓ_1 (dashed), ℓ_2 (dot-dash) and elastic net (solid) penalties.

for some $t \geq 0$. Zou and Hastie (2005) refer to the function $\alpha \sum_j \beta_j^2 + (1 - \alpha) \sum_j |\beta_j|$ as the elastic net penalty. This function constitutes a compromise between the ℓ_1 and ℓ_2 penalties (see figure 2.2). When $\alpha = 1$, the naïve elastic net estimator becomes the ridge estimator, while when $\alpha = 0$, it becomes the lasso estimator.

Making the required arrangements see (Zou and Hastie, 2005, Lemma 1) the naïve elastic net solution can be obtained by solving the computationally efficient lasso (due to the lars of Efron et al. (2004)). Zou and Hastie (2005) assert based upon empirical evidence that the naïve elastic net does not perform satisfactorily unless it is very close to either the ridge estimator or the lasso which is why they call it “naïve”. This estimator obtained through two stages of shrinkage, one via the ℓ_1 -penalty and the other through the ℓ_2 . They claim that this double shrinkage does not help to reduce the variances much while introducing unnecessary extra bias compared with pure lasso or ridge regression. The prediction is improved by correcting this double shrinkage. The “corrected” elastic net estimate is given by $(1 + \lambda_2)\hat{\beta}(\text{naïve elastic net})$.

Adaptive Lasso

Zou (2006) introduces yet another variant of the lasso, the adaptive lasso. He constructs certain scenarios in which the lasso selection is not consistent, i.e. it does not improve the model selection accuracy as the sample size grows larger. Adaptive lasso is aimed to fixed this consistency problem by differently weighting the ℓ_1 -penalty for different coefficients.

Zou (2006) considers a weighted version of lasso

$$\arg \min_{\boldsymbol{\beta}} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j w_j |\beta_j|, \quad (2.37)$$

where $\mathbf{w}$ is a known vector of weights. If the weights are data-dependent and cleverly chosen, then the weighted lasso can achieve the oracle properties.

Suppose that $\hat{\boldsymbol{\beta}}$ is a root- n -consistent estimator, e.g. OLS estimator. For some $\gamma > 0$, define $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$. The adaptive lasso estimates are given by the solution to

$$\arg \min_{\boldsymbol{\beta}} \sum_i \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \hat{w}_j |\beta_j|. \quad (2.38)$$

Given a proper choice of λ , Zou (2006) shows that adaptive lasso has the oracle properties. It is also shown that it gives near-minimax optimality (see Zou (2006) for details).

Chapter 3

Adaptive Ridge Selector

3.1 Extending the Normal Linear Regression Model

Beginning with a standard hierarchical linear model (Sorensen and Gianola, 2002, Section 6.2.2) we propose a basic modification. In this case, the joint posterior of the parameters is proportional to,

$$p(\boldsymbol{\beta}, \mathbf{v}^{-1}, \sigma^2 | \mathbf{y}, H) \propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} | \sigma^2, \mathbf{v}^{-1}) p(\mathbf{v}^{-1} | \mu, \eta) p(\sigma^2). \quad (3.1)$$

Here a normal likelihood is assumed, $p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$, along with a conjugate normal prior on the regression coefficients, $\boldsymbol{\beta}$, and a typical Jeffreys prior on the error variance σ^2 ,

$$p(\boldsymbol{\beta} | \sigma^2, \mathbf{v}^{-1}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{V}) \quad (3.2)$$

$$p(\sigma^2) \propto 1/\sigma^2. \quad (3.3)$$

As with the relevance vector machine discussed in Section 2.3, the vector $\mathbf{v}^{-1} = \text{diag}(\mathbf{V}^{-1})$ where $\mathbf{V}$ is a diagonal matrix with elements v_j , $j = 1, \dots, p$ the reciprocals of which are independent and identically distributed from a gamma distribution,

$$p(v_j^{-1}) = \frac{\mu^{\eta+1}}{\Gamma(\eta+1)} v_j^{-\eta} \exp(-\mu v_j^{-1}) \quad (3.4)$$

where μ is the inverse scale parameter, and η is the shape parameter. By definition $v_j > 0$, $\mu > 0$ and $\eta > -1$. Notice that when $\eta = 0$, this becomes an exponential distribution which we'll consider as a special case.

Tipping (2001) sets $\eta = -1$ and $\mu = 0$ which leads to a scale invariant improper prior. He then derives a marginal likelihood $p(\mathbf{y}|\sigma^2, v_1^{-1}, \dots, v_p^{-1})$ through direct integration which is then maximized with respect to σ^2 and v_j^{-1} . Hypothetically as the algorithm proceeds some v_j s will tend toward 0 which correspond to the irrelevant variables in the model. Tipping (2001) does not check the validity of the joint posterior density having assumed an improper prior on v_j^{-1} .

In contrast to Tipping (2001), we choose not to integrate the regression coefficients out of the joint posterior distribution, but instead proceed to find the modal value given the data and η . Here we fix μ to be a very small number (e.g. machine epsilon) and adjust η to control shrinkage.

Sparsity is obtained via the combination of these two particular priors, $p(\beta_j|\sigma^2, v_j^{-1})$ and $p(v_j^{-1})$. Integration over v_j^{-1} in the joint prior distribution $p(\beta_j, v_j^{-1})$ will reveal that the marginal prior density of the regression coefficients is a product of univariate t densities, with a scale of $\sqrt{(\mu\sigma^2)/(\eta+1)}$ and degrees of freedom of $2\eta+2$. It is important to note that the product of univariate t -densities is not equivalent to a multivariate t and does not have elliptical contours but instead produces ridges along the axes. These ridges can be made more drastic by choosing the scale parameter

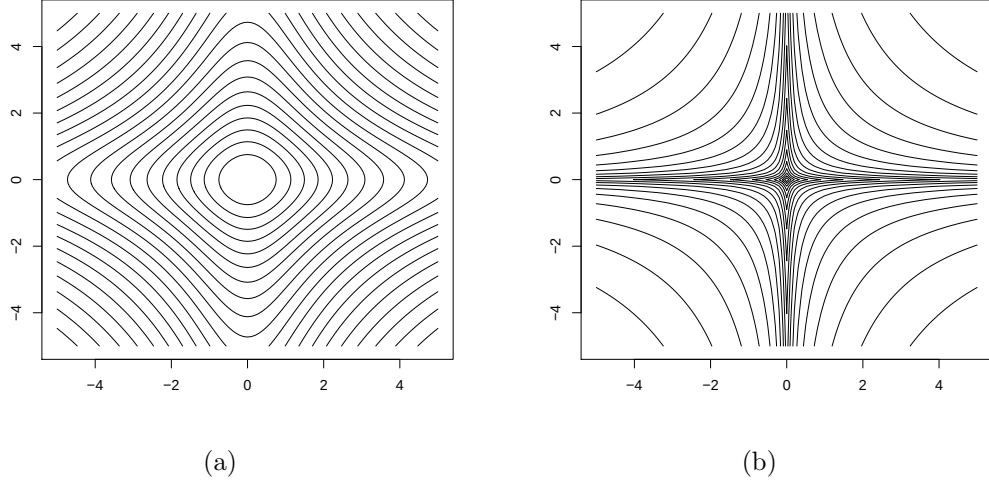


Figure 3.1: Contours of the penalty imposed by the independent (log) t -priors for $\mu = 1$ and $\mu = 0.001$.

to be small; see Figure 3.1. Then the posterior will be maximized where ever these ridges first touch the contours of the likelihood. The parameter η plays a very similar role to the regularization parameter of the ridge regression, lasso, etc., with larger values encouraging further shrinkage. Hence the proposed hierarchical structure implies independent t priors being placed on each regression coefficient. Direct use of such t priors would obscure the conjugate nature of the model. From an optimization perspective, a direct use of such priors leads to a non-convex objective function which would not be desirable. As we will discuss, within the hierarchical structure each iteration solves a simpler convex problem leading to an efficient solution.

Once v_j^{-1} s are integrated out of the joint posterior, the problem can be seen analogously as a regularized least squares problem as $\mu \rightarrow 0$ which solves

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p \log(\beta_j^2). \quad (3.5)$$

Although it may seem attractive to minimize this function, this is a multi-modal objective which we would like to avoid.

Parameter estimation and model selection are simultaneously accomplished in this model by maximizing the joint posterior distribution of the data and all random parameters, $p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \mathbf{y}|H)$. This solution is desirable for two reasons:

- i. This value is the mode of the posterior density and is asymptotically equivalent to the MLE. Besides asymptotic efficiency, the mode is also the Bayes estimator under a 0 – 1 type loss function, see (Sorensen and Gianola, 2002, pg. 999).

Suppose that the loss function has the following form Sorensen and Gianola (2002):

$$\mathcal{L}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \begin{cases} 0, & \text{if } |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \leq \mathbf{c} \\ 1, & \text{if } |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \mathbf{c} \end{cases} \quad (3.6)$$

where $\hat{\boldsymbol{\theta}}$ is a point estimator of $\boldsymbol{\theta} \in \mathbb{R}^d$, and $\mathbf{c} \in \mathbb{R}^d$ is a constant. The expected posterior loss is then

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y})} \mathcal{L}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= 0 \times P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \leq \mathbf{c}|\mathbf{y}) + 1 \times P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \mathbf{c}|\mathbf{y}) \\ &= 1 - P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \leq \mathbf{c}|\mathbf{y}) \\ &= 1 - \int_{\hat{\boldsymbol{\theta}} - \mathbf{c}}^{\hat{\boldsymbol{\theta}} + \mathbf{c}} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \end{aligned} \quad (3.7)$$

One can see that for some $\mathbf{c}$ chosen to be sufficiently small, this quantity will be minimized by the posterior mode.

The modal value will also provide information on which variables are active (which regression coefficients are non-zero) under the support of the likelihood. Once a model has been selected, other point estimators will also be available for parameters of interest such as posterior means.

- ii. In Section 3.1.3, we approximate the marginal likelihood integral using a Laplace approximation Sorensen and Gianola (2002). The Laplace approximation requires a Taylor series expansion around the mode of the function, hence it is useful to compute this estimator.

The joint posterior can be written as the product of the posterior distribution of the parameters and the marginal distribution of the data,

$$p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \mathbf{y}|H) = p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}, H)p(\mathbf{y}) \quad (3.8)$$

Since $p(\mathbf{y})$ is only a function of the observed measurements, it follows that

$$\arg \max_{\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}} p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \mathbf{y}|H) = \arg \max_{\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}} p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}, H). \quad (3.9)$$

Having specified a complete hierarchical model, the joint posterior distribution of the parameters is obtained by the product of the likelihood and the specified priors up to a normalizing constant as

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}, H) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}|\sigma^2, \mathbf{v}^{-1})p(\sigma^2)p(\mathbf{v}^{-1}|H) \\ &\propto (\sigma^2)^{-(\frac{n+p}{2}+1)} \prod_{j=1}^p v_j^{-1/2-\eta} \mu^{p(\eta+1)} \Gamma(\eta+1)^{-p} \\ &\quad \times \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta}}{2\sigma^2} \right\} \\ &\quad \times \exp \left\{ -\mu \sum_{j=1}^p v_j^{-1} \right\} \end{aligned} \quad (3.10)$$

where $H = (\eta, \mu)$.

Theorem 1. *Given the priors in (3.2), (3.3), (3.4), the product of these prior densities and the normal likelihood, (3.10) is the kernel of a posterior density function for $\boldsymbol{\beta}$, σ^2 , and $\mathbf{v}^{-1}$.*

Proof for this theorem can be found in Appendix C. The conditional distributions of the parameters can now easily be derived from this joint distribution.

- i. The regression coefficients are distributed as multivariate normal conditional on the error variance σ^2 and the prior covariance of the regression coefficients $\mathbf{V}$.

$$\boldsymbol{\beta} | \sigma^2, \mathbf{v}^{-1}, \mathbf{y} \sim N_p \left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{V}}^{-1} \sigma^2 \right), \quad (3.11)$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y}$ and $\tilde{\mathbf{V}} = \mathbf{X}'\mathbf{X} + \mathbf{V}^{-1}$.

- ii. The error variance is distributed as inverse gamma conditional upon all other parameters.

$$p(\sigma^2 | \boldsymbol{\beta}, \mathbf{v}^{-1}, \mathbf{y}) \propto (\sigma^2)^{-(\frac{n+p}{2}+1)} \times \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta}}{2\sigma^2} \right\} \quad (3.12)$$

Thus,

$$\sigma^2 | \boldsymbol{\beta}, \mathbf{v}^{-1}, \mathbf{y} \sim \text{inverse - gamma} (\nu^*, \lambda^*), \quad (3.13)$$

where $\nu^* = (n + p)/2$ and $\lambda^* = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta}}{2}$.

- iii. The prior precisions of the regression coefficients, conditional on all other parameters, follow a gamma distribution.

$$p(\mathbf{v}^{-1} | \boldsymbol{\beta}, \sigma^2, \mathbf{y}, H) \propto \prod_{j=1}^p (v_j^{-1})^{(\frac{1}{2} + \eta)} \exp \left\{ -\frac{\beta_j^2 + 2\sigma^2 \mu}{2\sigma^2} v_j^{-1} \right\}$$

$$\propto \prod_{j=1}^p (v_j^{-1})^{(\frac{3}{2}+\eta-1)} \exp \left\{ -\frac{\beta_j^2 + 2\sigma^2\mu}{2\sigma^2} v_j^{-1} \right\}. \quad (3.14)$$

Thus,

$$v_j^{-1} | \beta_j, \sigma^2, \mathbf{y}, H \sim \text{gamma}(\eta^*, \mu_j^*), \quad (3.15)$$

where $\eta^* = 3/2 + \eta$ and $\mu_j^* = (\beta_j^2 + 2\sigma^2\mu)/2\sigma^2$.

Deriving the full set of conditional distributions has several uses. As is frequently done, we may utilize these to simulate from the posterior distribution using Gibbs sampling. Such an approach would allow us to compute traditional Bayes estimators for the regression coefficients. In Section 3.1.1 we show how to use the conditional distributions to maximize the joint posterior in a surprisingly simple and effective way. Maximization will also facilitate computation of the Laplace approximation to the marginal likelihood (Tierney and Kadane, 1986).

3.1.1 Computing Posterior Modes

Lindley and Smith (1972) proposed an optimization algorithm to find the joint posterior modes; see also Chen et al. (2001). Once the fully conditional densities of the model parameters are obtained, it is possible to maximize the joint posterior distribution by iteratively maximizing these conditional densities.

Since the conditional posterior distributions obtained in equations (3.11), (3.13), and (3.15) are well-known distributions with readily available modes, the Lindley-Smith optimization algorithm becomes rather appealing to implement. The modes for the distributions in equations (3.11), (3.13), and (3.15) respectively are

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y} \quad (3.16)$$

$$\tilde{\sigma}^2 = \frac{\lambda^*}{\nu^* + 1} \quad (3.17)$$

$$\tilde{v}_j = \frac{\beta_j^2 + 2\sigma^2\mu}{(1 + 2\eta)\sigma^2}, \quad j = 1, 2, \dots, p \quad (3.18)$$

where ν^* , λ^* were defined in (3.13). The maximization proceeds through sequential re-estimation of $\tilde{\boldsymbol{\beta}}^{(l-1)}$, $\tilde{\sigma}^{2(l)}$, $\tilde{v}_j^{(l)}$, where $l = 1, \dots, m$, m is the number of iterations, and $\tilde{\boldsymbol{\beta}}^{(0)}$ is the OLS estimator.

For the Laplace approximation we will also need the negative Hessian evaluated at the posterior mode. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1})'$. The negative Hessian, $\mathbf{H}_{\boldsymbol{\theta}}$, is given by

$$-\frac{\partial^2}{\partial \beta_k \partial \beta_m} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = \begin{cases} \frac{1}{\sigma^2} (\sum_{i=1}^n x_{ik}^2 + v_k^{-1}), & k = m \\ \frac{1}{\sigma^2} (\sum_{i=1}^n x_{ik} x_{im}), & k \neq m \end{cases} \quad (3.19)$$

$$-\frac{\partial^2}{(\partial \sigma^2)^2} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = -\frac{\nu^* + 1}{\sigma^4} + \frac{2\lambda^*}{\sigma^6} \quad (3.20)$$

$$-\frac{\partial^2}{\partial v_k^{-1} \partial v_m^{-1}} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = \begin{cases} v_k^2 (\frac{1}{2} + \eta), & k = m \\ 0, & k \neq m \end{cases} \quad (3.21)$$

$$-\frac{\partial^2}{\partial \beta_k \partial v_m^{-1}} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = \begin{cases} \frac{\beta_k}{\sigma^2}, & k = m \\ 0, & k \neq m \end{cases} \quad (3.22)$$

$$-\frac{\partial^2}{\partial \sigma^2 \partial \beta_k} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = \frac{1}{\sigma^4} \left[\sum_{i=1}^n x_{ik} \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right) - \beta_k v_k^{-1} \right] \quad (3.23)$$

$$-\frac{\partial^2}{\partial \sigma^2 \partial v_k^{-1}} \log p(\mathbf{y}, \boldsymbol{\theta} | \eta) = -\frac{\beta_k^2}{2\sigma^4}. \quad (3.24)$$

3.1.2 Relation to Regularized LS

Given $\mathbf{V}$, the modal value of $\boldsymbol{\beta}$ can be obtained as a solution to a penalized least squares problem as with ridge regression. Since we have an iterative procedure, let

$v_j^{(l)}$ be the j th diagonal of $\mathbf{V}^{(l)}$ at the l th iteration and be given. Then the l th iterate for $\boldsymbol{\beta}$ is the solution to a similar penalized least squares problem:

$$\boldsymbol{\beta}^{(l)} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + \sum_{j=1}^p \frac{\beta_j^2}{v_j^{(l)}}. \quad (3.25)$$

If we substitute v_j with the estimate from (3.18) and let $\mu \rightarrow 0$, we obtain

$$\boldsymbol{\beta}^{(l)} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 + (1 + 2\eta) \sum_{j=1}^p \frac{\beta_j^2}{\omega_j^{(l)}}, \quad (3.26)$$

where $\omega_j^{(l)} = +\sqrt{\beta_j^{2(l-1)}/\sigma^{2(l)}}$. This procedure is essentially re-weighting the predictor variables by the positive square root of the ratio between the current estimate of the coefficients and the residual variance due to them. After this re-weighting procedure, the problem takes the form of a standard ridge regression,

$$\boldsymbol{\beta}^{*(l)} = \arg \min_{\boldsymbol{\beta}^*} \sum_{i=1}^n \left(y_i - \mathbf{x}_i^{*(l-1)} \boldsymbol{\beta}^* \right)^2 + (1 + 2\eta) \sum_{j=1}^p \beta_j^{*2}, \quad (3.27)$$

where $\beta_j^* = \beta_j/\omega_j$, $\mathbf{x}_i^{*(l)} = \mathbf{x}_i^{*(l-1)} \boldsymbol{\omega}^{*(l)}$ and $\mathbf{x}_i^{*(0)} = \mathbf{x}_i$. The solution to the problem above at iteration l is given by

$$\boldsymbol{\beta}^{*(l)} = (\mathbf{X}^{*'(l)} \mathbf{X}^{*(l)} + (1 + 2\eta) \mathbf{I})^{-1} \mathbf{X}^{*'(l)} \mathbf{y}. \quad (3.28)$$

Hence the mode is computed through a sequence of re-weighted ridge estimators. The final estimate $\tilde{\boldsymbol{\beta}}^{(m)}$ then can be recovered as $\tilde{\boldsymbol{\beta}}^{*(m)} \times \left(\prod_{l=1}^m \omega_1^{(l)}, \dots, \prod_{l=1}^m \omega_p^{(l)} \right)'$ (this multiplication is understood component wise). Note that when $\eta = -1/2$, this procedure results in the OLS estimator.

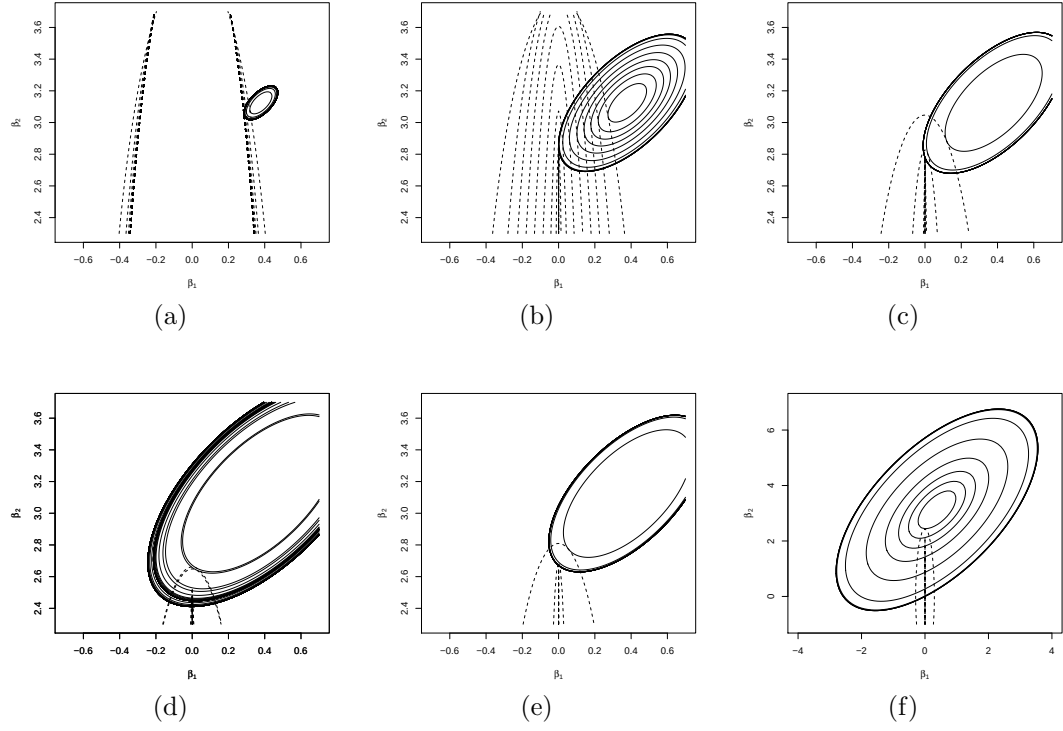


Figure 3.2: Progression of the ARiS algorithm with η values of $-0.25, 0, 2, 5, 10, 20$. Solid and dashed lines indicate the level sets of the squares residuals and the adapted constrained region at each iteration respectively.

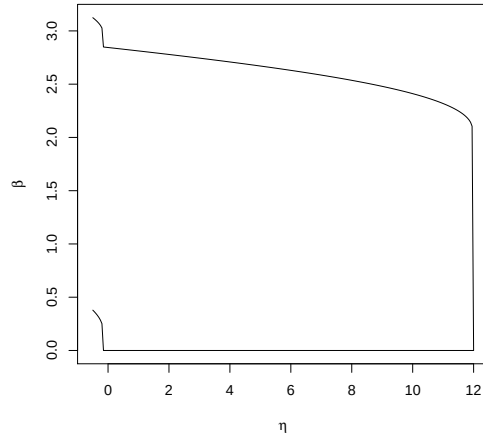


Figure 3.3: The solution path for the toy example with respect to η .

Table 3.1: Estimators for each η

η	β_1	β_2
-0.5	0.3787	3.1242
-0.25	0.2850	3.0521
0	0	2.8441
2	0	2.7789
5	0	2.6703
10	0	2.4110
≥ 12	0	0

We construct a two-dimensional example to illustrate the method . Consider the model $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \mathcal{N}(0, 1)$, where the true $\beta_1 = 0$ and $\beta_2 = 3$, and the correlation between two predictors is -0.5 . We generate 30 observations and run ARiS for $\eta = 0$. ARiS iteratively updates the constrained region and converges to a solution. Figure 3.2 (a) - (f) clearly demonstrates how the shrinkage occurs through the adaptation of constrained regions for different η values. The constrained region eventually becomes singular along the dimension which has no contribution to the response in the underlying model. Figure 3.3 also shows the solution path of the algorithm with respect to η . This path is obtained by increasing η values in increments of 0.1. As η grows sufficiently large, both coefficients are zeroed. Table 3.1 also presents the estimated values of β_1 and β_2 for $\eta = \{-0.5, -0.25, 0, 2, 5, 10\}$. Recall that $\eta = -0.5$ case corresponds to the OLS estimator.

3.1.3 Approximations for the marginal likelihood.

Critical to the ARiS procedure is the choice of the η . We propose empirical Bayes estimation of η through the maximization of the marginal likelihood $p(\mathbf{y}|\eta)$.

Hence, we must integrate the joint posterior over all parameters,

$$p(\mathbf{y}|\eta) = \int_{\boldsymbol{\theta}} p(\mathbf{y}, \boldsymbol{\theta}|\eta) d\boldsymbol{\theta}, \quad (3.29)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1})'$. In the case of the hierarchical model developed in Section 3.1, the direct calculation is intractable. Below we propose both analytical and simulation-based approximations.

Laplace approximation

A standard analytical approximation of the marginal likelihood can be computed using the Laplacian method (Tierney and Kadane, 1986). The approximation is obtained as

$$\log(p(\mathbf{y}|\eta)) \approx \log \left[p(\mathbf{y}, \tilde{\boldsymbol{\theta}}|\eta) \right] + \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{H}_{\tilde{\boldsymbol{\theta}}}|, \quad (3.30)$$

where $\tilde{\boldsymbol{\theta}}$ is the mode of the joint posterior and $\mathbf{H}_{\tilde{\boldsymbol{\theta}}}$ is the Hessian matrix evaluated at the posterior mode.

Recall that the ARiS is designed to drive the values of β_j and v_j to zero for those independent variables $\mathbf{x}_j$ which provide no explanatory value. As $\mu \rightarrow 0$, the prior precisions and the regression coefficients related to irrelevant independent variables will tend toward ∞ and 0 respectively. In fact we can see in (3.19) that along these β_j the curvature approaches ∞ as we converge to the solution thus driving their variance to 0. At the joint posterior mode, the corresponding dimensions of $\mathbf{X}$ do not contribute and become irrelevant. Under the support of the data, we claim these variables to be insignificant and suggest their removal from the model. The integration follows removal of these irrelevant variables from the model.

The resulting Laplace approximation to the log-marginal likelihood is

$$\log p(\mathbf{y}|\eta) \approx \log p(\tilde{\boldsymbol{\beta}}^\dagger, \tilde{\sigma}^2, \tilde{\mathbf{v}}^{-1\dagger}, \mathbf{y}|H) + \frac{p^\dagger}{2} \log(2\pi) - \frac{1}{2} \log \left| \mathbf{H}_{\tilde{\boldsymbol{\beta}}^\dagger, \tilde{\sigma}^2, \tilde{\mathbf{v}}^{-1\dagger}} \right|, \quad (3.31)$$

where $(.)^\dagger$ represents the reduced model after the removal of the irrelevant variables at the mode.

Marginal Likelihood Calculation via Importance Sampling

As will be seen in Section 4.1, Laplace approximation may fail in certain circumstances. We therefore consider other approaches to computing the marginal likelihood. $\boldsymbol{\beta}$ and σ^2 can be analytically integrated out of the joint posterior given in (3.10). The resulting marginal likelihood conditioned on the prior variances is,

$$p(\mathbf{y}|\mathbf{v}^{-1}) \propto |\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1}|^{-1/2} |\mathbf{V}|^{-1/2} (\lambda + S^2)^{-(n+\nu)/2}, \quad (3.32)$$

where

$$S^2 = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y}; \quad (3.33)$$

see also (Chipman et al., 2001, eqs 3.11,3.12).

The marginal likelihood conditioned over η can now be obtained through integration as

$$\begin{aligned} p(\mathbf{y}|\eta) &= \int_{\mathbf{v}^{-1}} p(\mathbf{y}|\mathbf{v}^{-1}) p(\mathbf{v}^{-1}|\eta) d\mathbf{v}^{-1} \\ &= \mathbb{E}_{\mathbf{v}^{-1}|\eta} [p(\mathbf{y}|\mathbf{v}^{-1})] \end{aligned} \quad (3.34)$$

where the expectation is taken over the prior distribution of $\mathbf{v}^{-1}$.

The quantity of interest is the marginal likelihood conditioned over the parameter η which controls the amount of shrinkage imposed on the model size. Although the integration over β and σ^2 is analytically tractable as shown above, the integration over $\mathbf{v}^{-1}$ is not. The marginal likelihood of the data given η can then be expressed as

$$\begin{aligned} p(\mathbf{y}|\eta) &= \frac{\int p(\mathbf{y}|\mathbf{v}^{-1})p(\mathbf{v}^{-1}|\eta)d\mathbf{v}^{-1}}{\int p(\mathbf{v}^{-1}|\eta)d\mathbf{v}^{-1}} \\ &= \frac{\int p(\mathbf{y}|\mathbf{v}^{-1})\frac{p(\mathbf{v}^{-1}|\eta)}{g(\mathbf{v}^{-1})}g(\mathbf{v}^{-1})d\mathbf{v}^{-1}}{\int \frac{p(\mathbf{v}^{-1}|\eta)}{g(\mathbf{v}^{-1})}g(\mathbf{v}^{-1})d\mathbf{v}^{-1}} \end{aligned} \quad (3.35)$$

since $\int p(\mathbf{v}^{-1}|\eta)d\mathbf{v}^{-1} = 1$. Here $g(\mathbf{v}^{-1})$ represents a convenient proposal density. In this particular case, because the prior distribution over v_j^{-1} is very flat, we use a uniform distribution over the joint space of $\prod v_j^{-1}$. Once the samples are generated from this uniform distribution, they are then re-weighted according to a gamma density. We choose to do this since directly sampling from the prior may result in inefficient estimates of $p(\mathbf{y}|\eta)$; see Kass and Raftery (1995). In cases where the posterior is concentrated relative to the prior, most of the samples obtained from the prior distribution will be in regions that are not supported by the data. Only a few samples will come from the likelihood region which will eventually lead to an estimate that is an average of a few “useful” samples. Let us write a consistent estimator for the marginal likelihood given above. The pieces in the numerator and the denominator of (3.35) can respectively be estimated as

$$\begin{aligned} \lim_{m \rightarrow \infty} \left[\frac{1}{m} \sum_{j=1}^m p(\mathbf{y}|\mathbf{v}^{-1[j]}) \frac{p(\mathbf{v}^{-1[j]}|\eta)}{g(\mathbf{v}^{-1[j]})} \right] &= \int p(\mathbf{y}|\mathbf{v}^{-1}) \frac{p(\mathbf{v}^{-1}|\eta)}{g(\mathbf{v}^{-1})} g(\mathbf{v}^{-1}) d\mathbf{v}^{-1} \\ \lim_{m \rightarrow \infty} \left[\frac{1}{m} \sum_{j=1}^m \frac{p(\mathbf{v}^{-1[j]}|\eta)}{g(\mathbf{v}^{-1[j]})} \right] &= \int \frac{p(\mathbf{v}^{-1}|\eta)}{g(\mathbf{v}^{-1})} g(\mathbf{v}^{-1}) d\mathbf{v}^{-1}. \end{aligned} \quad (3.36)$$

Setting $w^{[j]} = p(\mathbf{v}^{-1[j]}|\eta)/g(\mathbf{v}^{-1[j]})$, for large m ,

$$\hat{p}(\mathbf{y}|\eta) = \frac{\sum_{j=1}^m w^{[j]} p(\mathbf{v}^{-1[j]}|\eta)}{\sum_{j=1}^m w^{[j]}}, \quad (3.37)$$

which is a weighted average of the density of the sampling distribution evaluated at the corresponding samples values of the parameter.

In order to ensure efficient sampling, we define a hypercube around the mode of the joint posterior in order to obtain a sampling region over $\prod_j v_j^{-1}$. The sampling region is the set $\left\{ v_j^{-1} | \max(0, \tilde{v}_j^{-1} - k\sigma_{v_j^{-1}}) < v_j^{-1} < \tilde{v}_j^{-1} + k\sigma_{v_j^{-1}} \right\}$ where $\tilde{v}_j^{-1}$ is the modal value of v_j^{-1} , $\sigma_{v_j^{-1}}$ is the inverse curvature at the mode and k is to be chosen to adjust the width of the box.

3.1.4 The Marginal Posterior Mode of β via EM

An expectation-maximization approach may be used to obtain the marginal posterior mode of β . Consider the identity

$$p(\beta|\mathbf{y}) = \frac{p(\beta, \sigma^2, \mathbf{v}^{-1}|\mathbf{y})}{p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \beta)}. \quad (3.38)$$

Taking the logarithm and then taking the expectations of both sides with respect to $p(\sigma^2, \mathbf{v}^{-1}|\beta^{(l)})$ yields

$$\begin{aligned} \log p(\beta|\mathbf{y}) &= \log p(\beta, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}) - \log p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \beta) \\ &= \int \log p(\beta, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}) p(\sigma^2, \mathbf{v}^{-1}|\beta^{(l)}) d\sigma^2 d\mathbf{v}^{-1} \\ &\quad - \int \log p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \beta) p(\sigma^2, \mathbf{v}^{-1}|\beta^{(l)}) d\sigma^2 d\mathbf{v}^{-1}, \end{aligned} \quad (3.39)$$

where $\boldsymbol{\beta}^{(l)}$ is the current guess of $\boldsymbol{\beta}$ (Sorensen and Gianola, 2002, pg. 446). The EM algorithm involves working with the first term of (3.39). The EM procedure in our case would consist of the following two steps: (i) expectation of $\log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y})$ with respect to $p(\sigma^2, \mathbf{v}^{-1}|\boldsymbol{\beta}^{(l)})$, (ii) maximization of the expected value with respect to $\boldsymbol{\beta}$. An iterative procedure results by replacing the initial guess $\boldsymbol{\beta}^{(l)}$ with the solution of the maximization procedure $\boldsymbol{\beta}^{(l+1)}$ and repeating (i) and (ii) until convergence.

With a slight change in the hierarchical model used, the above expectation will become quite trivial. Unlike (3.2), let us not condition the prior density of $\boldsymbol{\beta}$ on σ^2 . Under such a setup, $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)}) = p(\sigma^2|\mathbf{y}, \boldsymbol{\beta}^{(l)}) p(\mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)})$. Notice that the conditional posteriors in (3.13) and (3.15) now become

$$p(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right\}, \quad (3.40)$$

$$p(v_j^{-1}|\boldsymbol{\beta}, \mathbf{y}, H) \propto (v_j^{-1})^{(\frac{3}{2}+\eta-1)} \exp \left\{ -\left(\frac{\beta_j^2}{2} + \mu \right) v_j^{-1} \right\}. \quad (3.41)$$

Given the new prior, let us re-write (3.10) in the log form excluding the terms that do not depend on $\boldsymbol{\beta}$, σ^2 and $\mathbf{v}^{-1}$:

$$\begin{aligned} \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}, H) &\propto -(n/2 + 1) \log(\sigma^2) + (1/2 + \eta) \sum_{j=1}^p \log v_j^{-1} \\ &\quad - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{\boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta}}{2} - \mu \sum_{j=1}^p v_j^{-1} \end{aligned} \quad (3.42)$$

Next we compute $\mathbb{E}_{\mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \mathbb{E}_{\sigma^2|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}|\mathbf{y}, H)$ as $\mu \rightarrow 0$:

$$\begin{aligned} &-(n/2 + 1) \mathbb{E}_{\sigma^2|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \log(\sigma^2) + (1/2 + \eta) \mathbb{E}_{\mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \left(\sum_{j=1}^p \log v_j^{-1} \right) \\ &- \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2S^{2(l)}/n} - \left(\eta + \frac{3}{2} \right) \sum_{j=1}^p \frac{\beta_j^2}{\beta_j^{2(l)}}, \end{aligned} \quad (3.43)$$

where $S^{2(l)} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)})$.

Having completed the expectation step, the maximization of (3.43) with respect to $\boldsymbol{\beta}$ yields an estimator as the solution to the following sequence of convex minimization problems:

$$\boldsymbol{\beta}^{(l+1)} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (2\eta + 3) \sum_{j=1}^p \frac{\beta_j^2}{n\beta_j^{2(l)}/S^{2(l)}} \quad (3.44)$$

Hence, the marginal posterior mode of $\boldsymbol{\beta}$ is extremely similar in form to the joint posterior mode. Note that we can still adopt the weighting perspective mentioned earlier. Recall from Section 3.1 that the integration over v_j^{-1} in the prior distribution of β_j results in a univariate t -density with degrees of freedom $2\eta + 2$. Therefore, a value of $\eta = -3/2$ will actually lead to a flat prior over β_j resulting in the OLS estimator (note that when $\eta = -3/2$, the kernel of the t density has power 0 resulting in a flat density). Also, the solution to the marginal when $\eta = -1$ will be identical to the maximization of the joint posterior when $\eta = 0$.

Harville (1977) mentions that the estimator of Lindley and Smith (1972) based on joint maximization may be far from the Bayes estimator and suggests that the maximization of the marginal mode of the variance components would be a superior approach. Above we have shown that in our case the mode of the marginal density has the same form as the joint mode justifying the conditional maximization approach. In fact, Tipping adopts the approach suggested by Harville, maximizing the joint posterior mode of $\mathbf{v}^{-1}$ and σ^2 after integrating over $\boldsymbol{\beta}$ but still achieves the zeroing effect.

We needed a slight change in the model to ease our work for the expectation step above, that is, we made the prior distribution of the regression coefficients independent of the error variance. One may think that while we were trying to show the

equivalence of these two solutions (the joint and the marginal solutions), we actually created two different models and show that their solutions are identical in form, yet, they do not follow the same model. In the traditional Bayesian analysis of the linear regression models, the regression coefficients are conditioned over the noise variance. This provides an estimator for the regression coefficients that is independent of the noise variance. We followed this convention when we were forming our hierarchical model. However, in our case, there is no such thing as independence between the solutions of the regression coefficients and the noise variance. Although in an explicit statement such as (3.16) it may seem that the solution for the regression coefficients does not depend on the error variance, there exists an implicit dependence through the solution of v_j^{-1} s. We could have very well constructed our hierarchical model using a prior on regression coefficients independent of the noise variance. This would lead to a solution that is only slightly different. The re-estimation equation in (3.16), (3.17) and (3.18) would become

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y} \quad (3.45)$$

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}{n + 2} \quad (3.46)$$

$$\tilde{v}_j = \frac{\beta_j^2 + 2\mu}{1 + 2\eta}, \quad j = 1, 2, \dots, p. \quad (3.47)$$

Now, the solution for v_j is independent of σ^2 , yet the solution of $\boldsymbol{\beta}$ explicitly depends on σ^2 . That said, the implicit dependence has become an explicit one. Thus the iterative solution for the regression coefficients as $\mu \rightarrow 0$ can be written as

$$\boldsymbol{\beta}^{(l+1)} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (2\eta + 1) \sum_{j=1}^p \frac{\beta_j^2}{\beta_j^{2(l)} / \sigma^{2(l)}} \quad (3.48)$$

in which, apart from the tuning quantity $(2\eta + 1)$, the only difference with (3.44) is the plug-in estimator used for the noise variance.

Having shown that these procedures are fundamentally identical in form to each other, let us discuss another important point, the choice of initial values to start the algorithm. Let us consider the solution following (3.44) with $\eta = -1$:

$$\boldsymbol{\beta}^{(l+1)} = \left(\mathbf{X}'\mathbf{X} + \begin{bmatrix} S^{2(l)}/n\beta_1^{2(l)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & S^{2(l)}/n\beta_p^{2(l)} \end{bmatrix} \right)^{-1} \mathbf{X}'\mathbf{y} \quad (3.49)$$

We cannot just plug any $\boldsymbol{\beta}^{(0)}$ as an initial estimator. Consider $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. In this case all the regression coefficients will be zeroed. Or let only a subset of $\boldsymbol{\beta}^{(0)}$ be zero. Then in the solution those components will remain zero. Although we are solving a series of simple convex problems, the dependency of the solution to the initial value proves that here we are dealing with a multi-modal objective function as would be expected. Thus using the OLS estimator as an initial value will take us to a local stationary point which is most likely under the support of the data in hand.

To gain further intuition, let us consider an orthogonal case and let the predictors be scaled so that they have unit 2-norm, i.e. $\mathbf{X}'\mathbf{X} = \mathbf{I}$. In such a case the OLS estimator for $\boldsymbol{\beta}$ would have a variance-covariance matrix $\hat{\sigma}^2\mathbf{I}$ where $\hat{\sigma}^2$ is a plug-in estimator, e.g. the maximum likelihood estimator or the bias corrected estimator. Testing the null hypothesis $H_0 : \beta_j = 0$, a t -statistic can be computed for a component $\hat{\beta}_j$ as $\hat{\beta}_j/\hat{\sigma}$. Notice in (3.49) the quantities at the diagonal of the second piece under the matrix inverse operation, $S^{2(l)}/n\beta_j^{2(l)}$, resemble the inverse of a squared t -statistic. In fact, recall from Section 3.1.2 that we formed a sequence of ridge regression problems

out of this procedure by re-weighting our predictors by $|\beta_j^{(l-1)}/\sigma^{(l)}|$ which is the absolute value of a t -statistic. Thus, following a conventional testing procedure, those predictors which correspond to coefficients with larger t -statistics will be given more importance. This is yet another point that intuitively explains our procedure.

3.2 Developing ARiS with a Scale Mixtures of Normals

There may be cases where a normal likelihood assumption is not feasible. If the data is contaminated by outliers, one may consider to use a likelihood that is going to be more tolerating to these extreme observations. A Student- t likelihood quickly comes to mind with its heavier tails and ability to accommodate observations that would not be likely to observe under a normal process. In Section 2.2 we introduced a Bayesian hierarchical model by Geweke (1993) that is equivalent to a joint posterior under a Student- t likelihood. Now we will exploit this nice hierarchical structure.

Consider the following hierarchical model:

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}, \mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Omega}) \quad (3.50)$$

$$\boldsymbol{\beta}|\sigma^2, \mathbf{v}^{-1} \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{V}) \quad (3.51)$$

$$\sigma^2 \propto 1/\sigma^2 \quad (3.52)$$

$$v_j^{-1}|\eta, \mu \sim \text{gamma}(\eta + 1, \mu) \quad (3.53)$$

$$\omega_i|\xi \sim \text{inv-gamma}(\nu/2, \nu/2) \quad (3.54)$$

where $i = 1, \dots, n$ and $j = 1, \dots, p$, the vector $\mathbf{v}^{-1} = \text{diag}(\mathbf{V}^{-1})$ where $\mathbf{V}$ is a diagonal matrix with elements v_j , $j = 1, \dots, p$ and the vector $\boldsymbol{\omega} = \text{diag}(\boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ is a

diagonal matrix with elements ω_i , $i = 1, \dots, n$.

The joint posterior density can be written up to a normalizing constant as

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega} | \mathbf{y}, H) &\propto p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta} | \sigma^2, \mathbf{v}^{-1}) p(\sigma^2) p(\mathbf{v}^{-1} | H) p(\boldsymbol{\omega}) \\
&\propto (\sigma^2)^{-(\frac{n+p}{2}+1)} \prod_{j=1}^p v_j^{-1/2-\eta} \mu^{p(\eta+1)} \Gamma(\eta+1)^{-p} \\
&\quad \times \prod_{i=1}^n \omega_j^{-(\nu+1)/2} \nu^{n\nu/2} \Gamma(\nu/2)^{-n} \\
&\quad \times \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{V}^{-1}\boldsymbol{\beta}}{2\sigma^2} \right\} \\
&\quad \times \exp \left\{ -\mu \sum_{j=1}^p v_j^{-1} \right\} \exp \left\{ -\frac{\nu}{2} \sum_{i=1}^n \omega_j^{-1} \right\} \quad (3.55)
\end{aligned}$$

where $H = (\eta, \mu, \nu)$.

This hierarchical structure leads to an analytically intractable posterior, yet another very attractive structure from which the fully conditional posterior densities of the unobservables can be obtained.

- i. The regression coefficients are distributed as multivariate normal conditional on all the other unobservables.

$$\boldsymbol{\beta} | \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega}, \mathbf{y} \sim N_p \left(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{V}}^{-1} \sigma^2 \right), \quad (3.56)$$

where $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ and $\tilde{\mathbf{V}} = \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} + \mathbf{V}^{-1}$.

- ii. The error variance is distributed as inverse gamma conditional on all other unobservables.

$$\begin{aligned}
p(\sigma^2 | \boldsymbol{\beta}, \mathbf{v}^{-1}, \boldsymbol{\omega}, \mathbf{y}) &\propto (\sigma^2)^{-(\frac{n+p}{2}+1)} \\
&\quad \times \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right\}
\end{aligned}$$

$$\times \exp \left\{ -\frac{\boldsymbol{\beta}' \mathbf{V}^{-1} \boldsymbol{\beta}}{2\sigma^2} \right\} \quad (3.57)$$

Thus,

$$\sigma^2 | \boldsymbol{\beta}, \mathbf{v}^{-1}, \boldsymbol{\Omega}, \mathbf{y} \sim \Gamma^{-1}(\nu^*, \lambda^*), \quad (3.58)$$

where $\nu^* = (n + p)/2$ and $\lambda^* = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}' \mathbf{V}^{-1} \boldsymbol{\beta}}{2}$.

- iii. The prior precisions of the regression coefficients, conditional on all other unobservables, follow a gamma distribution.

$$\begin{aligned} p(\mathbf{v}^{-1} | \boldsymbol{\beta}, \sigma^2, \mathbf{y}, H) &\propto \prod_{j=1}^p (v_j^{-1})^{(\frac{1}{2} + \eta)} \exp \left\{ -\frac{\beta_j^2 + 2\sigma^2 \mu}{2\sigma^2} v_j^{-1} \right\} \\ &\propto \prod_{j=1}^p (v_j^{-1})^{(\frac{3}{2} + \eta - 1)} \exp \left\{ -\frac{\beta_j^2 + 2\sigma^2 \mu}{2\sigma^2} v_j^{-1} \right\}. \end{aligned} \quad (3.59)$$

Thus,

$$v_j^{-1} | \beta_j, \sigma^2, \mathbf{y}, H \sim \Gamma(\eta^*, \mu_j^*), \quad (3.60)$$

where $\eta^* = 3/2 + \eta$ and $\mu_j^* = (\beta_j^2 + 2\sigma^2 \mu)/2\sigma^2$.

- iv. The weights, conditional on all other unobservables, follow an inverse gamma distribution.

$$p(\boldsymbol{\omega} | \boldsymbol{\beta}, \sigma^2, \mathbf{y}, H) \propto \prod_{i=1}^p \omega_i^{-(\nu+3)/2} \exp \left\{ -\frac{(y_i - \mathbf{x}_i \boldsymbol{\beta})^2 / \sigma^2 + \nu}{2\omega_i} \right\}. \quad (3.61)$$

Thus,

$$\omega_i | \boldsymbol{\beta}, \sigma^2, \mathbf{y}, H \sim inv - \Gamma(\nu^\dagger, \nu_i^\dagger), \quad (3.62)$$

where $\nu^\dagger = (\nu + 1)/2$ and $\nu_i^\dagger = \frac{(y_i - \mathbf{x}_i \boldsymbol{\beta})^2 / \sigma^2 + \nu}{2}$.

3.2.1 Computing Posterior Modes

We can use the Lindley-Smith procedure (Lindley and Smith, 1972) to maximize this joint posterior density since we have all the fully conditional densities available. The modes for the distributions in equations (3.56), (3.58), and (3.60) and (3.62) respectively are

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} \quad (3.63)$$

$$\tilde{\sigma}^2 = \frac{\lambda^*}{\nu^* + 1} \quad (3.64)$$

$$\tilde{v}_j = \frac{\beta_j^2 + 2\sigma^2\mu}{(1 + 2\eta)\sigma^2}, \quad j = 1, 2, \dots, p \quad (3.65)$$

$$\tilde{\omega}_i = \frac{(y_i - \mathbf{x}_i\boldsymbol{\beta})^2/\sigma^2 + \nu}{\nu + 3} \quad (3.66)$$

where ν^* , λ^* were defined in (3.58). The maximization proceeds through sequential re-estimation of $\tilde{\boldsymbol{\beta}}^{(l-1)}$, $\tilde{\sigma}^{2(l)}$, $\tilde{v}_j^{(l)}$, $\tilde{\omega}_i^{(l)}$ where $l = 1, \dots, m$, m is the number of iterations, and $\tilde{\boldsymbol{\beta}}^{(0)}$ is the OLS estimator.

3.2.2 The Joint Posterior Mode of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ via EM

An expectation-maximization approach may be used to obtain the joint posterior mode of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. This can be obtained as

$$\begin{aligned} \log p(\boldsymbol{\beta}, \boldsymbol{\omega}|\mathbf{y}) &= \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega}|\mathbf{y}) - \log p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}) \\ &= \int \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega}|\mathbf{y}) p(\sigma^2, \mathbf{v}^{-1}|\boldsymbol{\beta}^{(l)}, \boldsymbol{\omega}^{(l)}) d\sigma^2 d\mathbf{v}^{-1} \\ &\quad - \int \log p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}) p(\sigma^2, \mathbf{v}^{-1}|\boldsymbol{\beta}^{(l)}, \boldsymbol{\omega}^{(l)}) d\sigma^2 d\mathbf{v}^{-1}, \quad (3.67) \end{aligned}$$

where $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{\omega}^{(l)}$ are the current guesses for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. The EM procedure in this case would follow these two steps: (i) expectation of $\log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega}|\mathbf{y})$ with

respect to $p(\sigma^2, \mathbf{v}^{-1}|\boldsymbol{\beta}^{(l)}, \boldsymbol{\omega}^{(l)})$, (ii) maximization of the expected value with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. An iterative procedure results by replacing the initial guesses $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{\omega}^{(l)}$ with the solution of the maximization procedure $\boldsymbol{\beta}^{(l+1)}$ and $\boldsymbol{\omega}^{(l+1)}$ repeating (i) and (ii) until convergence.

We make the same slight change to (3.51) as we have in Section 3.1.4. This helps us again write $p(\sigma^2, \mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)}) = p(\sigma^2|\mathbf{y}, \boldsymbol{\beta}^{(l)}) p(\mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)})$. Notice that the conditional posteriors in (3.58) and (3.60) now become

$$p(\sigma^2|\boldsymbol{\beta}, \mathbf{y}) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp \left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right\}, \quad (3.68)$$

$$p(v_j^{-1}|\boldsymbol{\beta}, \mathbf{y}, H) \propto (v_j^{-1})^{(\frac{3}{2}+\eta-1)} \exp \left\{ -\left(\frac{\beta_j^2}{2} + \mu \right) v_j^{-1} \right\}. \quad (3.69)$$

Given the new prior, let us re-write (3.55) in the log form excluding the terms that do not depend on $\boldsymbol{\beta}$, σ^2 , $\mathbf{v}^{-1}$ and $\boldsymbol{\omega}$:

$$\begin{aligned} \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{V}|\mathbf{y}, H) &\propto -(n/2 + 1) \log(\sigma^2) + (1/2 + \eta) \sum_{j=1}^p \log v_j^{-1} \\ &\quad - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{\boldsymbol{\beta}' \mathbf{V}^{-1} \boldsymbol{\beta}}{2} - \mu \sum_{j=1}^p v_j^{-1} \\ &\quad - \frac{\nu + 3}{2} \sum_{i=1}^n \log(\omega_i) - (\nu/2) \sum_{i=1}^n \omega_i^{-1} \end{aligned} \quad (3.70)$$

Next we compute $\mathbb{E}_{\mathbf{v}^{-1}|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \mathbb{E}_{\sigma^2|\mathbf{y}, \boldsymbol{\beta}^{(l)}} \log p(\boldsymbol{\beta}, \sigma^2, \mathbf{v}^{-1}, \boldsymbol{\omega}|\mathbf{y}, H)$. Using the identity $p(\log \alpha) = \alpha p(\alpha)$ for a real valued random variable α , this expectation is obtained as

$$\begin{aligned} &- \frac{n+2}{2n-8} S^{2(l)} + \left(\frac{1}{2} + \eta \right) (2\eta + 5) \sum_{j=1}^p \frac{1}{\beta_j^{2(l)}} - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2S^{2(l)}/n} \\ &- \left(\eta + \frac{3}{2} \right) \sum_{j=1}^p \frac{\beta_j^2}{\beta_j^{2(l)}} - \frac{\nu + 3}{2} \sum_{i=1}^n \log(\omega_i) - (\nu/2) \sum_{i=1}^n \omega_i^{-1}, \end{aligned} \quad (3.71)$$

where $S^{2(l)} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(l)})$.

Having completed the expectation step, the maximization of (3.71) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ yields a re-weighted estimator:

$$\boldsymbol{\beta}^{(l+1)} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1(l)} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (2\eta + 3) \sum_{j=1}^p \frac{\beta_j^2}{n\beta_j^{2(l)}/S^{2(l)}} \quad (3.72)$$

$$\omega_i^{(l+1)} = \left[\frac{(y_i - \mathbf{x}_i\boldsymbol{\beta}^{(l+1)})^2}{S^{2(l)}/n} + \nu \right] / (\nu + 3), \quad (3.73)$$

where $\boldsymbol{\beta}^{(0)}$ is the OLS estimator and $\boldsymbol{\omega}^{(0)} = \mathbf{1}$.

Chapter 4

Experimental Results

4.1 Simulated Data Sets

Normal Setup

This section reports the results of a simulation study comparing the ARiS estimates with a number of computationally efficient penalized least squares methods. In the study we consider a model of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathcal{N}(0, \sigma^2)$. For each data set, we center $\mathbf{y}$ and scale the columns of $\mathbf{X}$ so that they have unit 2-norm. The lasso and elastic net were fit using the `lars` and `elasticnet` libraries in R.

Model 0: This model, adopted from Zou (2006), is a special case where the lasso estimate fails to improve as sample size increases. The true regression coefficients are $\boldsymbol{\beta} = (5.6, 5.6, 5.6, 0)$. The predictors $\mathbf{x}_i$ ($i = 1, \dots, n$) are iid $\mathcal{N}(\mathbf{0}, \mathbf{C})$ where $\mathbf{C}$ is defined in Zou (2006) (Corollary 1, pg. 1420) with $\rho_1 = -.39$ and $\rho_2 = .23$. Under this scenario, $\mathbf{C}$ does not allow consistent lasso selection. Zou (2006) proposes the adaptive lasso for consistent model selection. In this setting, we simulate 1000 data sets from the above model for different combinations of sample size and error variance. Table 4.1 reports the proportion of the cases where the solution paths included the

Table 4.1: Results for model 0.

	$n = 60, \sigma = 9$	$n = 120, \sigma = 5$	$n = 300, \sigma = 3$
<i>Lasso</i>	0.499	0.489	0.498
<i>AdaLasso</i>	0.724	0.935	0.996
<i>ARiS</i>	0.671	0.927	1
<i>ARiS</i> ($\eta = 0$)	0.462	0.895	0.944

true model for ARiS, lasso and adaptive lasso. We also report the results for the ARiS algorithm in the special case of $\eta = 0$. The results indicate that the ARiS algorithm performs nearly as well as the adaptive lasso and far better than the ordinary lasso in terms of consistent model selection under this particular setting. For $\eta = 0$, the ARiS produces a consistent estimate and does not require a search over the solution path. For medium and large values of n we can see that it significantly outperforms the lasso. Results for the lasso and adalasso agree with those of Zou (2006).

We now compare prediction accuracy and model selection consistency using the following three models which are drawn from Tibshirani (1996).

Model 1: In this example, we let $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ with iid normal predictors $\mathbf{x}_i$ ($i = 1, \dots, n$). The pairwise correlation between the predictors $\mathbf{x}_j$ and $\mathbf{x}_k$ are adjusted to be $(.5)^{|j-k|}$.

Model 2: We use the same setup as model 1 with $\beta_j = 0.85$ for all j .

Model 3: We use the same setup as model 1 with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)'$.

We test models 1,2, and 3 for two different sample sizes ($n = 20, 100$) and two noise levels ($\sigma = 3, 6$). This experiment is conducted 100 times under each setting. In Tables 4.2, 4.3 and 4.4 we report the median prediction error (MSE) on a test set of 10,000 observation for each of the 100 cases. The values in the parentheses give the bootstrap standard error of the median MSE values obtained. C, I and CM respectively stand for the number of correctly chosen predictors, number of incorrectly

Table 4.2: Results for model 1.

			MSE (Sd)	C	I	CM
$\sigma = 3$	$n = 20$	$ARiS(\eta = 0)$	14.3414 (0.4198)	2.23	0.89	0.15
		$ARiS - eB_{Lap}$	16.3220 (0.3434)	1.42	0.10	0.04
		$ARiS - eB_{k=10}$	14.1294 (0.5490)	2.05	0.53	0.17
		$Lasso$	13.8329 (0.4078)	2.69	1.79	0.08
		$Lasso(cml)$	13.7349 (0.4959)	2.37	1.07	0.09
		$AdaLasso$	15.0272 (0.4686)	2.26	1.2	0.13
		$ElasticNet(\lambda_2 = 1)$	13.7353 (0.3343)	2.73	1.89	0.06
		$nn - Garrote$	14.0934 (0.4435)	2.58	2.48	0.02
		$Ridge$	13.7727 (0.4166)	3	5	0
		Ols	15.4568 (0.5224)	3	5	0
	$n = 100$	$ARiS(\eta = 0)$	9.3409 (0.0660)	2.97	0.29	0.75
		$ARIS - eB_{Lap}$	9.4427 (0.0724)	2.96	0.5	0.64
		$ARiS - eB_{k=3}$	9.3887 (0.0432)	2.98	0.42	0.67
		$Lasso$	9.6631 (0.0631)	3	1.96	0.13
		$Lasso(cml)$	9.6605 (0.0555)	2.99	0.9	0.41
		$AdaLasso$	9.7004 (0.0939)	2.85	1.08	0.31
		$ElasticNet(\lambda_2 = 0.1)$	9.5607 (0.0671)	3	2.02	0.14
		$nn - Garrote$	9.4919 (0.0901)	3	2.14	0.02
		$Ridge$	9.8615 (0.0755)	3	5	0
		Ols	9.7112 (0.0596)	3	5	0
$\sigma = 6$	$n = 20$	$ARiS(\eta = 0)$	53.3474 (1.5041)	1.37	1.07	0.05
		$ARiS - eB_{Lap}$	52.3918 (1.2303)	0.89	0.48	0.01
		$ARiS - eB_{k=1000}$	50.3332 (1.3107)	0.87	0.48	0.01
		$Lasso$	48.9462 (1.1343)	1.73	1.43	0.03
		$Lasso(cml)$	48.1166 (1.1363)	1.57	1.05	0.04
		$AdaLasso$	52.9084 (0.8642)	1.42	1.26	0.06
		$ElasticNet(\lambda_2 = 100)$	46.5830 (0.9027)	2.01	1.50	0.02
		$nn - Garrote$	58.5472 (1.5686)	2.44	3.68	0
		$Ridge$	48.4516 (0.9443)	3	5	0
		Ols	60.1073 (1.5030)	3	5	0
	$n = 100$	$ARiS(\eta = 0)$	38.6604 (0.3078)	2.45	0.35	0.39
		$ARIS - eB_{Lap}$	40.2355 (0.6259)	1.92	0.24	0.23
		$ARiS - eB_{k=3}$	38.5913 (0.2623)	2.53	0.43	0.41
		$Lasso$	38.8449 (0.1967)	2.93	2.07	0.15
		$Lasso(cml)$	38.6644 (0.3805)	2.67	0.79	0.33
		$AdaLasso$	39.1548 (0.2348)	2.61	1.25	0.19
		$ElasticNet(\lambda_2 = 100)$	38.4698 (0.1214)	2.96	1.98	0.10
		$nn - Garrote$	39.5005 (0.2327)	2.92	3.77	0
		$Ridge$	38.9526 (0.1671)	3	5	0
		Ols	39.2768 (0.2282)	3	5	0

Table 4.3: Results for model 2.

			MSE (Sd)	C	I	CM
$\sigma = 3$	$n = 20$	$ARiS(\eta = 0)$	15.3053 (0.4332)	3.4	0	0
		$ARiS - eB_{Lap}$	19.0261 (0.1610)	1.60	0	0
		$ARiS - eB_{k=3}$	15.2739 (0.3484)	3.24	0	0
		$Lasso$	14.0350 (0.3963)	5.21	0	0.08
		$Lasso(cml)$	14.7502 (0.5382)	3.61	0	0
		$AdaLasso$	16.4863 (0.5305)	3.66	0	0.01
		$ElasticNet(\lambda_2 = 1)$	13.0765 (0.2780)	6.40	0	0.29
		$nn - Garrote$	14.5337 (0.4887)	5.09	0	0
		$Ridge$	11.7124 (0.2210)	8	0	1
		Ols	14.2135 (0.3473)	8	0	1
	$n = 100$	$ARiS(\eta = 0)$	10.6279 (0.0781)	5.73	0	0.01
		$ARiS - eB_{Lap}$	10.6008 (0.1936)	4.96	0	0.28
		$ARiS - eB_{k=3}$	10.5673 (0.0920)	5.86	0	0.02
		$Lasso$	9.7986 (0.0428)	7.83	0	0.84
		$Lasso(cml)$	10.3712 (0.0664)	7.24	0	0.44
		$AdaLasso$	10.0627 (0.0889)	7.18	0	0.53
		$ElasticNet(\lambda_2 = 0.1)$	9.7212 (0.0624)	7.87	0	0.87
		$nn - Garrote$	10.0068 (0.0635)	7.38	0	0.49
		$Ridge$	9.6199 (0.0649)	8	0	1
		Ols	9.7262 (0.0596)	8	0	1
$\sigma = 6$	$n = 20$	$ARiS(\eta = 0)$	49.7997 (0.7488)	2.13	0	0
		$ARiS - eB_{Lap}$	49.3095 (0.5579)	1.31	0	0
		$ARiS - eB_{k=100}$	47.9480 (0.7287)	1.38	0	0
		$Lasso$	47.3209 (0.7402)	2.7	0	0.01
		$Lasso(cml)$	46.5628 (0.5432)	1.95	0	0
		$AdaLasso$	48.7509 (0.5405)	2.39	0	0
		$ElasticNet(\lambda_2 = 1000)$	46.7312 (0.7713)	3.19	0	0
		$nn - Garrote$	57.1654 (2.3273)	5.71	0	0.06
		$Ridge$	45.6485 (0.8320)	8	0	1
		Ols	60.2328 (2.0051)	8	0	1
	$n = 100$	$ARiS(\eta = 0)$	40.8476 (0.1875)	3.22	0	0
		$ARiS - eB_{Lap}$	45.3506 (0.2318)	1.5	0	0
		$ARiS - eB_{k=3}$	40.8015 (0.1975)	3.46	0	0
		$Lasso$	38.8809 (0.2259)	6.45	0	0.18
		$Lasso(cml)$	40.8431 (0.3779)	3.74	0	0
		$AdaLasso$	40.4044 (0.2428)	4.41	0	0.02
		$ElasticNet(\lambda_2 = 0.01)$	38.6808 (0.1883)	6.4	0	0.17
		$nn - Garrote$	39.0697 (0.1628)	6.79	0	0.29
		$Ridge$	38.4051 (0.1647)	8	0	1
		Ols	38.6823 (0.1705)	8	0	1

Table 4.4: Results for model 3.

			MSE (Sd)	C	I	CM
$\sigma = 3$	$n = 20$	$ARiS(\eta = 0)$	11.2573 (0.3805)	1	1.09	0.41
		$ARiS - eB_{Lap}$	9.8811 (0.1401)	1	0.10	0.92
		$ARiS - eB_{k=1000}$	10.0642 (0.1829)	1	0.07	0.95
		$Lasso$	11.5735 (0.3479)	1	1.62	0.31
		$Lasso(cml)$	10.6312 (0.3642)	1	1.59	0.34
		$AdaLasso$	11.5925 (0.4178)	1	1.29	0.43
		$ElasticNet(\lambda_2 = 0)$	11.5735 (0.3479)	1	1.62	0.31
		$nn - Garrote$	12.8139 (0.4729)	1	3.43	0.01
		$Ridge$	15.1850 (0.4721)	1	7	0
		Ols	15.3540 (0.3310)	1	7	0
	$n = 100$	$ARiS(\eta = 0)$	9.2237 (0.0404)	1	0.36	0.71
		$ARIS - eB_{Lap}$	9.1452 (0.0172)	1	0.04	0.97
		$ARiS - eB_{k=1000}$	9.1531 (0.0172)	1	0.05	0.96
		$Lasso$	9.3343 (0.0503)	1	1.99	0.21
		$Lasso(cml)$	9.2238 (0.0437)	1	1.28	0.40
		$AdaLasso$	9.3025 (0.0578)	1	1.27	0.31
		$ElasticNet(\lambda_2 = 0)$	9.3343 (0.0503)	1	1.99	0.21
		$nn - Garrote$	9.5324 (0.0460)	1	3.35	0.01
		$Ridge$	9.8868 (0.0610)	1	7	0
		Ols	9.7112 (0.0596)	1	7	0
$\sigma = 6$	$n = 20$	$ARiS(\eta = 0)$	45.6378 (0.7751)	0.89	1.26	0.28
		$ARiS - eB_{Lap}$	40.3920 (0.7669)	0.87	0.28	0.76
		$ARiS - eB_{k=1000}$	41.4490 (0.7836)	0.86	0.23	0.80
		$Lasso$	45.0416 (1.0000)	0.96	1.72	0.23
		$Lasso(cml)$	42.2038 (1.0688)	0.96	1.61	0.30
		$AdaLasso$	45.1020 (1.5670)	0.9	1.64	0.30
		$ElasticNet(\lambda_2 = 0)$	45.0416 (1.0000)	0.96	1.72	0.23
		$nn - Garrote$	55.4879 (2.8182)	0.98	5.15	0
		$Ridge$	53.4027 (1.1883)	1	7	0
		Ols	60.8385 (2.5876)	1	7	0
	$n = 100$	$ARiS(\eta = 0)$	36.9999 (0.1633)	1	0.35	0.72
		$ARIS - eB_{Lap}$	36.5987 (0.1946)	1	0.06	0.97
		$ARiS - eB_{k=1000}$	36.8319 (0.1794)	1	0.04	0.96
		$Lasso$	37.7736 (0.1582)	1	1.99	0.24
		$Lasso(cml)$	37.0283 (0.1826)	1	1.1	0.47
		$AdaLasso$	37.7555 (0.2429)	1	1.24	0.50
		$ElasticNet(\lambda_2 = 0.01)$	37.5619 (0.1634)	1	1.75	0.31
		$nn - Garrote$	38.5477 (0.2468)	1	5.07	0
		$Ridge$	38.7256 (0.2456)	1	7	0
		Ols	38.8450 (0.1913)	1	7	0

chosen predictors and the proportion of cases (out of 100) where the correct model was found by the method.

The bootstrap standard error was calculated by generating 500 bootstrap samples from each of the 100 cases, finding the median MSE for each case, and then calculating the standard error of these medians. Lasso, adalasso (adaptive lasso), elastic net, nonnegative garrote, ridge and ordinary least squares estimates are computed along with the ARiS estimate. ARiS hyper-parameter η is determined both by the Laplace approximation and the numerical integration to the marginal likelihood. In each example the numerical integration step of ARiS-eB is carried out for values of $k = 3, 10, 100, 1000$ and only the best result is reported. This is a rather arbitrary choice and will depend upon the number of samples drawn. Model 3 is the only example where the best value chosen consistently is $k = 1000$. For the ridge estimator, the ridge parameter is determined by a GCV (generalized cross-validation) type statistic, while for all the others we use 10-fold cross-validation for the choice of the tuning parameters. We also consider the lasso where the tuning parameter is chosen by the method of Yuan and Lin (2005). We also report the results for the particular case of $\eta = 0$.

Model 3 demonstrates the most striking feature of the ARiS algorithm, the ability to identify the correct model under sparse setups. When utilized along with the empirical Bayes step, it is able to identify the correct model in a very large proportion of cases with very low prediction error. This is especially surprising for the cases where $n = 20$ ($\sigma = 3, 6$). In the case where $n = 100$ and $\eta = 0$ the algorithm still outperforms all other methods in terms of correct model choice and MSE.

Among all the variants of lasso (lasso, adaptive lasso, elastic net, lasso(CML)), lasso(CML) is optimal in terms of prediction accuracy. In the case of $n = 100$,

its prediction error is almost identical to that of $\text{ARiS}(\eta = 0)$ but correct model identification is significantly weaker.

One interesting observation here is that a cross-validation approach may not accurately choose the tuning parameter for the lasso-variants. For example, as we moved from $n = 20$ to $n = 100$, the proportion of cases where the correct model was chosen decreased for all the lasso-variants except $\text{lasso}(\text{CML})$ where the tuning parameter is chosen via an empirical Bayes step similar to our approach. The nonnegative garrote estimator performs quite poorly in this situation along with the ridge and ols estimators. Results indicate that the ARiS provides superior performance for model 3.

In the case of model 1, for $n = 100$, $\sigma = 3$ ARiS performs best in terms of prediction accuracy and strongly outperforms other algorithms in terms of model selection accuracy. $\text{ARiS}(\eta = 0)$ outperformed other versions which required a search over the solution path. Both the Laplace approximation and the numerical integration fail to detect this value of η . For the case of $n = 20$, ARiS performs within a standard error of all the other estimators in terms of prediction accuracy, yet does better in terms of model selection accuracy. The ridge estimator does almost as well as the lasso-variants in terms of prediction accuracy. Similar results follow for the case $n = 100$, $\sigma = 6$. However, elastic net seems to have slightly lower prediction error. The case $n = 20$, $\sigma = 6$ shows fairly weak results across all estimators.

Model 2 demonstrates the biggest weakness of the ARiS and several other estimators. When there are many small effects present in the underlying model, these estimators do not perform well since they favor parsimony. For all cases the clear winner is the ridge estimator.

Here we also conduct an experiment to investigate the model selection consistency of our special case when $\eta = 0$ for the joint maximization or the equivalent $\eta = -1$ case

Table 4.5: Model selection accuracy of ARiS.

		$\sigma = 1$	$\sigma = 3$	$\sigma = 6$
Model 1	$n = 20$	0.567	0.178	0.014
	$n = 100$	0.777	0.734	0.388
	$n = 500$	0.800	0.785	0.792
Model 2	$n = 20$	0.184	0	0
	$n = 100$	0.999	0.024	0
	$n = 500$	1	0.990	0.080
Model 3	$n = 20$	0.579	0.583	0.509
	$n = 100$	0.746	0.737	0.714
	$n = 500$	0.750	0.784	0.740
Model 4	$n = 20$	0.522	0.116	0.039
	$n = 100$	0.725	0.549	0.176
	$n = 500$	0.768	0.759	0.636

for the marginal maximization. Recall that these two values respectively correspond to a flat and a scale-invariant prior on v_j^{-1} in the limit as $\mu \rightarrow 0$. We have previously experimented with three models: (i) one with moderate number of moderate effects, (ii) one with many small effect, and (iii) one with one large effect. Here we add a fourth model which has one large effect and one small effect.

Model 4: We use the same setup as model 1 with $\beta = (5, 0, 0, 0, 0.85, 0, 0, 0)'$.

One thousand data sets for each combination of model, sample size ($n = 20, 100, 500$) and noise variance ($\sigma = 1, 3, 6$). Table 4.5 gives the proportion of times the correct model was identified by ARiS under each circumstance. Figure 4.1 also contrasts the $\mathcal{MSE}$ s resulting from the OLS and ARiS estimators for these four different models. Following Tibshirani (1996) the mean-squared error of an estimate $\mathbf{X}\hat{\beta}$ is computed as $\mathcal{MSE} = \mathbb{E} \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta \right)^2$. The expectation of this quantity with respect to the joint distribution of $\mathbf{X}$ and $\mathbf{y}$ with fixed $\hat{\beta}$ is $\left(\hat{\beta} - \beta \right)' \Sigma \left(\hat{\beta} - \beta \right)$ where Σ is the population covariance matrix of $\mathbf{X}$. In fact, assuming $\hat{\beta}$ is a fixed quantity, $\mathcal{MSE}$ only consists of a squared bias term.

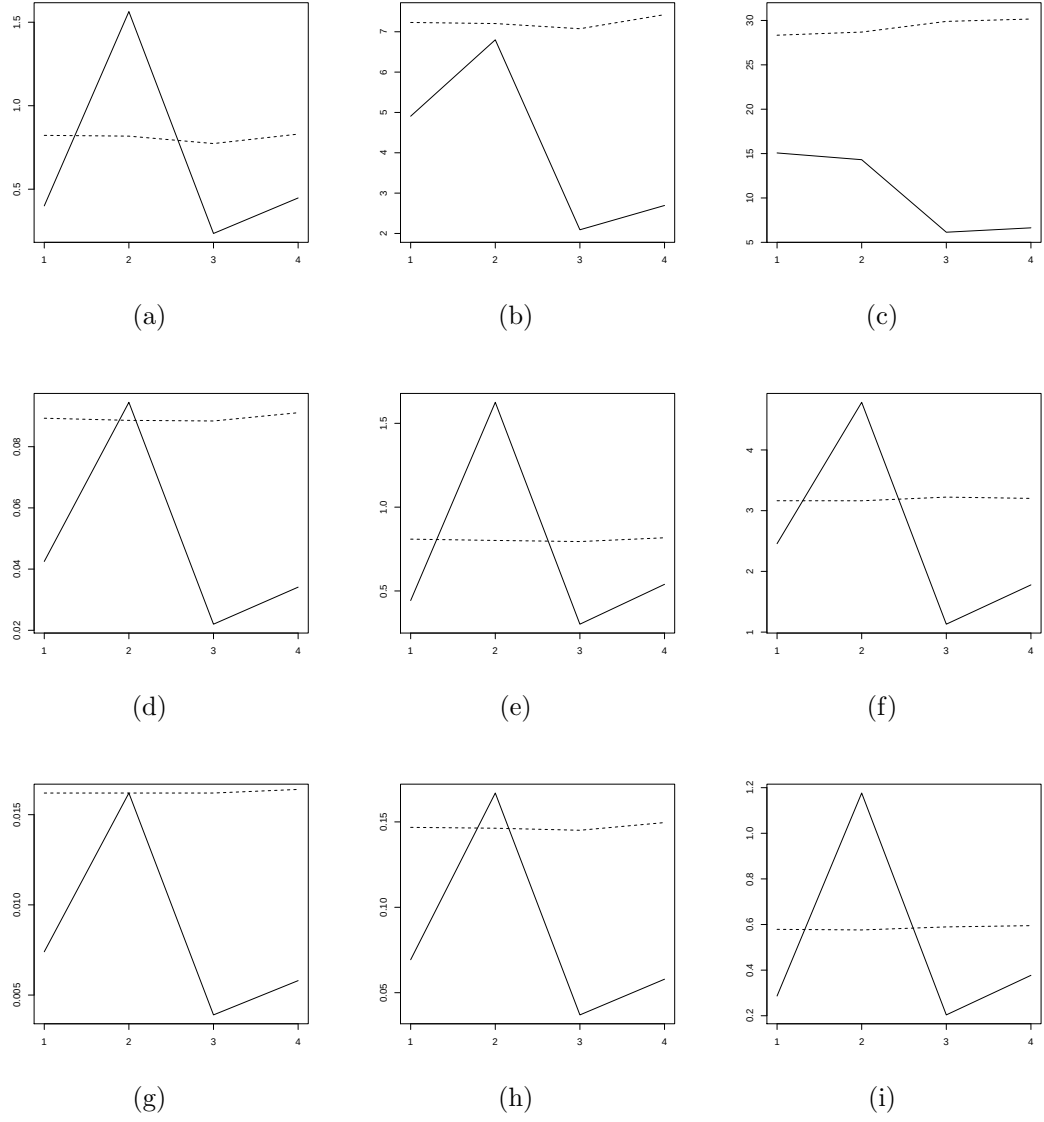


Figure 4.1: $\mathcal{MSE}$ values for ARiS($\eta = 0$) (solid lines) vs. OLS (dashed lines) estimator for model 1,2,3,4. Sample size increases in the direction $\downarrow$ as 20, 100 and 500 while residual standard deviation increases in the direction $\rightarrow$ as 1, 3 and 6.

Student- t Setup

Having contrasted the ARiS estimator with a number of others under the normal setup, here we contrast it with ARiS- t which is obtained via modeling the data with a scale mixtures of normals as explained in Section 3.2. In this section we utilize a 10-fold cross-validation procedure for both methods to choose from a course grid of $\eta(0, 0.001, 0.01, 0.1, 1)$ and $v = (.5, 2, 5, 10, 30, 100, 1000)$ values. The underlying models are generated by the same protocols that were used before and are again named Model 1,2 and 3. This time we contaminate 10% of the data by extra noise introduced by $\mathcal{N}(0, \sigma_c^2)$ for two σ_c values, 10 and 20. As in the earlier examples we use two different levels of sample size $n = 20, 100$, and add an additional level of noise variance, $\sigma = 1, 3, 6$. Results are presented in tables 4.6 and 4.7.

In the case of $\sigma_c = 10$, the only significant improvement is observed when $\sigma = 1$. This is due to the absorption of the contamination by the underlying noise for the cases of $\sigma = 3$ and $\sigma = 6$. In the case of $\sigma_c = 20$, the improvement carries onto the $\sigma = 3$ case as well. As the magnitude of the contamination increases ARiS- t becomes more advantageous.

4.2 Real Data Sets

4.2.1 Diabetes Data

Here we use the diabetes data originally analyzed in Efron et al. (2004). There are ten baseline variables, age, sex, body mass index (bmi), average blood pressure (map) and six blood serum measurements (tc,ldl,hdl,tch,ltg,glu) as well as a response variable, a quantitative measure of disease progression one year after baseline; $n = 442$ cases

Table 4.6: Results for $\sigma_c = 10$.

				MSE (Sd)	C	I	CM
Model 1	$\sigma = 1$	$n = 20$	<i>ARiS</i>	5.2732 (1.0725)	2.47	1.35	0.28
			<i>ARiS</i> - t	2.8307 (0.1584)	2.82	1.73	0.37
		$n = 100$	<i>ARiS</i>	1.7719 (0.0798)	2.99	0.94	0.62
			<i>ARiS</i> - t	1.2441 (0.0299)	3	1.31	0.62
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	17.8392 (0.6190)	2.19	1.52	0.12
			<i>ARiS</i> - t	16.9201 (0.9723)	2.26	1.67	0.11
		$n = 100$	<i>ARiS</i>	10.5035 (0.1598)	2.87	1.01	0.52
			<i>ARiS</i> - t	9.9816 (0.1257)	2.94	0.96	0.60
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	57.8068 (1.6903)	1.41	1.34	0
			<i>ARiS</i> - t	58.9853 (2.2114)	1.46	1.46	0
		$n = 100$	<i>ARiS</i>	39.9456 (0.3487)	2.61	1.59	0.21
			<i>ARiS</i> - t	39.7621 (0.3473)	2.67	1.74	0.19
Model 2	$\sigma = 1$	$n = 20$	<i>ARiS</i>	7.8928 (0.6474)	4.88	0	0.21
			<i>ARiS</i> - t	3.7783 (0.1903)	6.42	0	0.38
		$n = 100$	<i>ARiS</i>	2.1779 (0.1261)	7.68	0	0.80
			<i>ARiS</i> - t	1.2621 (0.0267)	7.92	0	0.95
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	20.1101 (0.5071)	3.26	0	0.08
			<i>ARiS</i> - t	19.4910 (0.7102)	3.66	0	0.16
		$n = 100$	<i>ARiS</i>	11.3175 (0.2062)	6.95	0	0.61
			<i>ARiS</i> - t	10.5497 (0.0744)	7.31	0	0.73
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	55.9146 (1.4794)	2.59	0	0.11
			<i>ARiS</i> - t	55.4070 (1.7295)	2.68	0	0.14
		$n = 100$	<i>ARiS</i>	41.1967 (0.1974)	5.61	0	0.31
			<i>ARiS</i> - t	41.0093 (0.3300)	5.24	0	0.27
Model 3	$\sigma = 1$	$n = 20$	<i>ARiS</i>	2.2117 (0.2175)	1	1.15	0.58
			<i>ARiS</i> - t	1.5880 (0.1031)	1	1.1	0.69
		$n = 100$	<i>ARiS</i>	1.5335 (0.0642)	1	0.56	0.73
			<i>ARiS</i> - t	1.2724 (0.0671)	1	1.18	0.71
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	12.2746 (0.7873)	0.99	1.37	0.51
			<i>ARiS</i> - t	11.5989 (0.5967)	0.99	1.57	0.55
		$n = 100$	<i>ARiS</i>	9.6844 (0.1055)	1	0.66	0.77
			<i>ARiS</i> - t	9.5581 (0.1156)	1	0.7	0.77
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	52.1999 (4.2667)	0.78	1.18	0.37
			<i>ARiS</i> - t	55.6111 (3.8153)	0.82	1.17	0.36
		$n = 100$	<i>ARiS</i>	37.3368 (0.2655)	1	1.14	0.66
			<i>ARiS</i> - t	37.2309 (0.1836)	1	0.98	0.67

Table 4.7: Results for $\sigma_c = 20$.

				MSE (Sd)	C	I	CM
Model 1	$\sigma = 1$	$n = 20$	<i>ARiS</i>	15.6786 (2.5786)	1.91	1.51	0.12
			<i>ARiS</i> - t	5.3514 (0.9097)	2.55	1.58	0.23
		$n = 100$	<i>ARiS</i>	4.2165 (0.4098)	2.85	1.18	0.51
			<i>ARiS</i> - t	1.6617 (0.1412)	2.94	1.53	0.50
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	27.0979 (2.1684)	1.79	1.57	0.02
			<i>ARiS</i> - t	17.9485 (0.9062)	2.03	1.42	0.07
		$n = 100$	<i>ARiS</i>	13.0048 (0.3020)	2.67	1.24	0.32
			<i>ARiS</i> - t	10.5156 (0.1742)	2.84	1.22	0.43
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	65.8497 (3.5447)	1.13	1.17	0
			<i>ARiS</i> - t	60.6089 (2.5987)	1.4	1.45	0
		$n = 100$	<i>ARiS</i>	43.4620 (0.6791)	2.47	2.22	0.04
			<i>ARiS</i> - t	41.2204 (0.4121)	2.51	2.1	0.07
Model 2	$\sigma = 1$	$n = 20$	<i>ARiS</i>	15.2099 (1.3416)	3.4	0	0.16
			<i>ARiS</i> - t	7.7472 (1.0183)	4.79	0	0.24
		$n = 100$	<i>ARiS</i>	5.4633 (0.3394)	6.78	0	0.50
			<i>ARiS</i> - t	1.6485 (0.0988)	7.64	0	0.87
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	25.3197 (1.0193)	2.73	0	0.11
			<i>ARiS</i> - t	21.1522 (0.8861)	3.23	0	0.09
		$n = 100$	<i>ARiS</i>	14.6959 (0.3444)	6.05	0	0.41
			<i>ARiS</i> - t	10.9480 (0.1734)	7.09	0	0.63
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	56.8565 (1.7079)	2.12	0	0.09
			<i>ARiS</i> - t	54.2452 (1.4603)	2.06	0	0.03
		$n = 100$	<i>ARiS</i>	44.4648 (0.6366)	4.28	0	0.15
			<i>ARiS</i> - t	42.1942 (0.3727)	4.99	0	0.28
Model 3	$\sigma = 1$	$n = 20$	<i>ARiS</i>	6.0621 (1.3163)	0.93	1.21	0.54
			<i>ARiS</i> - t	3.0622 (0.4213)	1	1.43	0.63
		$n = 100$	<i>ARiS</i>	2.9015 (0.2510)	1	0.51	0.72
			<i>ARiS</i> - t	2.0073 (0.1474)	1	1.5	0.65
	$\sigma = 3$	$n = 20$	<i>ARiS</i>	17.0510 (1.7611)	0.94	1.41	0.51
			<i>ARiS</i> - t	13.2317 (0.7918)	0.97	1.5	0.59
		$n = 100$	<i>ARiS</i>	10.8986 (0.2119)	1	0.62	0.70
			<i>ARiS</i> - t	10.2499 (0.1591)	1	1.05	0.73
	$\sigma = 6$	$n = 20$	<i>ARiS</i>	61.3606 (2.7082)	0.74	1.48	0.39
			<i>ARiS</i> - t	59.7560 (2.0970)	0.8	1.79	0.34
		$n = 100$	<i>ARiS</i>	38.6686 (0.3668)	1	0.85	0.65
			<i>ARiS</i> - t	38.0756 (0.4615)	1	1.28	0.61

Table 4.8: Models for ARiS and lasso solution paths.

ARiS	lasso
Full model	Full model
sex,bmi,map,tc,ldl,tch,ltg,glu	age,sex,bmi,map,tc,ldl,tch,ltg,glu
sex,bmi,map,tc,ldl,tch,ltg	Full model
sex,bmi,map,tc,tch,ltg	sex,bmi,map,tc,ldl,tch,ltg,glu
bmi,map,tc,ltg	sex,bmi,map,tc,tch,ltg,glu
bmi,map,ltg	sex,bmi,map,tc,ltg,glu
bmi,ltg	sex,bmi,map,ltg,glu
bmi	sex,bmi,map,ltg
Null model	bmi,map,ltg
-	bmi,ltg
-	bmi
-	Null model

were observed. Predictors are standardized to have zero mean and unit length, and the response is centered. This data can be obtained in R under the package `lars`.

We fit a linear model to the response using all the previously considered methods for the simulated data examples under the normal case. Results presented in table 4.9 are obtained by randomly assigning 42 observations to a testing set and using the remaining as a training set. This is repeated 100 times and median values for prediction error and coefficient of determination are presented. Also an average number of variables kept in the model is given for these 100 cases along with bootstrap standard errors (in parentheses, for 500 bootstrap samples) for MSE and R^2 .

There are no significant differences among the prediction errors achieved by the different methods considering the standard errors. However, ARiS manages to fit a sparser model than the competing estimators while keeping the prediction error within a comparable region. Figure 4.2 shows the solution path of the algorithm with respect to η . This path is obtained by increasing η values in increments of 0.1. This also yields a list of potential subsets to be considered. Table 4.8 contrasts the potential models along the solution path of ARiS with those of lasso.

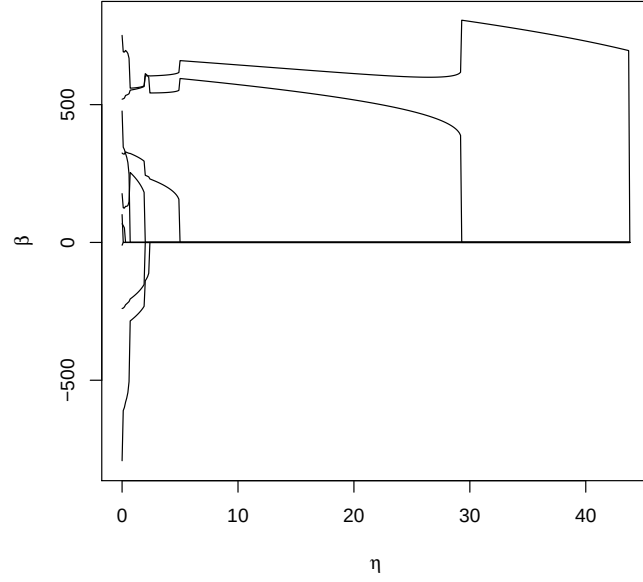


Figure 4.2: Solution path for the diabetes data with respect to η .

Table 4.9: Results for Diabetes data.

	MSE (Sd)	R^2 (Sd)	No. of Var.
$ARiS(\eta = 0)$	3020.9308 (42.9675)	0.4853 (0.0168)	6
$ARiS - eB_{Lap}$	3062.5430 (59.2229)	0.4761 (0.0198)	6.7
$ARiS - eB_{k=3}$	3028.1413 (37.5719)	0.4804 (0.0155)	6
<i>Lasso</i>	3003.2556 (49.2946)	0.4889 (0.0150)	9
<i>Lasso(cml)</i>	2978.8661 (46.0926)	0.4956 (0.0162)	7.8
<i>AdaLasso</i>	3048.0327 (52.5579)	0.4708 (0.0153)	7
<i>ElasticNet</i> ($\lambda_2 = 0.01$)	2990.9332 (47.5049)	0.4909 (0.0160)	8.6
<i>nn - Garrote</i>	3010.9086 (55.6293)	0.4842 (0.0159)	10
<i>Ridge</i>	3037.0677 (55.8977)	0.4887 (0.0166)	10
<i>Ols</i>	3010.8947 (51.6668)	0.4842 (0.0160)	10

4.2.2 Prostate Cancer Data

The prostate cancer data come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen and a number of clinical measures, in men who were about to receive a radical prostatectomy. The factors were $\log(\text{cancer volume})$ (lcavol), $\log(\text{prostate weight})$ (lweight), age, $\log(\text{benign prostatic hyperplasia amount})$ (lbph), seminal vesicle invasion (svi), $\log(\text{capsular penetration})$ (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (ppg45). A linear model is fit to $\log(\text{prostate specific antigen})$ (lpsa) having scaled the predictors to have unit norm and centered the response. Results are presented in table 4.11.

Results presented in table 4.11 are obtained by randomly assigning 7 observations to a testing set and using the remaining as a training set. The same procedure is followed as for diabetes data set.

Again in this example, the prediction errors are not significantly different from each other. However ARiS and nonnegative garrote are able to achieve the same level of prediction error with much sparser models than what other methods offer. Figure 4.3 shows the solution path of the algorithm with respect to η . Table 4.10 again contrasts the potential models along the solution path of ARiS with those of lasso.

Table 4.10: Models for ARiS and lasso solution paths.

ARiS	lasso
Full model	Full model
lcavol,lweight,age,lbph,svi,lcp,ppg45	lcavol,lweight,age,lbph,svi,gleason,ppg45
lcavol,lweight,age,lbph,svi,ppg45	lcavol,lweight,age,lbph,svi,ppg45
lcavol,lweight,age,lbph,svi	lcavol,lweight,lbph,svi,ppg45
lcavol,lweight,lbph,svi	lcavol,lweight,lbph,svi
lcavol,lweight,svi	lcavol,lweight,svi
lcavol	lcavol,svi
Null Model	lcavol
-	Null model

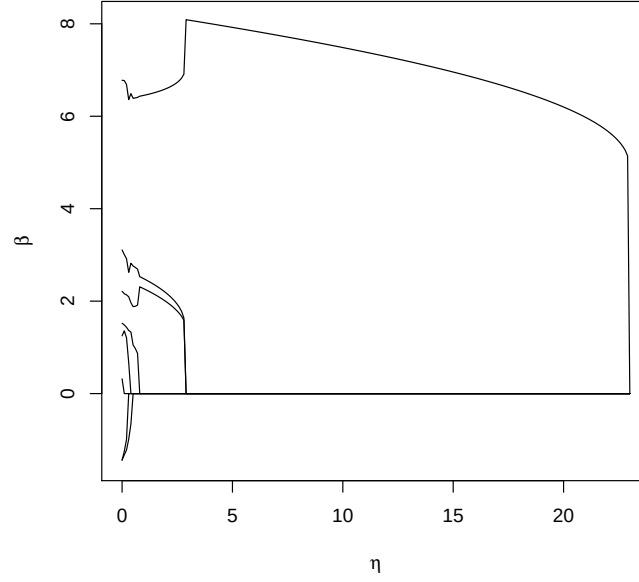


Figure 4.3: Solution path for the prostate cancer data with respect to η .

Table 4.11: Results for Prostate Cancer data.

	MSE (Sd)	R^2 (Sd)	No. of Var.
$ARiS(\eta = 0)$	0.4956 (0.0423)	0.4713 (0.0422)	3.1
$ARiS - eB_{Lap}$	0.5284 (0.0404)	0.4250 (0.0475)	1.1
$ARiS - eB_{k=3}$	0.4310 (0.0508)	0.4225 (0.0409)	3.2
<i>Lasso</i>	0.4796 (0.0497)	0.4897 (0.0413)	6.1
<i>Lasso(cml)</i>	0.4745 (0.0436)	0.4938 (0.0388)	6.1
<i>AdaLasso</i>	0.5159 (0.0228)	0.5654 (0.0289)	5.1
<i>ElasticNet</i> ($\lambda_2 = 0.1$)	0.4649 (0.0438)	0.5051 (0.0323)	6.3
<i>nn - Garrote</i>	0.5274 (0.0358)	0.4074 (0.0449)	2.6
<i>Ridge</i>	0.4567 (0.0490)	0.5161 (0.0362)	8
<i>Ols</i>	0.4410 (0.0430)	0.4898 (0.0504)	8

Chapter 5

Conclusion

We have introduced a new simultaneous shrinkage estimation and model selection procedure, ARiS, which makes use of a hierarchical model and enforces parsimony. The method combines an efficient optimization procedure which is tailored to the fully conditional posterior densities with various techniques to derive and maximize the marginal likelihood. This development, although radically different in detail, is similar in spirit to modern implementation of the lasso which has been described as a Bayesian procedure which combines a normal likelihood with a Laplace prior on the regression coefficients; see Tibshirani (1996).

Considering the simulation results of Section 4, we note two key features of the ARiS: (i) its superior prediction and model selection accuracy when the underlying model is sparse, and (ii) the significant improvement in performance accompanying an increase in the sample size indicating asymptotic consistency.

Computationally, for a specific η value, ARiS requires one matrix inversion at each iteration. This point is obvious from the description of the method as a series of ridge regressions. Thus the computational cost for each iteration of ARiS is at most $O(p^3)$. In practice, because variables are eliminated throughout the procedure, the cost often

decreases dramatically with each iteration. Our experiments indicate fast convergence of this procedure across sample sizes. Lasso methods offer a computational advantage due to the *lars* algorithm (Efron et al., 2004) which can compute the entire solution path of the lasso with the cost of a single OLS estimator. However, our experimental results indicate that these methods are often inferior in terms of model selection and prediction accuracy. The special case $\text{ARiS}(\eta = 0)^*$ also provides very promising results asymptotically. This is especially exciting since this procedure requires no search over η , thus is computationally much more efficient.

Large scale experiments have shown that the procedure remains computationally feasible in situations where the number of predictor variables is very large. Hence the proposed method offers the most advantage in problems where one is attempting to select a small or moderate number of variables from a large initial group, a common situation in many modern statistical and data mining applications. An open issue is the empirical Bayes step via numerical integration.

Due to the large scale simulations throughout our experiments we have only drawn 1000 samples for the integration of $\mathbf{v}^{-1}$. Obviously in practice a much larger set of samples could be drawn at little additional cost particularly for sparse models. The process should become more stable as we draw larger samples. In such a case, the choice of k may just be fixed at a larger value, i.e. $k = 1000$. We conducted an experiment for model 1 in Section 4.1, with $n = 100$ and $\sigma = 3$. We changed the number of samples drawn for the integration of v_j^{-1} from 1000 to 100000. The behavior of log-marginal likelihood can be seen in figure 5.1 for varying k and samples drawn. Panels (a)-(c) demonstrate the log-marginal likelihood profiles for 1000, 10000 and 100000 samples drawn for numerical integration. Observe that variation for $k = 1000$ case becomes progressively smaller, yet is not very smooth. The maximum appears

*or $\text{ARiS}(\eta = -1)$ for the solution obtained via the EM procedure, i.e. the marginal mode for β

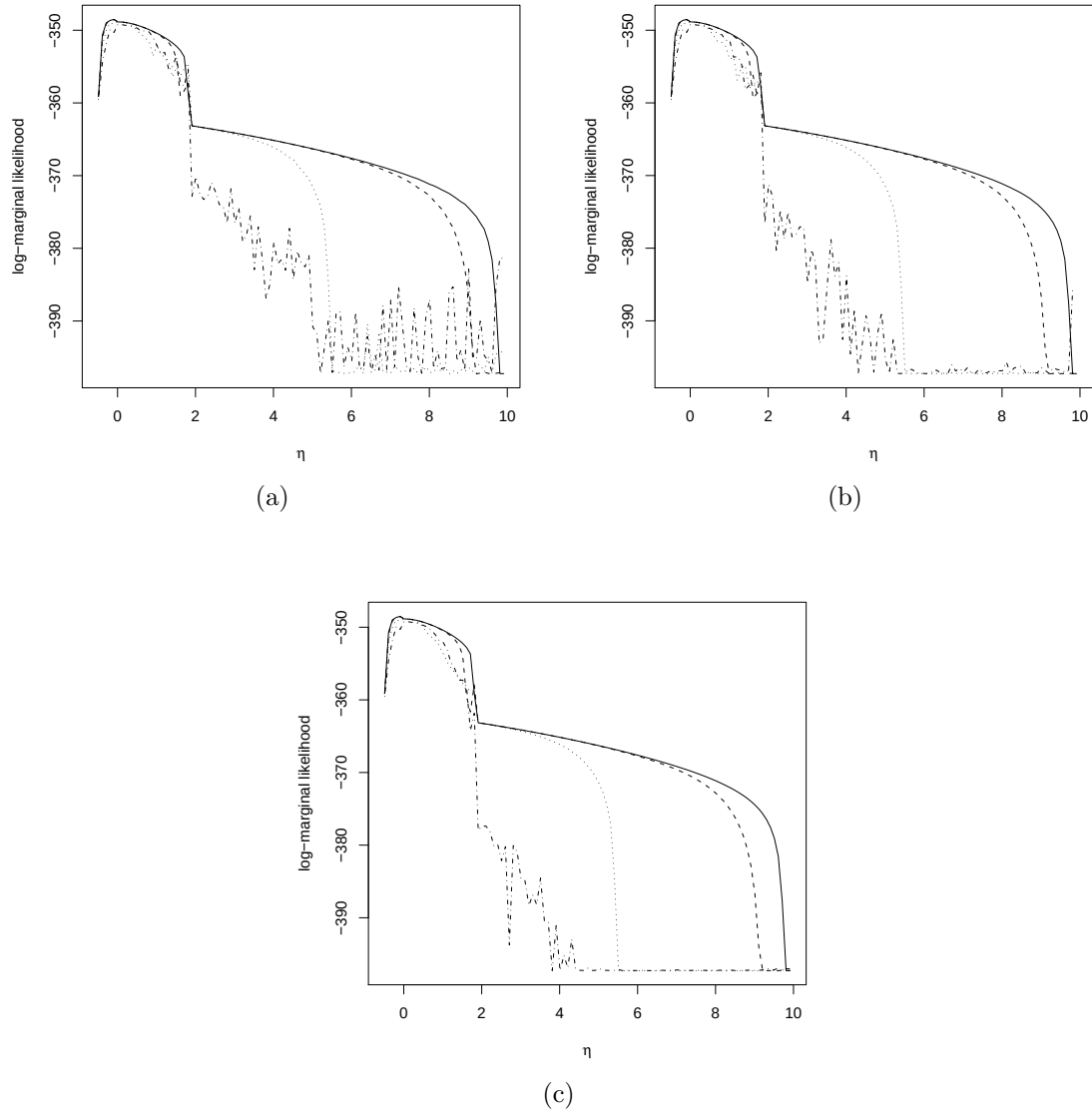


Figure 5.1: Log-marginal likelihood plotted against η values. Solid, dashed, dotted and dot-dash lines stand respectively for $k = 3, 10, 100, 1000$.

to occur around the same η value for $k = 3, 10, 100$ while the maximum for $k = 1000$ occurs at a slightly larger η value. The larger number of samples drawn, the more expensive and infeasible the procedure becomes computationally. In our examples, we contrasted different k values to understand which gave better results under different circumstances. $k = 1000$ gives the best results in situations where our method is most suitable for, i.e. sparse setups. This is sensible since larger k values appear to choose larger η as optimal.

Here are also two immediate research ideas that could follow this dissertation:

- i. I would like to explore further properties of the ARiS, such as whether it is an oracle procedure. Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and further assume $|\mathcal{A}| = p_0 < p$, i.e. the true model depends only on a subset of the predictors. Denote the estimator $\hat{\beta}(\delta)$ obtained by a fitting procedure δ . Using the language of Fan and Li (2001), δ is an *oracle* procedure if $\hat{\beta}(\delta)$ asymptotically has the following properties:
 - Identifies the right subset, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
 - Has the optimal estimation rate, $\sqrt{n} \left(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}}^* \right) \rightarrow_d \mathcal{N}(\mathbf{0}, \Sigma^*)$, where Σ^* is the covariance matrix knowing the true subset model.
- ii. This very hierarchical setup can be exploited to do variable selection in the Bayesian analysis of the probit model. Following the structure given in Albert and Chib (1993), we can introduce the prior given in (2.23) for the regression coefficients and obtain a similar shrinkage effect. This will help us eliminate the redundant predictors from the model.

Bibliography

Bibliography

- Akaike, H. “A New Look at the Statistical Model Identification.” *IEEE Transactions on Automatic Control*, 19:716–723 (1974).
- Albert, J. H. and Chib, S. “Bayesian Analysis of Binary and Polychotomous Response Data.” *Journal of the American Statistical Association*, 88 (1993).
- Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer (2006).
- Breiman, L. “Better Subset Regression Using the Nonnegative Garrote.” *Technometrics*, 37(4):373–384 (1995).
- Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. *Monte Carlo Methods in Bayesian Computation*. Springer (2001).
- Chipman, H., George, E. I., and McCulloch, R. E. “The Practical Implementation of Bayesian Model Selection.” *IMS Lecture Notes - Monograph Series*, 38 (2001).
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. “Least Angle Regression.” *The Annals of Statistics*, 32(2):407–499 (2004).
- Fan, J. and Li, R. “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.” *Journal of the American Statistical Association*, 96:1348–1360 (2001).

- George, E. I. “The Variable Selection Problem.” *Journal of the American Statistical Association*, 95 (2000).
- George, E. I. and Foster, D. P. “Calibration and empirical Bayes Variable Selection.” *Biometrika*, 87(4) (2000).
- Geweke, J. “Bayesian Treatment of the Independent Student- t Linear Model.” *Journal of Applied Econometrics*, 8 (1993).
- Harville, D. A. “Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems.” *Journal of the American Statistical Association*, 72(358):320–338 (1977).
- Hoerl, A. E. and Kennard, R. W. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*, 12(1):55–67 (1970).
- Kass, R. E. and Raftery, A. E. “Bayes Factors.” *Journal of the American Statistical Association*, 90(430) (1995).
- Lindley, D. V. *The Estimation of Many Parameters*, in V. P. Godambe and D. A. Sprott (eds), *Foundations of Statistical Inference*. Holt Rinehart and Winston (1971).
- Lindley, D. V. and Smith, A. F. M. “Bayes Estimates for the Linear Model.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 (1972).
- Mallows, C. L. “Some Comments on C_P .” *Technometrics*, 15 (1973).
- Schwarz, G. “Estimating the Dimension of a Model.” *The Annals of Statistics*, 6 (1978).

- Sorensen, D. and Gianola, D. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. New York: Springer (2002).
- Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. “Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii. Radical prostatectomy treated patients.” *J. Urol.*, 141(5):1076–1083 (1989).
- Tibshirani, R. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288 (1996).
- Tierney, L. and Kadane, J. B. “Accurate Approximations for Posterior Moments and Marginal Densities.” *Journal of the American Statistical Association*, 81(393):82–86 (1986).
- Tipping, M. E. “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research*, 1 (2001).
- Yuan, M. and Lin, Y. “Efficient Empirical Bayes Variable Selection and Estimation in Linear Models.” *Journal of the American Statistical Association*, 100(472):1215–1225 (2005).
- Zou, H. “The Adaptive Lasso and Its Oracle Properties.” *Journal of the American Statistical Association*, 101:1418–1429 (2006).
- Zou, H. and Hastie, T. “Regularization and variable selection via the elastic net.” *Journal Of The Royal Statistical Society Series B*, 67(2):301–320 (2005).

Appendix

Appendix

A Some Useful Probability Density Functions

Univariate Normal Density

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} (\theta - \mu)^2 \right] \quad (\text{A.1})$$

$\mu \in \mathbb{R}$ location parameter, $\sigma \in \mathbb{R}^+$ scale parameter, $\mathbb{E}(\theta) = \mu$, $\mathbb{V}(\theta) = \sigma^2$, $\text{mode}(\theta) = \mu$

Multivariate Normal Density

$$p(\theta) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp \left[-\frac{1}{2} (\theta - \mu)' \Sigma^{-1} (\theta - \mu) \right] \quad (\text{A.2})$$

μ p -dimensional location parameter, Σ $p \times p$ symmetric positive definite variance-covariance matrix, $\mathbb{E}(\theta) = \mu$, $\mathbb{V}(\theta) = \Sigma$, $\text{mode}(\theta) = \mu$

Univariate Student- t Density

$$p(\theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\nu\pi}\sigma} \left[1 + \frac{1}{\nu} \left(\frac{\theta - \mu}{\sigma} \right)^2 \right]^{-(\nu+1)/2} \quad (\text{A.3})$$

$\mu \in \mathbb{R}$ location parameter, $\sigma \in \mathbb{R}^+$ scale parameter, $\nu \in \mathbb{R}^+$ degrees of freedom, $\mathbb{E}(\theta) = \mu$ for $\nu > 0$, $\mathbb{V}(\theta) = \frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$, $\text{mode}(\theta) = \mu$

Multivariate Student- t Density

$$p(\theta) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}} |\Sigma|^{-1/2} \left[1 + \frac{1}{\nu} (\theta - \mu)' \Sigma^{-1} (\theta - \mu)\right]^{-(\nu+p)/2} \quad (\text{A.4})$$

μ p -dimensional location parameter, Σ $p \times p$ symmetric positive definite scale matrix, $\nu \in \mathbb{R}^+$ degrees of freedom, $\mathbb{E}(\theta) = \mu$ for $\nu > 1$, $\mathbb{V}(\theta) = \frac{\nu}{\nu-2} \Sigma$ for $\nu > 2$, $\text{mode}(\theta) = \mu$

Laplace (Double Exponential) Density

$$p(\theta) = \frac{1}{2\sigma} \exp\left(-\frac{|\theta - \mu|}{\sigma}\right) \quad (\text{A.5})$$

$\mu \in \mathbb{R}$ location parameter, $\sigma \in \mathbb{R}^+$, $\mathbb{E}(\theta) = \mu$, $\mathbb{V}(\theta) = 2\sigma^2$, $\text{mode}(\theta) = \mu$

Gamma Density

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\beta\theta) \quad (\text{A.6})$$

$\alpha \in \mathbb{R}^+$ shape parameter, $\beta \in \mathbb{R}^+$ inverse scale parameter, $\mathbb{E}(\theta) = \alpha/\beta$, $\mathbb{V}(\theta) = \frac{\alpha}{\beta^2}$, $\text{mode}(\theta) = \frac{\alpha-1}{\beta}$ for $\alpha \geq 1$ (Note that in our formulation $\eta = \alpha - 1$)

Inverse Gamma Density

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} \exp(-\beta/\theta) \quad (\text{A.7})$$

$\alpha \in \mathbb{R}^+$ shape parameter, $\beta \in \mathbb{R}^+$ scale parameter, $\mathbb{E}(\theta) = \frac{\beta}{\alpha-1}$ for $\alpha > 1$, $\mathbb{V}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$, $\text{mode}(\theta) = \frac{\beta}{\alpha+1}$ for $\alpha \geq 1$

χ^2 Density

$$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{\nu/2-1} \exp(-\theta/2) \quad (\text{A.8})$$

$\nu \in \mathbb{R}^+$ degrees of freedom, $\mathbb{E}(\theta) = \nu$, $\mathbb{V}(\theta) = 2\nu$, $\text{mode}(\theta) = \nu - 2$ for $\nu \geq 2$

Inverse χ^2 Density

$$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} \theta^{-(\nu/2+1)} \exp(-1/2\theta) \quad (\text{A.9})$$

$\nu \in \mathbb{R}^+$ degrees of freedom, $\mathbb{E}(\theta) = \frac{1}{\nu-2}$ for $\nu > 2$, $\mathbb{V}(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}$ for $\nu > 4$,
 $\text{mode}(\theta) = \frac{1}{\nu+2}$

Scaled Inverse χ^2 Density

$$p(\theta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} s^\nu \theta^{-(\nu/2+1)} \exp(-\nu s^2/2\theta) \quad (\text{A.10})$$

$\nu \in \mathbb{R}^+$ degrees of freedom, $s \in \mathbb{R}^+$ scale parameter, $\mathbb{E}(\theta) = \frac{\nu}{\nu-2} s^2$ for $\nu > 2$,
 $\mathbb{V}(\theta) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} s^4$ for $\nu > 4$, $\text{mode}(\theta) = \frac{\nu}{\nu+2} s^2$

B Conditional Maximization

Let

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (\text{B.1})$$

be the unnormalized joint posterior distribution of the parameter set $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)'$, $d = 1, 2, \dots$ and D denotes the data. If $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_d)'$ is the joint posterior mode, then these values satisfy the equations

$$\left. \frac{\partial}{\partial \theta_1} p(\theta_1, \theta_2, \dots, \theta_d | D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \dots = \left. \frac{\partial}{\partial \theta_d} p(\theta_1, \theta_2, \dots, \theta_d | D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = 0. \quad (\text{B.2})$$

These equations may be rewritten in terms of conditional and marginal distributions.

$$\begin{aligned} \left. \frac{\partial}{\partial \theta_1} p(\theta_1 | \theta_2, \dots, \theta_d, D) p(\theta_2, \dots, \theta_d | D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} &= 0 \\ &\vdots \\ \left. \frac{\partial}{\partial \theta_d} p(\theta_d | \theta_1, \dots, \theta_{d-1}, D) p(\theta_1, \dots, \theta_{d-1} | D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} &= 0 \end{aligned} \quad (\text{B.3})$$

Provided that $p(\tilde{\theta}_2, \dots, \tilde{\theta}_d | D) \neq 0, \dots, p(\tilde{\theta}_1, \dots, \tilde{\theta}_{d-1} | D) \neq 0$,

$$\begin{aligned} \left. \frac{\partial}{\partial \theta_1} p(\theta_1 | \theta_2, \dots, \theta_d, D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} &= 0 \\ &\vdots \\ \left. \frac{\partial}{\partial \theta_d} p(\theta_d | \theta_1, \dots, \theta_{d-1}, D) \right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} &= 0. \end{aligned} \quad (\text{B.4})$$

Equation (B.4) implies that the joint posterior mode $\tilde{\boldsymbol{\theta}}$ can be obtained via iterative conditional maximizations.

C Proofs of Theorems

Proof of Theorem 1. β and σ^2 can tractably be integrated out of (??). As a result of this integration, the only remaining terms that are dependent upon v_j are

$$|\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1}|^{-1/2} |\mathbf{V}|^{-1/2} \left(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y} \right)^{-n/2} \prod_{j=1}^p p(v_j^{-1}). \quad (\text{C.1})$$

It will suffice to show that (C.1) is finitely integrable with respect to v_j^{-1} .

$$|\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1}|^{-1/2} |\mathbf{V}|^{-1/2} = |\mathbf{X}'\mathbf{X}\mathbf{V} + \mathbf{I}|^{-1/2} < |\mathbf{X}'\mathbf{X}\mathbf{V}|^{-1/2} = |\mathbf{X}'\mathbf{X}|^{-1/2} |\mathbf{V}|^{-1/2}, \quad (\text{C.2})$$

and

$$\mathbf{y}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \mathbf{V}^{-1})^{-1} \mathbf{X}'\mathbf{y} \leq \mathbf{y}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}. \quad (\text{C.3})$$

Eliminating the terms again that are not dependent upon v_j , we reduce (C.1) to

$$|\mathbf{V}|^{-1/2} \prod_{j=1}^p p(v_j^{-1}). \quad (\text{C.4})$$

Integrating (C.4) is equivalent to $\prod_{j=1}^p \mathbb{E} v_j^{-1/2}$. This expectation is taken with respect to (3.4) and is finite for $\mu > 0$. $\square$

Vita

Artin Armagan received his BS degree in Mechanical Engineering in 2001 from Istanbul Technical University. After having worked as a Mechanical Engineer for a year he then received an MS degree in 2004 and a PhD degree in 2008 in Statistics from the University of Tennessee, Knoxville. He will be starting a 2-year appointment as a visiting assistant professor at the Department of Statistical Science at Duke University in Fall 2008. His research interests include Bayesian shrinkage estimation and model selection for linear models, application of Bayesian statistical models in engineering etc.