Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Ritabrata Dutta

Entitled
MODEL SELECTION : BAYES AND FREQUENTIST PERSPECTIVE

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Jayanta K Ghosh - Co-chair
Chair

Guang Cheng - Co-chair

Alan Qi

Xiao Wang

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Jayanta K Ghosh

Guang Cheng

Approved by: Jun Xie 05/04/2012
Head of the Graduate Program Date

Graduate School Form 20
(Revised 9/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

MODEL SELECTION : BAYES AND FREQUENTIST PERSPECTIVE

For the degree of Doctor of Philosophy

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Ritabrata Dutta

Printed Name and Signature of Candidate

05/04/2012

Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

MODEL SELECTION : BAYES AND FREQUENTIST PERSPECTIVE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ritabrata Dutta

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2012

Purdue University

West Lafayette, Indiana

To my family and friends

ACKNOWLEDGMENTS

I would like to thank my advisors Professor Jayanta K Ghosh and Professor Guang Cheng for being a continual source of support, guidance and good ideas. Additionally, I would like to thank Professor Xiao Wang and Alan Qi for their patience and the helpful suggestions while serving on my committee. Special thanks to Professor Rebecca Doerge and Professor Anirban Dasgupta for all the discussions and encouragements. My gratitude also extends to everyone in the Statistics Department to make my four years in Purdue wonderful. At this scope, I would also like convey my love to all the wonderful friends I met in Purdue campus and in the United States of America. I can not express how grateful I am to them for their friendship, as well as for aiding my growth as who I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Dutta, Ritabrata Ph.D., Purdue University, August 2012. Model Selection : Bayes and Frequentist Perspective. Major Professor: Professor Jayanta K Ghosh and Professor Guang Cheng.

We discuss model selection, both from a Bayes and Classical point of view. Our presentation introduces a novel point of view about goals and methods of model selection. Our hope is that this will introduce a logical overview connecting all model selection rules. We start with old Bayes and Classical rules like AIC and BIC, then develop some new theory and a new novel Bayes strategy within this discussion. We introduce some new definitions of consistency and results and conjectures about consistency in high dimensional model selection problems. For model selection with Cross-validatory Bayes Factor, we show that when the number of parameters tends to infinity at a smaller rate than sample size, to achieve consistency it is best to use most of the data for inference and only a negligible proportion to result in a proper prior. In Chapter 2 we take up a major method of Bayes model selection, namely, Path Sampling (PS) for detailed study via new theorems and novel diagnostics, finally leading to a new method PS-SC. Most of this chapter contains new material.

In the last two chapters we consider classical model selection. We first provide a new approach to selection for SNP's of great interest in Bioinformatics as well as model selection methodology in high dimensional linear models. Our last chapter is a contribution to model selection in nonparametric regression. We take up model selection for variable selection in this context. We believe this has the potential to improve model selection in more traditional areas where LASSO and its modification are very popular. This will require much further work. Works presented here have already been published as journal articles Dutta and Ghosh [2011a], Dutta et al. [2011], Dutta and Ghosh [2011b].

1. MODEL SELECTION - A BAYES OVERVIEW AND SOME NEW RESULTS

1.1 Early history of Bayes and classical model selection

To the extent that Bayes model selection, with the goal of choosing the correct model, is an extension of or identical with testing of two hypotheses, Bayes model selection may be said to have begun with the work of Jeffreys in 1939, see Jeffreys [1961]. The Bayes test, as well as an approximation to it due to Lindley, is mentioned in Cox [1961]. Cox [1961, 1962] pioneered model selection in classical statistics, but without bringing in the novel concept of penalization for complexity of a model.

Both in Cox [1961] and Cox [1962], one considers two separated hypotheses, i.e. hypotheses or models having the property that no density in one is obtainable as a limit of densities in the other. Cox does not specify the notion of convergence but in a follow up study, Ghosh and Subramanyam [1975] suggest that the limit may be taken in the sense of convergence in L_1-norm. This is equivalent to requiring the two sets of densities can be covered by two disjoint L_1-open sets. An alternative definition, depending on n, is given towards the end of the previous reference.

Essentially, Cox's model selection rule is based on the maximized likelihood of data under each model. Subramanyam and Ghosh, vide Subramanyam [1979] show that the true model is rejected with exponentially small probability. To show this, one has to use results on large deviations. They verify the conditions for one of Cox's examples, where the true model is either Geometric or Poisson.

It is remarkable that both in Jeffreys [1961] and Cox [1961, 1962] the goal is the same - to choose the correct model. It is equally remarkable that, starting with AIC, choosing the correct model has not been a goal in classical model selection, at least not explicitly.

If one reads Akaike's early papers and resolves the ambiguities in the light of the subsequent pioneering theoretical papers of Shibata [1994], it becomes clear that the objective is to choose a model so that one predicts optimally the data on the dependent variables in an exact new replicate of the design for the given data. Assuming the true model is more complex than the assumed linear models, Shibata's calculations, and also the calculations in Li [1987] and Shao [1997], show that the special form of penalty associated with AIC has a special role in this kind of optimal prediction. Optimality is proved through a lower bound to the predictive loss or risk of all so called penalized likelihood rules that may be used in this problem, and then showing AIC attains the lower bound asymptotically because of its special penalty. This lower bound would now be called an oracle, a term which was probably introduced later formally in the model selection literature by Johnstone and Silverman [2004]. It has been shown by Van de Geer [2006] that there is a close connection between the penalty of a predictive model selection criteria and the oracle it may attain.

In Bayes model selection, the picture is far more mixed. One sees papers of both kinds, selecting the correct model or predicting well. The rich literature on Bayes model selection is reviewed in Clyde and George [2004].

A historian of model selection may well speculate about this curious divergence in development. To us it seems this may be due to the clarity of all aspects, including goals of inference, that one finds readily in the Bayesian paradigm.

In the first of the next two subsections, we first introduce Bayes model selection through Bayes factors and then Schwarz's BIC. In the next subsection we discuss a few popular model selection rules based on penalized likelihood.

1.1.1 Bayes rule for selection of true model: BIC, consistency in Bayes model selection

Assume first that we have two models $M_i,\ i = 1, 2$. With M_i, we associate a parameter space Θ_i, a density $f_i(x_1, x_2, \cdots, x_n|\theta_i)$ for the data $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in$

$I\!R_n$ and a prior $\pi_i(\theta_i)$ for θ_i. We assume the families $\{f_i(\mathbf{x}|\theta_i), \theta_i \in \Theta_i\}, i = 1, 2$ are separated in the sense explained earlier.

Assume the Bayesian has also prior probabilities λ_1, λ_2 for M_1 and M_2, $\lambda_1+\lambda_2 = 1$. Let

$$m_i(\mathbf{x}) \;=\; \text{marginal density of } \mathbf{x} \text{ under } M_i = \int_{\Theta_i} f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i)d\theta_i.$$

Then the posterior probability of M_i, given $\mathbf{x}$, is proportional to $\lambda_i m_i(\mathbf{x})$. Hence one would select M_2 if it is more likely than M_1 given data, i.e., if

$$\frac{\lambda_2 m_2(\mathbf{x})}{\lambda_1 m_1(\mathbf{x})} > 1,$$

and the other way if the ratio is < 1. Jeffreys [1961] has suggested a scale of numerical values for different levels of relative credibility of the two models. The Bayes rule is easily extended when we have k separated models.

The usual default choice for λ_1, λ_2 is $\lambda_1 = \lambda_2 = \frac{1}{2}$, in which case our selection criterion becomes:

$$\text{Select } M_2, \text{ if } \frac{m_2}{m_1} > 1.$$

The ratio $\frac{m_2}{m_1}$ is called the Bayes factor and denoted by BF_{21}.

We usually assume that each model specifies iid models $f_i(\mathbf{x}|\theta_i) = \prod_1^n f_i(x_j|\theta_i)$.

While Bayes model selection is straightforward in principle, numerical calculation of the Bayes factor is not. One needs Reversible Jump MCMC or Path Sampling, Andrieu et al. [2004], Dutta and Ghosh [2011a]. Schwarz [1978]'s BIC provides a convenient approximation, assuming fixed dimension p_i of Θ_i and suitable regularity conditions, and letting sample size $n \to \infty$,

$$\log m_i = \sum_{j=1}^{n} \log f_i(x_j|\hat{\theta}_i) - \frac{p_i}{2}\log n + O_p(1),$$

where p_i, equal to the dimension of Θ_i, is a measure of the complexity of M_i, and $\hat{\theta}_i$ is the MLE under model M_i. Thus $\log m_i$ can be approximated by

$$BIC(M_i) = \sum_{j=1}^{n} \log f_i(x_j|\hat{\theta}_i) - \frac{p_i}{2}\log n, \tag{1.1}$$

which may be interpreted as a penalized maximum likelihood corresponding to a model. The bigger the dimension of Θ_i the bigger is the penalty, in tune with the scientific principle of parsimony. One selects a model by maximizing equation 1.1 with respect the model M_i, i.e., by choosing suitable density f_i. This is the user's guess for the unknown true model.

Thus Bayesian model selection through Bayes factors automatically obeys the principle of parsimony. It tries to compensate for the fact that the bigger the Θ_i, the bigger we expect the maximum likelihood to be.

As theoretical validation of use of Bayes factor in a problem, one tries to prove consistency in some sense. Below is the usual definition adopted by Bayesians, but we do not know of a reference. We first consider the case of two separated models M_1, M_2 with our usual notation. The nested case will be discussed in the next section.

Definition of Consistency. We say consistency holds at $\theta \in \Theta_2$ if under θ, $\log BF_{21} \to \infty$, in probability. On the other hand consistency holds at $\theta \in \Theta_1$, if $\log BF_{21} \to -\infty$, in probability.

Our definition of consistency here agrees with that in Liang et al. [2008] and Fernandez et al. [2001]. Consistency in this sense will appear in our discussions many times. Also in Section 1.2 a definition of an alternative notion will be given. It will be used only in the example following that definition.

1.1.2 AIC tries to predict and BIC to choose the correct model

We start this subsection by digressing a little on nested models and some issues that arise when one of two nested models have to be selected. M_1 is said to be nested in M_2 if $\Theta_1 \subset \Theta_2$ and for $\theta \in \Theta_1$, $f_1(x_j|\theta) = f_2(x_j|\theta)$. This is a very different situation from the separated models that we have been considering earlier. If $\theta \in \Theta_2$ but $\notin \Theta_1$, we may say M_2 is true and the density is $f_2(x|\theta)$. But what about the case where $\theta \in \Theta_1$ and hence $\theta \in \Theta_2$ also? Here both models are correct, so which one do we choose? Bayesians suggest that on grounds of parsimony one should choose the smaller model M_1.

What about priors over M_1 and M_2? Typically, we have priors $\pi_1(\theta_1)$ on M_1 and $\pi_2(\theta_2)$ on M_2 such that $\pi_2(\theta_2)$ is a density with respect to Lebesgue measure on Θ_2 and hence assigns zero probability to Θ_1. Typically, $\pi_1(\theta_1)$ has a density with respect to Lebesgue measure on $\Theta_1 \subset \Theta_2$, so it assigns zero probability to $\Theta_2 \cap \Theta_1^c$.

The problem of selection of one of two nested models has puzzled philosophers of science. They feel one should always choose the bigger model M_2 as true. How can M_2 be rejected if M_1 is true? For their views, as well as what Bayesians have to say in response about logical consistency of nested model selection, see Chakrabarti and Ghosh [2011]. They try to explain through Galileo's famous experiment at Pisa. The smaller model represents Gallileo's view, while the bigger model represents the prevalent view at that time about falling bodies. The two priors represent idealized versions of two such views. Typically when we deal with linear models, choosing means or choosing non-zero regression coefficients in regression problems, we have to select from among models, some of which are nested in others. This is the most common model selection problem.

We now turn to a comparison of AIC and BIC. Historically BIC (Bayes Information Criterion) is the second penalized maximum likelihood rule. The first is AIC (Akaike Information Criteria). For linear models with normal error and known variance σ^2, AIC proposed by Akaike [1974a] is

$$AIC(M_i) = 2\sum_{j=1}^{n} \log f_i(x_j|\hat{\theta}_i) - 2p_i. \tag{1.2}$$

One has to multiply BIC by two or divide AIC by two to make them comparable. The expression (1.2) is maximized over models to get the best predictive model. Prediction is made by using the mle under the chosen model.

The first term of AIC is taken to be twice the more intuitive first term of BIC, because the difference of the first terms for AIC for two models has asymptotically a

χ^2-distribution if the smaller model is true. We could have done the same for BIC by doubling the penalty.

We recall that though AIC and BIC look similar, they are meant to do very different things, a fact that seems to have been first noted in Mukhopadhyay [2000]. AIC predicts optimally, while, as pointed out above, BIC is a Bayesian criterion for selecting the true model. Each will perform poorly if used for a purpose for which it was not meant.

We now turn to an example of a simple linear model, with normal error N(0,1)

$$Y_i = \mu + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

Suppose we wish to choose between $M_1 : \mu = 0$ and $M_2 : \mu$ is arbitrary, i.e., $\Theta_1 = \{0\}$, $\Theta_2 = I\!R$. If we use AIC to select the true model, then it is easy to verify that we are actually using a test with Type 1 error $= P\{\chi^2_1 > 2\} > .05$, i.e., the test is more liberal than the usual most liberal test. So, there is no parsimony if we use AIC in a testing problem.

On the other hand suppose we use BIC. BIC is very parsimonious and can be shown to choose the correct model with probability tending to one, at least for fixed p and $n \to \infty$. If we have linear models and our goal is prediction, AIC will be predictively optimal as discussed earlier. For ease of reference we consider Shao [1997]. Note that Shao's theorem is proved in the context of general linear models, and the main assumption is that the true linear model is not in the model space. Shao's theorem is an illustration of Box's famous remark that all models are false but some are useful. Moreover Shao's theorem shows AIC helps one identify the most useful model from the point of view of prediction. The proof is based on an oracle. Chakrabarti [2007] have used a similar oracle property of AIC to show heuristically that it is an adaptive, asymptotically minimax, estimate of the unknown non-parametric regression function f in the following model,

$$Y_i = f(t_i) + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \cdots, n, \quad t_i = \frac{i}{n}.$$

Computation in Chakrabarti [2004] shows AIC is competetive with popular best methods in non-parametric regression.

On the basis of comparison of AIC and BIC, we suggest tentatively that model selection rules should be used for the purpose for which they were introduced. If they are used for other problems, a fresh justification is desirable. In one case, justification may take the form of a consistency theorem, in the other some sort of oracle inequality. Both may be hard to prove. Then one should have substantial numerical assessment over many different examples.

1.1.3 Predictive Bayesian model selection - model average

If the object of Bayes model selection is not to select the true model but predict future observations well, the utility or loss changes dramatically from 0-1 loss to a conditional expectation of squared error loss of predictor of future observations x_f. Instead of the squared error one may choose other suitable loss functions for prediction. The Bayesian solution of this problem is straightforward but very different from the model with highest posterior probability, which is what we have discussed so far.

For prediction it is best not to select a single model but rather average over best predictions from all models, i.e., use the so called predictor based on model average. Assuming conditional independence of current data x and future data x_f, given θ, calculate the model average prediction

$$\frac{\sum P(M_j|X_j) \int E(x_f|\theta_j, M_j) p(\theta_j|M_j) d\theta_j}{\sum P(M_j|X_j)}.$$

If, on the other hand, one wants to choose a suitable model M_j and then use the predictor assuming M_j is true, then it turns out that the median model of Barbieri and Berger [2004] is optimal under orthogonality assumptions. A quick introduction to both model average and median model is available in [Ghosh et al., 2006, Ch. 9]. As mentioned earlier, a posterior quantile model selection rule and its optimality is studied in Mukhopadhyay and Ghosh [2003].

Finally computation is a major concern, specially when both the dimensions of models and the number of models are very large. Many algorthims are available, see e.g. Clyde and George [2004].

1.2 Consistency in high dimensional model selection

Technically, consistency of model selection in high dimensional problems is usually very hard to prove, because our knowledge of the asymptotic behavior of the Bayes factor under a fixed θ is still quite meagre. For fixed (or slowly increasing) dimensions we have approximations like the BIC. These are no longer generally available in the high dimensional case.

In this section we try to do three things. We first discuss in detail a high dimensional model selection problem proposed by Stone [1974] and discussed in Berger et al. [2003]. Then we explore tentatively a weaker definition that may be both more satisfactory and often verifiable. The new definition reduces model selection to a test of two simple hypotheses. In at least one problem we show, we can get quite definitive results for a whole class of priors (Theorem 1).

We have also made a couple of general conjectures involving positive results on consistency (in the usual sense) under M_1. Consistency under M_1 is very important from the point of view of parsimony.

In Subsection 1.2.1 we try to initiate a change in our perspectives. We first attempt to pose a few general questions, make conjectures and suggest how they may be answered. It is no more than a modest attempt to inspire others to take up the challenges of consistency in model selection and, through that effort, get a better insight about high dimensional model selection.

1.2.1 Generalized BIC for Stone's high dimensional example and posterior consistency

Common inference procedures for model selection in high dimensional examples may behave in a very different way from those in low dimensional examples. This subsection gives such an example, essentially due to Stone, studied in Berger et al. [2003]. This subsection is based on Berger et al. [2003]. We try to initiate new work, as indicated earlier.

The following is a slight modification of a high dimensional linear model due to Stone, showing (as in Stone's original example) that in high dimensional problems, BIC need not be consistent but AIC is. Consider the model

$$y_{ij} = \mu_i + \epsilon_{ij},\ i = 1, \ldots, p,\ j = 1, \ldots, r,\ \textit{and}\ \epsilon_{ij}\ \textit{are i.i.d.}\ N(0, 1). \tag{1.3}$$

The two competing models are $M_1 : \mu = 0$ and $M_2 : \mu \in R^p$. The dimension $p \to \infty$.

Berger et al. [2003] show that the BIC is a very bad approximation to the marginal under one of the models, and hence its inconsistency is unrelated to consistency of Bayes factors in general. We reproduce a table (Table 1.2.1) showing the poor approximation obtained from BIC and the much better approximations obtained from the generalization, called GBIC, in Berger et al. [2003]. A naive application of the relatively standard Laplace approximation due to Kass, Wasserman and Pauler, denoted in Berger et al. [2003] as Lap_{KWP} provides a good approximation but not as good as the best. The best approximation is provided by the rigorous Laplace approximation for the BF based on high dimensional Cauchy, given in the last column. But even this approximation is poor near small values of c_p, i.e., when the alternative model (M_2) is likely to be close to the null model (i.e., M_1).

We now turn to consistency issues, our main interest in this section. Berger et al. [2003] consider a family of priors including the popular Zellner-Siow multivariate Cauchy prior and the Smooth Cauchy prior due to Berger and Pericchi [1996]. It turns out that the Bayes factor for the multivariate Cauchy prior is consistent under both M_1 and M_2, but is not consistent under the smooth prior. At the time Berger

Table 1.1
Log (Bayes factor) and its approximations under the Cauchy prior. (Table taken from Mukhopadhyay [2000])

c_p	$\log BF^c$	BIC	GBIC	Lap^c_{KWP}	$\log BF^{HD_c}$
.1	-8.53	-110.12	-1.95	15.12	-8.57
.5	-3.82	-90.12	-1.95	-2.26	-3.908
1.0	6.03	-65.12	5.71	5.55	5.92
1.5	20.82	-40.12	20.57	20.38	20.75
2.0	38.48	-15.12	38.38	39.13	38.44
2.1	42.23	-10.12	42.16	41.89	42.19
10.0	397.36	384.87	398.15	397.29	397.36

et al. [2003] was written, the paper pioneered a study of consistency of Bayesian model selection as well as use of BIC to approximate a Bayes factor in high-dimensional problem. In retrospect one understands consistency much better and some more insight as well as new results can be provided. We do this below.

We first note that the two models are nested, the parameter space under M_1 is a singleton, namely it has only the point with all coordinates equal to zero, and finally though this point is not contained in Θ_2, it lies in $\bar{\Theta}_2$. This is basically like testing a sharp null and with a disjoint but not topologically separated alternative, i.e., we can not find two disjoint open sets, one containing $\theta_0 = 0$ and the other Θ_2. Thus any open set containing θ_0 will have non-empty overlap with Θ_2. For a discussion of the same point in Bayesian nonparametrics, see Tokdar et al. [2010].

That, the posterior for the Zellner-Siow prior (denoted as Z-S prior or π_c) is consistent inspite of lack of separation of M_1 and M_2 may be intuitively explained by examining the structure of this prior as well as that of the Smooth Cauchy prior (denoted as π_{sc}). The Z-S prior may be represented as a mixture of multivariate normals with a gamma prior for the common precision parameter $t = \frac{1}{\sigma}$, eqn. (6) of Berger et. al.

$$\begin{aligned}\pi_c(\mu) &= \frac{\Gamma((p+1)/2)}{\pi^{(p+1)/2}}(1+\mu'\mu)^{-(p+1)/2} \\ &= \int_0^\infty \frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\mu'\mu} \frac{1}{\sqrt{2\pi}} e^{-t/2} t^{-\frac{1}{2}} dt.\end{aligned}$$

The gamma mixing measure puts positive mass near precision parameter $t = \infty$, i.e., in the σ-space, near $\sigma = 0$, making even small differences detectable (under both M_1 and M_2). This seems to be the reason why $\pi_c(\mu)$ is a consistent prior. On the other hand it appears that

$$\pi_{sc}(\mu) = \int_0^1 \frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\mu'\mu} \frac{1}{\pi\sqrt{t(1-t)}} dt$$

has inconsistent posterior (under M_2) because it is supported on the set (0,1), with $t = \infty$, i.e. $\sigma = 0$ (in the σ-space) not in the support of the mixing measure. Some comments on these facts and proofs in Berger et al. [2003] are also in order.

Consider the general family of priors

$$\pi_g(\mu) = \int_0^\infty (\frac{t^{p/2}}{(2\pi)^{p/2}} e^{-(t/2)\mu'\mu}) g(t) dt. \tag{1.4}$$

All priors with the general structure given by Berger et al. [2003] in (1.4) ensure posterior consistency under M_1 even if $g(t)$ is strictly positive only on $(0, T]$, for arbitrary $T > 0$. The condition $g(t) > 0$ on $(0, \infty)$ is needed for consistency under M_2, but not under M_1. (In the proof of Theorem 3.1 in Berger et al. [2003], without this assumption the set S_ϵ is empty, but that does not matter under M_1.) Thus the smooth Cauchy with support of g equal to $(0, 1]$ has consistent posterior under M_1, and under M_2 it is inconsistent if $0 < \tau^2 < 2\log(2) - 1$, where $\tau^2 = \lim_{p\to\infty} \frac{1}{p} \sum \mu_i^2$.

A similar result holds if the support of g is $(0, T]$. The most interesting fact that emerges is the rather general consistency theorem under M_1. For those of us who believe in parsimony, this is a very pleasant fact.

In Scott and Berger [2010], we note a similar fact under the global null, provided the global null is given a positive prior probability. One gets posterior consistency under this model by straightforward applications of Doob's Theorem ([Ghosh et al., 2006, p.22]), as in the proof of similar results for Bayesian Nonparametric models by Dass et al. [2004].

Conjecture. All this suggests a general consistency theorem under the global null for linear models remains to be discovered and that one would need to try to prove a general version of Doob's theorem for independent but not identically distributed r.v.'s. A simple example appears in Choi and Ramamoorthi [2008].

So far in this chapter we have been using consistency as defined in Section 1.1.1. We now define a new notion of consistency, which is simply consistency of Bayes factors under the two marginals of $\{X_i\}$ under M_1 and M_2.

Definition of $(\mathbf{P_1}, \mathbf{P_2})$-consistency Let P_1 and P_2 be the infinite-dimensional marginal distributions of Y_i's under M_1 and M_2. We say consistency holds, iff under the true infinite-dimensional marginal distributions $P_1,\ P_2$, $\log BF_{21} \to -\infty$ and ∞ in probability. We call this (P_1, P_2)-consistency to distinguish from usual consistency.

This should be particularly attractive, if the data $(Y_1, \cdots, Y_n)$ have been generated as assumed in Scott and Berger [2010], i.e., under P_1 or P_2 as true, not an unknown θ. If this model is correct, a large enough data is expected to choose the correct model with very high probability.

We prove (P_1, P_2)-consistency for Stone's example.

Theorem 1 *Assume the general family of priors in equation (1.4) with* $0 < t < \infty$ *w.p.* 1 *under* $g(t)$. *Then* (P_1, P_2)*-consistency holds.*

Proof Let $\bar{Y}_i = \frac{1}{r}\sum_{j=1}^{r} Y_{ij}$. Conditionally, for fixed value of the precision parameter t, under M_2, $\frac{1}{p}\sum_1^p \bar{Y}_i^2 \to \frac{1}{r} + \frac{1}{t}$ a.s. Since $t > 0$ with probability one, under P_2 (obtained by integrating out t),

$$\text{liminf}\frac{1}{p}\sum_1^p \bar{Y}_i^2 > \frac{1}{r},\ a.s.$$

On the other hand, under P_1,

$$\text{liminf}\frac{1}{p}\sum_1^p \bar{Y}_i^2 \to \frac{1}{r},\ a.s.$$

The above facts show P_1 and P_2 are orthogonal (i.e., P_1 and P_2 are supported on disjoint subsets of the sample space $\{Y_i\}$) and hence, by standard facts about likelihood ratios (Kraft [1955]) the Bayes factor

$$\log BF_{21} \ \to \ \infty \quad a.s.(P_2), \qquad \log BF_{21} \to -\infty \ \ a.s.(P_1),$$

proving posterior consistency under both M_1 and M_2.

■

It seems plausible to us that just as consistency under M_1 may hold rather generally (by Doob's theorem) for linear models, posterior consistency under P_2 may also be true rather generally. Moreover even when the question cannot be settled theoretically, a Bayesian simulation as in Scott and Berger [2010] will throw light on whether consistency is to be expected or not. All this will bypass the need to have good Laplace approximations to $m_1(\mathbf{x})$, $m_2(\mathbf{x})$ in high dimensional cases.

The general linear model, of which Berger et al. [2003] is a very simple special case, has been studied in a greater detail in Liang et al. [2008] from the point of view of choice of new priors, calculation of Bayes factor and consistency (for the fixed p case). However, posterior consistency for the high dimensional case doesn't seem to be have been studied.

On the other hand Moreno et al. [2010] study posterior consistency for some general high dimensional regression problems when intrinsic priors are used. The results are quite interesting but the formulations of consistency are somewhat different.

1.3 Cross-Validatory Bayes factor

1.3.1 General issues

It has been known for quite some time that Bayesian estimation of parameters or prediction of future observations is quite robust with respect to the choice of prior while model selection (for 0-1 loss) based on Bayes factors is not robust. Draper and Krnjajic [2010] discuss instability of Bayes factors and suggest replacing them with cross-validatory Bayes factors.

Very roughly speaking, in estimation one uses diffuse improper or diffuse proper priors, so that most of the information in the posterior come from the data, not the prior. In particular the undetermined constants in an improper prior gets canceled because it appears in both the numerator and denominator of the basic Bayes Formula:

$$p(\theta|x) = \frac{c.p(\theta)p(x|\theta)}{\int c.p(\theta)p(x|\theta)d\theta}.$$

On the other hand testing procedures or model selection methods based on Bayes factor do not have these good properties, see e.g., Ghosh and Samanta [2002] and Ghosh et al. [2006]. A standard way of solving both problems is to use data based priors as follows. One uses a part of the data, say a vector x_k to calculate the posterior for each model which is then used as a prior for the corresponding model. For each model the corresponding data based prior is then combined with the remaining data set x_{-k} consisting of the remaining $(n-k)$ x_i's, to produce a marginal based on x_{-k}. With these modifications, both difficulties, namely the appearence of the arbitrary constant in an improper prior and lack of robustness with respect to the prior disappear. All of the new priors are really posteriors and, hence, robust if based on substantial data.

There is a huge literature on these cross-validatory Bayes factors. We mention a few, based partly on their importance and partly on our familiarity, Geisser [1975], Berger and Pericchi [1996], Ghosh and Samanta [2002], O'Hagan [1995], Chakrabarti [2007], Draper and Krnjajic [2010]. Berger and Pericchi [1996, 2004] adopt the same procedure but condition w.r.t. what they call the minimal training sample, that is, a smallest subsample such that conditioning w.r.t. it makes the posterior proper. Then averages are taken over minimal training samples. We note in passing, that this method has actually led to construction of objective priors, which Berger and Pericchi [2004] call intrinsic priors. The papers of Berger and Pericchi [2004] and other colleagues provide many details and interpretations. We need to know only the basic facts stated above. *Berger and Pericchi are not trying to get a stable Bayes factor, they are trying to get a Bayes factor which may be called "objective".*

These new Bayes factors are more stable but raise a new issue which is still not fully understood. How much of the data should be used to make the prior stable (by computing the posterior and treating the posterior as a data based prior) and how much of the data should be used for inference? We first heard this question from Prof. L. Pericchi. This is a deep and difficult question. It has been discussed in Chakrabarti [2007]. In the next subsection we turn to the problem involving what level of cross-validation should be chosen if we are in the M-closed case of Bernardo and Smith [1994]. This is the usual case where all the models considered are in the model space.

The cross-validatory Bayes factors have a long history, suggested by Bernardo and Smith [1994]. They were extended by Gelfand and Dey [1994], who in turn had drawn on Geisser [1975]. Consistency issues were studied by Chakrabarti [2007], assuming these Bayes factors as given - a major motivation was to throw some light on Pericchi's question but not settle any of the other basic normative issues, specially in the context of selecting a true model or one close to it in some sense. Many interesting alternative approaches were suggested by discussants of Chakrabarti [2007], namely Lauritzen, Pericchi, Draper, Vehtari and others, which are still not explored. In the next subsection we content ourselves with revisiting a partly heuristic treatment of the high dimensional M-closed nested case, i.e. , with either M_1 or M_2 true. We provide a relatively simple proof of consistency under M_2 and M_1. A partly heuristic proof of consistency under M_1 is given in Chakrabarti [2007]. Generalizing this result we also suggest how the sample size r is to be allocated between stabilizing the posterior and model selection under more general assumptions than Chakrabarti [2007].

A very recent paper is Draper and Krnjajic [2010], who suggest the cross-validatory Bayes factors be used for model choice in the M-closed case to ensure stable inference. Though they report the results of a few simulations which are promising, one would need a much more extensive study of complex varying dimensions and different sizes before drawing firm conclusions.

We make a few tentative comments about their work. Draper and Krnjajic [2010] seem to be giving up the usual Bayes factors or replace them with cross-validatory Bayes factors. Also they seem to make simplistic assumptions about asymptotics in model selection. When n goes to infinity, p will usually tend to infinity, but not necessarily so. Moreover the rate of growth of p/n can vary a lot, leading to very different kinds of asymptotics. In particular if p tends to infinity sufficiently slowly, the results will be like those for fixed p and $n \to \infty$.

It will be interesting to compare the cross-validatory BF of Draper and Krnjajic [2010] and Intrinsic Bayes factor in the same problem. Also, our general view for fixed dimensional parameter and moderate n, is that the cross-validatory BF does not differ much from the usual BF. The following heuristics might clarify why this is likely. We consider $X_1, \ldots, X_n \sim N(\theta, 1)$ and competing models $M_1 : \theta = 0$ and

$M_2 : \theta > 0$, with a standard prior $N(\mu, \sigma^2)$ for θ. The leading term of the marginal under M_1, by Laplace approximation can be written as,

$$-\frac{1}{2}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

A heuristic cross validatory replacement of this will be

$$-\frac{1}{2}\sum_{j=1}^{n}(x_j - \bar{x}_{-j})^2, \qquad \bar{x}_{-j} = \sum_{i \neq j} x_i/(n-1).$$

The fact $\sum(x_j - \bar{x}_{-j})^2 = \sum(x_j - \bar{x})^2(\frac{n}{n-1})^2$ has the effect of reducing the marginal under alternative, making it conservative under $\theta = 0$. The leading term of cross-validatory BF increases a bit, suggesting it will work better. This has been confirmed by using cross-validatory BF and BF, with simulation studies. We have seen that under M_2 they differ by a very small value in log-scale, but that the small difference plays a critical role under M_1. Under M_1, $CVBF_{12}$ is greater than one for 80% cases choosing the correct model compared to none of them in case of BF_{12}, but we are looking at cases where the Bayes factor is very close to one.

We can justify the above claims for the cross-validatory Bayes factor of Draper and Krnjajic [2010], who replace the marginals in the Bayes factor by

$$\frac{1}{n}\sum_{i=1}^{n}\log p(y_i|y_{-i}, M_j). \tag{1.5}$$

Justification follows from a straight forward application of results in Mukhopadhyay et al. [2005]. Note that the above CVBF is the Bayes factor defined in the last reference with $s = n - 1$. Then the identity in [Mukhopadhyay et al., 2005, eq.11] reduces equation (1.5) to

$$-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n-1}{2}\log 2\pi - \frac{1}{2}\log n$$

The proof of the equation 11 in Mukhopadhyay et al. [2005] is given in the appendix of that paper.

1.3.2 How to choose the size of cross-validation: some preliminary results

Following Chakrabarti [2007], we will review the high-dimensional normal linear model setup as described in equation (1.3), with the same competetive models M_1 and M_2 described there. The study in Chakrabarti [2007] was done under the Zellner-Siow prior, but our results hold for any other priors under the following assumption. We assume that

Assumption 1. $\pi(\hat{\mu}_k) - \pi(\hat{\mu}_r) = o_p(1)$, as $k \to \infty$, $r \to \infty$, where $\pi(\hat{\mu}_{\mathbf{k}})$ and $\pi(\hat{\mu}_{\mathbf{r}})$ are the prior density evaluated at the mle of μ depending on k and r replicates.

For cross-validatory Bayes factor we use k out of r replicates for each μ_i to make the prior proper. A formal definition appears in equation (1.6) below. Here we will try to prove the consistency of a proxy to $CVBF_{21}$ obtained by using the popular KWP-Laplace (Kass-Wasserman-Pauler-Laplace) approximation for high dimensional problems. Under Assumption 1, below we apply this approximation to both BF_{21}^r and BF_{21}^k with p as dimension of parameter space and r and k as sample size.

$$\begin{aligned} \log CVBF_{21} &= \log \frac{BF_{21}^r}{BF_{21}^k} = \frac{p}{2}[(rC_p - kC_p') - \log \frac{r}{k}] + o_P(1) \\ &= \log CVBF_{21}^{ps} + o_P(1), \end{aligned} \tag{1.6}$$

where $C_p = \frac{1}{p}\sum_{i=1}^{p}[\frac{1}{r}\sum_{j=1}^{r} y_{ij}]^2$ and $C_p' = \frac{1}{p}\sum_{i=1}^{p}[\frac{1}{k}\sum_{j=1}^{k} y_{ij}]^2$. Here "ps" stands for pseudo.

We will prove posterior consistency under both models M_1 and M_2 for $CVBF_{21}^{ps}$. To prove the consistency under model M_2, we assume as in Chakrabarti [2007], the following.

Assumption 2. $\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \mu_i^2 = \tau^2 > 0$, under M_2.

Theorem 2 *Let p, k, $r \to \infty$. Under Assumptions 1 and 2, for both M_1 and M_2, $\log CVBF_{21}^{ps}$ chooses the correct model, with probability tending to one, for all $0 \le c < 1$, when $k/r \to c$.*

Proof Suppose first, model M_2 is true, then for fixed value of p,

$$\begin{aligned}\log CVBF_{21}^{ps} &= \frac{p}{2}[(rC_p - kC_p') + \log\frac{k}{r}] \\ &= \frac{p}{2}[(rC_p - kC_p') - \log\frac{rC_p}{kC_p'} + \log\frac{C_p}{C_p'}]\end{aligned}$$

The function $f(x) = x - \log x$ is increasing for $x > 1$. Using Assumption 2, when $k \to \infty, r \to \infty$, we have for sufficiently large p, r and k, $P(rC_p > 1) \to 1$, $P(kC_p' > 1) \to 1$. Then $\frac{rC_p}{kC_p'} > 1$ implies $(rC_p - kC_p') - \log\frac{rC_p}{kC_p'} > 0$. By definition of C_p and C_p', and also using Assumption A2, under $k \to \infty$ and $r \to \infty$

$$\lim_{k\to\infty, r\to\infty} \frac{p}{2}\log\frac{C_p}{C_p'} = \frac{p}{2}o_P(1)$$

which implies $\frac{C_p}{C_p'} = 1 + o_P(1)$. Hence, we obtain

$$\lim_{k\to\infty, r\to\infty} \frac{rC_p}{kC_p'} = \frac{1}{c}.$$

Now here $\frac{1}{c} > 1$ for any $0 \leq c < 1$, completing the proof of consistency under M_2.

We know,

$$\begin{aligned}&(rC_p - kC_p') \\ &\quad = (\frac{k^2}{r} - k)\frac{1}{p}\sum \bar{y}_i^{'2} + r(1-\frac{k}{r})^2\frac{1}{p}\sum \bar{y}_i^{''2} + 2k(1-\frac{k}{r})\frac{1}{p}\sum \bar{y}_i'\bar{y}_i'' \\ &\quad = (\frac{k}{r} - 1)\frac{1}{p}\sum(\sqrt{k}\bar{y}_i')^2 - (\frac{k}{r} - 1)\frac{1}{p}\sum(\sqrt{r-k}\bar{y}_i'')^2 \\ &\qquad\qquad +2\sqrt{\frac{k}{r}(1-\frac{k}{r})}\frac{1}{p}\sum\sqrt{k}\bar{y}_i'\sqrt{r-k}\bar{y}_i''\end{aligned}$$

Now assuming model M_1 is true and $r \to \infty$, $r - k \to \infty$, $\frac{k}{r} \to c$, for fixed value of p, we get,

$$\begin{aligned}&\lim_{r\to\infty, k\to\infty} rC_p - kC_p' \\ &\quad = (c-1)[\frac{1}{p}\sum W_j - \frac{1}{p}\sum U_j] + 2\sqrt{c(1-c)}\frac{1}{p}\sum w_i u_i,\end{aligned}$$

where $W_j \sim \chi_1^2$, $U_j \sim \chi_1^2$, $w_i \sim N(0,1)$ and $u_i \sim N(0,1)$ for $\forall\ 1 < j < p$. Hence,

$$\begin{aligned}&\lim_{k\to\infty, r\to\infty} \frac{p}{2}[(rC_p - kC'_p) - \log\frac{r}{k}]\\ &\quad= \frac{p}{2}[(c-1)[\frac{1}{p}\sum W_j - \frac{1}{p}\sum U_j] + 2\sqrt{c(1-c)}\frac{1}{p}\sum w_i u_i + \log c]\\ &\quad= \frac{p}{2}[o_p(1) + \log c] \sim \frac{p}{2}\log c < 0,\end{aligned}$$

for $0 \leq c < 1$, showing the consistency of $\log CVBF_{21}^{ps}$ under M_1. ■

Remark Extending the results proved in Chakrabarti [2007], we have shown the consistency of the cross-validatory Bayes factor under both M_1 and M_2 for $0 \leq c < 1$, when $\frac{k}{r} \to c$. Our proof is valid for any prior distribution satisfying the Assumption 1, which includes the Zellner-Siow prior used in Chakrabarti [2007] and many other commonly used. The proof also shows the smaller the value of c (i.e. smaller the k) we get a larger $CVBF_{21}$ under M_2 and a smaller $CVBF_{21}$ under M_1. This suggests one should have a relatively small value of c, $c = 0$ would be the best. This supports the choice of minimal training sample as in Berger and Pericchi [2004] but seems to contradict the conjecture of Chakrabarti [2007] that in high dimensional case c should tend to a positive constant. A possible explanation is that $r \to \infty$, $k \to \infty$ but p tends to infinity at a slower rate than k and $r - k$, this prevents it from becoming a real high-dimensional problem. Then further study is needed.

2. BAYES MODEL SELECTION WITH PATH SAMPLING: FACTOR MODELS AND OTHER EXAMPLES

2.1 Introduction

Advances in MCMC techniques to compute the posterior for many complex, hierarchical models have been a major reason for success in Bayes modeling and analysis of complex phenomena (Andrieu et al. [2004]). These techniques along with applications are surveyed in numerous papers, including Chen et al. [2000], Liu [2008] and Robert and Casella [2004]. Moreover, many Bayesian books on applications or theory and methods provide a quick introduction to MCMC such as Gelman et al. [2003], Ghosh et al. [2006], Gamerman and Lopes [2006] and Lynch [2007].

Just as the posterior for the parameters of a given model are important for calculating Bayes estimates, posterior variance, credibility intervals and a general description of the uncertainty involved, one needs to calculate Bayes Factors for selecting one of several models. Bayes factors are the ratio of marginals of given data under different models, when more than one model is involved and one wishes to choose one from among them, based on their relative or posterior probability. The ratio of marginals measures the relative posterior probability or credibility of one model with respect to the other if we make the usual objective choice of half as prior probability for each model.

Although there are many methods for calculating Bayes Factors, their success in handling complex modern models is far more limited than seems to be generally recognized. Part of the reason for lack of awareness of this is that model selection has become important relatively recently. Also one may think that, in principle, calculation of a BF can be reduced to the calculation of a posterior and hence solvable by the same methods as those for calculating the posterior. Reversible Jump MCMC

(RJMCMC) is an innovative methodology due to Green [1995], based on this simple fact. However, two models essentially lead to two different sets of states for any Markov chain that connects them. The state spaces for different models often differ widely in their dimension. This may prevent good mixing and may show up in the known difficulties of tuning RJMCMC. For a discussion of tuning difficulties see Robert and Casella [2004].

An other popular method for calculating BF is path sampling (PS), which is due to Gelman and Meng [1998], and recently re-examined by Lefebvre et al. [2009]. Our major goal is to explore PS further in the context of nested, relatively high dimensional covariance models, rather than non-nested low dimensional mean models, as in the last reference. The new examples show both similarities and sharp changes from the sort of behavior documented in Lefebvre et al. [2009].

We consider three paths, namely, the geometric mean path, the arithmetic mean path and the parametric arithmetic mean path, which appear in Gelman and Meng [1998], Lefebvre et al. [2009], Ghosh and Dunson [2008], Ghosh and Dunson [2009], Lee and Song [2002] and Song and Lee [2006]. Other applications of Path Sampling and Bridge Sampling(with some modifications) appear in Lartillot and Philippe [2006], Friel and Pettitt [2008], Xie et al. [2011] and Fan et al. [2011]. Our priors are usually the diffuse Cauchy priors, first suggested by Jeffreys [1961] and since then recommeded by many others, including Berger (personal communication), Liang et al. [2008], Gelman [2006] and Ghosh and Dunson [2009]. But we also examine other less diffuse priors too, going all the way to normal priors. Since Lefebvre et al. [2009] have studied applications of PS to mean like parameters, we focus on covariance models. We restrict ourselves generally to factor models for covariance, which have become quite popular in recent applications, e.g. Lopes and West [2004], Ghosh and Dunson [2008], Ghosh and Dunson [2009] and Lee and Song [2002]. The recent popularity of factor models is due to the relative ease with which they may be used to provide a sparse representation of the covariance matrix of multivariate normal data in many applied problems of finance, psychometry and epidemiology, see for example the last three references. Also, often it leads to interesting scientific insight, see Bartholomew et al. [2002].

In addition to prior, likelihood and path, there are other choices to be made before PS can be implemented, namely, a method of discretizing the path, e.g., by equispaced points or adaptively (Lefebvre et al. [2009]) and how to integrate the score function of Gelman and Meng [1998] at each point in the discrete path. A popular method is to use MCMC. These more technical choices are discussed later in the paper. Along with PS we will consider other methods like Importance Sampling (IS) and its descendents like Annealed Importance Sampling (AIS), due to Neal [2001], and Bridge Sampling (BS), due to Meng and Wong [1996].

We now summarize our contribution in this paper.

In Section 2.2 we review what is known about Path Sampling and Factor Models. We introduce Factor Models, a suitable path and suitable diffuse t-priors. The path we use was first introduced in Gelman and Meng [1998] for mean models and by Lee and Song [2002] and Ghosh and Dunson [2009] for Factor Models.

In Subsection 2.2.4 we prove a theorem (Theorem 3) which essentially shows that except for the convergence of MCMC estimates for expected score function $E_t(U(\theta, t))$ at each grid point t in the path, all other needed conditions for PS will hold for our chosen path, prior and likelihood for Factor Models. In one of the Remarks following the Theorem we generalize this result to other paths. Remark 3 points to the need for some finite moments for the prior not just for Theorem 3 to hold but for the posterior to behave well. Then in Remark 5 we provide a detailed, heuristic argument as to why the MCMC may fail dramatically by not mixing properly if the data has come from the bigger of the two models under consideration. If our heuristics is correct, and there is a small interval where $E_t(U(\theta, t))$ oscillates most, then a grid size that is a bit coarse will not only be a bit inaccurate, it will be very wrong. Even if the grid size is sufficiently small, one will need to do MCMC several times with different starting points just to realize PS will not work. Our new proposal avoids these problems but will require more time if many models are involved.

In Section 2.3, we give an argument as to why the above is unlikely to be true if the data has come from the smaller model. More importantly, in Subsection 2.3.3 we propose a modification of PS, which we call Path Sampling with Small Change (PS-SC) which is expected to do better.

Implementation of PS and PS-SC can be very time consuming due to the need of MCMC sampling for each grid point along the path. Time can be saved if we can implement PS and PS-SC by parallel computation, as noted by Gelman and Meng [1998].

In Subsection 2.3.4 we show MCMC output for the various cases discussed and validate our heuristics above. The diagnostics via projection into likelihood space should prove useful for other model selection problems. Our gold standard is PS-SC, based on an MCMC with the number of draws m=50,000 and burn in of 1,000, if necessary. But actually in our examples m=6,000 and burn in of 1000 suffice for PS-SC. For other model selection rules also we go up to m=50,000 if necessary. After Subsection 2.3.4, having shown our modified PS, namely PS-SC, is superior to PS under both models, we do not consider PS in the rest of the paper.

In the last two Sections we touch on the following related topics: effects of grid size, alternative path, alternative methods and performance of PS-SC and some other methods in very high dimensional simulated and real examples. PS-SC seems to choose the true models in the simulated cases and relatively conservative models for real data. In Section 2.5 we explore various real life and high dimensional factor models, with the object of combining PS-SC with two of the methods which do relatively well in Section 2.4 to reduce the time of PS-SC in problems with the number of factors rather high say 20 or 26, for which PS-SC can be quite slow. For these high dimensional examples, we use Laplace approximation to marginals for preliminary screening of models. A few general comments on Laplace approximation in high dimensional problems are in Section 2.5.

In Appendix A we introduce briefly a few other methods like Annealed Importance Sampling (AIS) which we have compared with PS-SC. Finally, Appendix B points to some striking differences between what we observe in Factor Models and what one might have expected from our familiar classical asymptotics for maximum likelihood estimates. Of course, as pointed out by Drton [2009], classical asymptotics does not apply here but it surprised us that the differences would be so stark. It is interesting and worth pointing out that the Bayes methods like PS-SC can be validated partly theoretically and partly numerically inspite of a lack of suitable asymptotic theory.

2.2 Path sampling and Factor Models

In the following subsections, we review some basic facts about PS, including definition of the three paths and the notion of an optimal path. More importantly, since our interest would be in model selection for covariance rather than mean, we introduce Factor models and then PS for Factor Models in Subsection 2.2.3 and 2.2.4.

Subsection 2.2.1 is mostly an introduction to PS and reviews previous work. After that we show the failure of PS-estimates in a toy problem related to the modelling of the covarince matrix in Subsection 2.2.2. In Subsection 2.2.3 we introduce Factor Models and our priors. Subsection 2.2.4 introduces paths that we consider for Factor Models and a Theorem showing the regularity conditions needed for validity of PS under Factor Models. Then in a series of remarks we extend the theorem and also study and explain how the remaining ingredient of PS, namely MCMC, can go wrong. We show a few MCMC outputs to support our arguments in Subsection 2.3.4. This particular theme is very important and will come up several times in later sections or subsections where related different aspects will be presented.

2.2.1 Path Sampling

Among the many different methods related to importance sampling, the most popular is path sampling (PS). However PS is best understood as a limit of the simpler bridge sampling (BS) (Gelman and Meng [1998]). So we first begin with BS.

It is well-known that unless the densities of the sampling and target distributions are close in relative importance sampling weights, Importance Sampling (IS) will have high variance as well as high bias. Due to the difficulty of finding a suitable sampling distribution for IS, one may try to reduce the difficulty by introducing a non-normalized intermediate density $f_{1/2}$ that acts like a bridge between the non-normalized sampling density f_1 and non-normalized target density f_0 (Meng and Wong [1996]). One can then use the identity $Z_1/Z_0 = \frac{Z_{1/2}/Z_0}{Z_{1/2}/Z_1}$ and estimate both the numerator and denominator by IS. Extending this idea, Gelman and Meng [1998] constructed a whole path $f_t : t \in [0, 1]$ connecting f_0 and f_1. This is also like a bridge. Discretizing this they get the identity $Z_1/Z_0 = \prod_{l=1}^{L} \frac{Z_{(l-1/2)}/Z_{(l-1)}}{Z_{(l-1/2)}/Z_{(l)}}$, which leads to a chain of IS estimates in the numerator and denominator. We call this estimate the Generalized Bridge Sampling (GBS) estimate.

More importantly, Gelman and Meng [1998] introduced PS, which is a new scheme, using the idea of a continuous version of GBS but using the log scale. The PS estimate is calculated by first constructing a path as in BS. Suppose the path is given by $p_t : t \in [0, 1]$ where for each t, p_t is a probability density. Then we have the following definition.

$$p_t(\theta) = \frac{1}{z_t} f_t(\theta), \tag{2.1}$$

where f_t is an unnormalized density and $z_t = \int f_t(\theta)d\theta$ is the normalizing constant. Taking the derivative of the logarithm on both sides, we obtain the following identity under the assumption of interchangeability of the order of integration and differentiation:

$$\frac{d}{dt} log(z_t) = \int \frac{1}{z_t} \frac{d}{dt} f_t(\theta)\mu(d\theta) = E_t[\frac{d}{dt} \log f_t(\theta)] = E_t[U(\theta, t)] \tag{2.2}$$

where the expectation E_t is taken with respect to $p_t(\theta)$ and $U(\theta, t) = \frac{d}{dt} \log f_t(\theta)$. Now integrating (2.2) from 0 to 1 gives the log of the ratio of the normalizing constants, i.e. log BF in the context of model selection:

$$log[\frac{Z_1}{Z_0}] = \int_0^1 E_t[U(\theta, t)]dt \tag{2.3}$$

To approximate the integral we discretize the path with k points $t_{(0)} = 0 < t_{(1)} < \ldots < t_{(k)} = 1$ and draw m MCMC samples converging to $p_t(\theta)$ at each of these k points. Then estimate $E_t[U(\theta, t)]$ by $\frac{1}{m} \sum U(\theta^{(i)}, t)$ where $\theta^{(i)}$ is the MCMC output. To estimate the final log Bayes factor, commonly numerical integration schemes are used. It is clear that MCMC at different points "t" on the path can be done in parallel. We have used this both for PS and for our modification of it, namely PS-SC introduced in Subsection 2.3.3.

Gelman and Meng [1998] showed there is an optimum path in the whole distribution space providing a lower bound for MCMC variance, namely

$$[\arctan \frac{H(f_0, f_1)}{\sqrt{4 - H^2(f_0, f_1)}}]^2 / m$$

where f_0 and f_1 are the densities corresponding the two models compared and $H(f_0, f_1)$ is their Hellinger distance. Unfortunately in nested examples f_0 and f_1 are mutually orthogonal, so $H(f_0, f_1)$ takes the trivial value of two. Moreover m is so large that the lower bound becomes trivial and unattainable. However in a given problem, one path may be more suitable or convenient than another.

Following Gelman and Meng [1998] and Lefebvre et al. [2009], we consider three paths generally used for the implementation of PS. Geometric Mean Path (GMP) and Arithmetic Mean Path (AMP) are defined by the mean ($f_t = f_0^{(1-t)} f_1^t$ and $f_t = t f_0 + (1 - t) f_1$ respectively) of the densities of two competing models for each model $M_t : t \in (0, 1)$ along the path. Our notation for Bayes Factor is given later in Equation 2.6.

One more common path is obtained by assuming a specific functional form f_θ for the density and then constructing the path in the parametric space ($\theta \in \Theta$) of the assumed density. If $\theta_t = t\theta_0 + (1 - t)\theta_1$, then f_{t,θ_t} is the density of the model M_t, where $f_{0,\theta_0} = f_0$ and $f_{1,\theta_1} = f_1$. We denote this third path as Parametric Arithmetic Mean Path (PAMP). The PAMP path was shown by Gelman and Meng [1998] to minimize the Rao distance in a path for model selection about normal means. More importantly it is very convenient for use of MCMC, as shown for some Factor models by Song and Lee [2006] and Ghosh and Dunson [2009], and for linear models by Lefebvre et al. [2009]. Implementation of PS with the paths mentioned above are denoted as GMP-PS, AMP-PS and PAMP-PS. In view of the discussion in Lefebvre et al. [2009] regarding the degeneracy of the AMP-PS, we will only consider PAMP-PS and GMP-PS.

Unlike Lefebvre et al. [2009], who study models about means, our interest is in studying model selection for covariance models, specifically Factor Models with different number of factors. These are discussed in the Subsections 2.2.3 and 2.2.4. Performance of PS for covariance models can be very different from the examples in Lefebvre et al. [2009]. In the next subsection we give a toy example of covariance model selection where PS fails and our proposed modification PS-SC is also not applicable.

2.2.2 Covariance Model : Toy Example

To illustrate the difficulties in calculation of the BF that we discuss later, we begin by considering a problem where we can calculate the true value of the Bayes Factor.

We assume $Y_p \sim N(0, \Sigma)$, for some $m < p$ we wish to test whether $Y_{1,\dots,m}$ and $Y_{m+1,\dots,p}$ are independent or not. If $\Sigma = \begin{pmatrix} A_{11} & A_{12} \\ A'_{12} & A_{22} \end{pmatrix}$ where $Y_{1,\dots,m} \sim N(0, A_{11})$ and $Y_{m+1,\dots,p} \sim N(0, A_{22})$, then the competetive models for a fixed m will be, $M_0 : A_{12} = 0$ vs $M_1 : A_{12} \neq 0$. Under M_1 we use a inverse-Wishart prior for the covariance matrix as it helps us to calculate the true BF, using the conjugacy property of the prior. Under M_0 we take A_{11}, A_{22} to be independent, each with a inverse Wishart prior.

We illustrate the above problem with p=10 and m=7 for a positive definite matrix $\Sigma^0 = \begin{pmatrix} A^0_{11} & A^0_{12} \\ (A^0_{12})' & A^0_{22} \end{pmatrix}$ (given in Appendix C). We implement the Path Sampling for this problem connecting M_0 and M_1, using a Parametric Arithmetic Mean Path :

$$M_t : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A^0_{11} & tA^0_{12} \\ t(A^0_{12})' & A^0_{22} \end{pmatrix}\right) \tag{2.4}$$

For every $0 \leq t \leq 1$, the Σ matrix is positive definite, being a convex combination of two positive definite matrices. For t=0 and t=1 we get the models M_0 and M_1.

We can estimate the Bayes Factor by using the Path Sampling schemes as described earlier. We simulated two datasets one each from M_0 and M_1 and report the true BF value with the PS estimate in Table 2.1. Here the reported Bayes Factor is defined as the ratio $\frac{m_1}{m_0}$, where m_1 and m_0 are the marginals under the models M_1 and M_0 respectively.

Table 2.1
Performance of PS in Toy Example modelling Covariance : Log Bayes Factor (MCMC-standard deviation)

Method	Data 1	Data 2
True BF value	258.38	-132.87
PS estimate of BF	184.59 (.012)	-20.11 (.008)

The values in Table 2.1 show us that estimated BF value is off by an order of magnitude when M_0 is true. The value is relatively stable as judged by the MCMC-standard deviation based on 10 runs and near to the true value for M_1.

2.2.3 Factor Models and Bayesian Specification of Prior

A factor model, with k factors is defined as n i.i.d. observed r.v.'s

$$y_i = \Lambda \eta_i + \epsilon_i, \epsilon_i \sim N_p(0, \Sigma),$$

where Λ is a $p \times k$ matrix of factor loadings,

$$\eta_i = (\eta_{i1}, \ldots, \eta_{ik})' \sim N_k(0, I_k)$$

is a vector of standard normal latent factors, and ϵ_i is the residual with diagonal covariance matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_p^2)$. Thus we may write the marginal distribution of y_i as $N_p(0, \Omega)$, $\Omega = \Lambda\Lambda' + \Sigma$. This model implies that the sharing of common latent factors explains the dependence in the outcomes and the outcome variables are uncorrelated given the latent factors.

A factor model, without any other constraints, is non-identifiable under orthogonal rotation. Post-multiplying Λ by an orthogonal matrix P, where P is such that $PP' = I_k$, we obtain exactly the same Ω as in the previous factor model. To avoid this, it is customary to assume that Λ has a full-rank lower triangular structure, restricting the number of free parameters in Λ and Σ to $q = p(k+1) - k(k-1)/2$, where k must be chosen so that $q \leq p(p+1)/2$. The reciprocal of diagonal entries of Σ form the precision vector here.

It is well-known that maximum likelihood estimates for parameters in factor models may lie on boundaries and hence likelihood equations may not hold (Anderson [1984]). The Bayes estimate of Ω defined as average over MCMC outputs is well-defined, easy to calculate and, being average of positive definite matrices, is easily seen to be positive definite. This fact is used to search for maximum likelihood estimate (mle) or maximum prior×likelihood estimates (mple) in a neighborhood of the Bayes estimate.

We also note for later use the following well-known simple fact, e.g. Anderson [1984]. If the likelihood is maximized over all positive definite matrices Ω, not just over factor models, then the global maximum for n independent observations exists and is given by

$$\hat{\Omega} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})'. \tag{2.5}$$

From the Bayes model selection perspective, a specification of the prior distribution for the free elements of Λ and Σ is needed. Truncated normal priors for the diagonal elements of Λ, normal priors for the lower triangular elements, and inverse-gamma priors for $\sigma_1^2, \ldots, \sigma_p^2$, have been commonly used in practice due to conjugacy and the resulting simplification in posterior distribution. Prior elicitation is not common.

Ghosh and Dunson [2009] addressed the above problems by using the idea of Gelman [2006] to introduce a new class of default priors for the factor loadings that have good mixing properties. They used Gibbs sampling scheme and showed there was good mixing and convergence. They use parameter expansion to induce a class of t or folded-t priors depending on sign constraints on the loadings. Suitable t-priors have been very popular. We use the same family of priors but consider a whole range of many degrees of freedom going all the way to the normal and use the same Gibbs sampler as in Ghosh and Dunson [2008]. We have used a modified version of their code.

In the factor model framework, we stick to the convention of denoting the Bayes factor for two models with latent factors $h-1$ and h as

$$BF_{h,h-1} = \frac{m_h(x)}{m_{h-1}(x)} \tag{2.6}$$

where $m_h(x)$ is the marginal under the model having h latent factors. *So the Bayes Factor for simpler model (defined as M_0) and complex model (defined as M_1) with $h-1$ and h latent factors will be defined as $BF_{h,h-1}$*. We choose the model with h and $h-1$ latent factors respectively, depending on the value of log Bayes factor being positive and negative. Alternatively one may choose a model only when the value of $\log BF$ is decisively negative or positive; say less than or greater than a chosen threshold.

2.2.4 Path Sampling for Factor Models

There are several variants of Path Sampling, which have been explored in different setups, depending on choice of path, prior and other tuning parameters (grid size and MCMC sample size). In the Factor model setup the parametric arithmetic mean path (PAMP) (used by Song and Lee [2006] and Ghosh and Dunson [2009]) seems to be the most popular one. We also consider Geometric Mean Path (GMP) along with the PAMP for Factor Model.

By constructing a GM path from corresponding prior to the posterior, we can estimate the value of the log-marginal under both M_0 and M_1, which in turn leads us to an estimate of the log-BF. The Gibbs Sampling scheme for these paths with parameter-expansion (for different priors) is described in the Appendix. We will first describe the two paths and their corresponding score functions to be estimated along the path.

i. **Parametric Arithmetic Mean Path** : Lee and Song [2002] used this path in factor models, following an example in Gelman and Meng [1998]. Ghosh and Dunson [2008] also used this path along with parameter expansion. Here we define M_0 and M_1 to be the two models corresponding to the factor model with factors $h-1$ and h, respectively and then connect them by the path $M_t : y_i = \Lambda_t \eta_i + \epsilon_i, \Lambda_t = (\lambda_1, \lambda_2, \ldots, \lambda_{h-1}, t\lambda_h)$ where λ_j is the j-th column of the loading matrix. So for t=0 and t=1 we get the models M_0 and M_1. The likelihood function at grid point t is a MVN which is denoted as $f(Y|\Lambda, \Sigma, \eta, t)$. We have independent priors $\pi(\Lambda), \pi(\Sigma), \pi(\eta)$ and a score function,

$$U(\Lambda, \Sigma, \eta, Y, t) = \sum_{i=1}^{n} (y_i - \Lambda_t \eta_i)' \Sigma^{-1} (0^{p\times(h-1)}, \lambda_h) \eta_i. \tag{2.7}$$

For fixed and ordered grid points along the path $t_{(0)} = 0 < t_{(1)} < \ldots < t_{(S)} < t_{(S+1)} = 1$, our path sampling estimate for the log Bayes factor is

$$log(\widehat{BF}_{h:h-1}) = \frac{1}{2} \sum_{s=0}^{S} (t_{s+1} - t_s)(\widehat{E}_{s+1}(U) + \widehat{E}_s(U)). \tag{2.8}$$

We simulate m samples of $(\Lambda_{t_s,i}, \Sigma_i, \eta_i : i = 1, \ldots, m)$ from the posterior distribution of $(\Lambda_{t_s}, \Sigma, \eta)$ at the point $0 \leq t_s \leq 1$ and use them to estimate $\widehat{E}_s(U) = \frac{1}{m} \sum U(\Lambda_{t_s,i}, \Sigma_i, \eta_i, y), \;\; \forall s = 1, \ldots, S+1$.

ii. **Geometric Mean Path** : This path is constructed over the distributional space (Gelman and Meng [1998]), hence we model the density for the model M_t at each point along the grid. We use the density $f_t(\Lambda, \Sigma, \eta | Y) = f(y|\Lambda, \Sigma, \eta)^t \pi(\Lambda, \Sigma, \eta)$

as the unnormalized density for the model M_t connecting the prior and the posterior, when $\pi(\Lambda, \Sigma, \eta)$ and $f(y|\Lambda, \Sigma, \eta)$ are prior and the likelihood function correspondingly. By using PS along this path we can find the log marginal for the models M_0 and M_1, as the normalizing constant for the prior is known. Hence the $\log BF$ can be estimated by using those estimates of the log marginal for those models. The score function $U(\Lambda, \Sigma, \eta, Y, t)$ will be the loglikelihood function $\log f(y|\Lambda, \Sigma, \eta)$.

The theorem below verifies the regularity conditions of Path Sampling for Factor Models. For PS to succeed we also need convergence of MCMC at each point in the path. That will be taken up after proving the theorem.

Theorem 3 *Consider Path Sampling for factor models with parametric arithmetic mean path (PAMP) and likelihood as given above for factor models. Assume prior is proper and the corresponding score function is integrable w.r.t. the prior,*

1. *The interchangeability of integration and differentiation in (2.2) is valid.*
2. *$E_t(U)$ is finite as $t \to 0$.*
3. *The path sampling integral for factor models, namely (2.3), is finite.*

Proof 1. Here for notational convenience, we write $(\Lambda, \Sigma, \eta) = \theta$. When $f(Y|\theta)$ and $\pi(\theta)$ are the likelihood function of the data and the prior density function for the corresponding parameter respectively, then the following is equivalent to showing eqn (2.2).

$$\frac{d}{dt}\int_{-\infty}^{\infty} f(Y|\theta, t)\pi(\theta)d\theta = \int_{-\infty}^{\infty} \frac{d}{dt} f(Y|\theta, t)\pi(\theta)d\theta$$

We can write the LHS as following,

$$\begin{aligned}
&= \lim_{\delta \to 0} \int_{-\infty}^{\infty} \frac{f(Y|\theta, t+\delta) - f(Y|\theta, t)}{\delta}\pi(\theta)d\theta \\
&= \lim_{\delta \to 0} \int_{-\infty}^{\infty} f'(Y|\theta, t')\pi(\theta)d\theta \ , t' \in [t, t+\delta] \\
&= \lim_{\delta \to 0} \int_{-\infty}^{\infty} U(Y|\theta, t')f(Y|\theta, t')\pi(\theta)d\theta
\end{aligned}$$

where $t' \in [t, t+\delta]$. U is a quadratic function in θ, and hence its absolute value is bounded above by a quadratic function in θ, free of t but depending on Y. $f(Y|\theta, t')$ is bounded by the global maximum of the MVN likelihood, say M, achieved at $\widehat{\Omega}$ (eqn 2.6). Now applying the moment assumptions for $\pi(\theta)$ we can use Dominated Convergence Theorem (DCT) and take the limit within the integral sign.

The rest of the statements, namely 2 & 3 follow similarly.

■

In Remark 1, we extend the Theorem 3 to other paths. Then in a series of Remarks we study various aspects like convergence and divergence of PS, that are closely related to the Theorem. All the remarks are related to the Theorem and insights gained from its proof. Remark 5 is the most important.

Remark 1 For PS with GMP, the score function is the loglikelihood function which can be bounded as before by using the RHS of equation (2.5). Also, $f(y|\Lambda, \Sigma, \eta)^t \leq (1 \vee f(y|\hat{\Omega}))$ with $\hat{\Omega}$ as in equation (2.5). We believe a similar generalization holds for most paths modeling means of two models. Now the proof of Theorem 3 applies exactly as before (i.e. as for PAMP). We exhibit performance of PS for this path in Section 2.4.

Remark 2 If we further assume the MCMC average at each point on the grid converges to the Expectation of score function of MCMC then the Theorem implies the convergence of PS. We showed the integrand is continuous on [0, 1]. So by continuity it can be approximated by a finite sum. Now take limit of MCMC average at each of these finitely many grid points. However, even if the MCMC converges in theory, the rate of convergence may be very slow or there may be a problem with mixing even for m=50,000, which we have taken as our gold standard for good MCMC. This problem will be apparent to some extent from high MCMC standard deviation.

Remark 3 As $t \to 0$ the likelihood is practically independent of the extra parameters of the bigger model, so that a prior for those parameters (conditional on other parameters) will not learn much from data. In particular, the posterior for these parameters will remain close to the diffuse prior one normally starts with. If the prior fails to have the required finite moment in the theorem, the posterior will also be likely to have large values for moments, which may cause convergence problems for the MCMC. That's why we chose a prior making the score function integrable. In the proof of the Thoerem, we have assumed the first two moments of the prior to be finite. In most numerical work our prior is a t with 5 to 10 d.f.

Remark 4 In the same vein we suggest that even when the integral at t near zero converges, the convergence may be slow for the following reason. Consider a fixed $(\Lambda_t, \Sigma, \eta)$ with a large posterior or negative value of $U(\Lambda_t, \Sigma, \eta)L(\Lambda_t, \Sigma, \eta)$ at point t, the same large value will occur at $(\frac{1}{t}\Lambda_t, \Sigma, \eta)$ with prior weight $\pi(\frac{1}{t}\Lambda_t, \Sigma, \eta)$. For priors like t-distribution with low degrees of freedom, $\pi(\frac{1}{t}\Lambda_t, \Sigma, \eta)$ will not decay rapidly enough to substantially diminish the contribution of the large value of $U(\Lambda_t, \Sigma, \eta)L(\Lambda_t, \Sigma, \eta)$ at $(\Lambda_t, \Sigma, \eta)$.

Remark 5 The structure of the likelihood and prior actually provides insight as to when the MCMC will not converge to the right distribution owing to bad mixing. To this end we sketch a heuristic argument below, which will be supported in Subsection 2.3.4 by MCMC figures.

1. The maximized likelihood remains the same along the whole path, because the path makes an one to one transformation of the parameter space as given below.

2. If the MLE of λ_h at t=1 is $\hat{\lambda}_h$, then MLE at $t = t'$ is $\frac{\hat{\lambda}_h}{t'}$ (subject to variation due to MCMC at two different points at the path), which goes to infinity as t goes to zero. This happens as the $\hat{\lambda}_h$ remains the vector among λ'_h (where λ'_h is the MCMC sample from model M_t at t) having the highest maximum likelihood. Hence as $t \to 0$, $\pi(\hat{\lambda}_h/t) \to 0$ at a rate determined by the tail of the prior. The conflict between prior and maximized likelihood may also be viewed as a conflict between the nested models, with the prior favoring the

parsimonious smaller model. This inherent conflict in model selection seems to have the following implications for MCMC.

We expect to see a range (say $[t_1, t_2]$) near zero showing a conflict between prior and maximized likelihood. Definitely the points t_1 and t_2 are not well-specified, but we treat them as such so as to understand some underlying issues of mixing and convergence here. On the set of points $t > t_2$ the MCMC samples are expected to be around the points maximizing likelihood, whereas for $t < t_1$ they will be nearly zero due to the concentration around a value λ_h which is both prior mode and the mle under M_0, namely $\lambda_h = 0$. But for any point in the range $[t_1, t_2]$, they will span a huge part of the parameter space, ranging from points maximizing likelihood to ones having higher prior probability, showing a lot of fluctuations from MCMC to MCMC. The MCMC outputs in Subsection 2.3.4 show both clusters but having highly fluctuating values (Figure 4, Subsection 2.3.4) for the proportions of the clusters.

Equation (2.7) tells us that the score function is proportional to $\frac{\lambda_h'}{t}$ (where λ_h' is the MCMC sample from model M_t at t). Hence we will see $E_t(U)$ as an increasing function while $t \to t_2$ from the right hand side ((2) in Remark 5). This leads to a lot of variation of the estimate of $E_t(U)$ for different MCMC samples in the range $[t_1, t_2]$ as explained above. Also, as explained above, for $t < t_1$, the score function will concentrate near zero.

The width of the zone of conflict (here $t_2 - t_1$) will shrink, if we have a relatively strong decaying tail of the prior. On the other hand for heavy-tailed priors we may see these above mentioned fluctuations for a longer range, causing a loss of mass from the final integration. These problems are aggravated by the high-dimension of the problem and the diffuse spread of the prior on the high-dimensional space. This may mean the usual BF estimated by PS will be off by an order of magnitude. We will see the implications reflected in some figures and tables in the next Section, when we study PAMP-PS for Factor Models in detail in Section 2.3.

Remark 6 We have checked that adaptive choice of grid points by Lefebvre et al. [2009], which improves accuracy in their two examples with GMP, doesn't help in the case of the very large fluctuations described above. It seems to us that adaptive choice would work better when the two models tested are less dissimilar than the models in Remark 5, e.g., when the smaller of two nested models is true (Subsection 2.3.1) or when our proposed modification of PS is used (Subsection 2.3.3). However, we have not verified this because even without adaptive choice, our new proposal worked quite well in our examples.

We note in passing that in both the examples of Lefebvre et al. [2009], the two models being tested have maximum likelihoods that differ by fifteen in the log scale, whereas for the models in Remark 5 they differ by much more, over hundred.

2.3 What do actual computations tell us?

Following the discussion in previous Section, we would like to study the effects of the theoretical observations in the previous Section for the implementation of Path Sampling. Here we only consider the PAMP for PS, and for notational convenience we will mention it as just PS. After studying estimated BF's in several simulated data sets (not reported here) from various factor models, we note a few salient features. Error in estimation of the BF or the discrepancy between different methods tends to be relatively large, if one of the following is true : the data has come from the complex model rather than the simpler model, the prior is relatively diffuse or the value of the precision parameters are relatively small. Different subsections study what happens if the complex or simpler model is true, the effect of the prior, the grid size and MCMC size. These are done in Subsections 2.3.1-2.3.2.

In Subsection 2.3.3, we introduce a new PS scheme, which operates through a chain of paths, each path involving two nested models with a small change between the contiguous pairs. The new scheme is denoted as Path sampling with Small Changes (PS-SC). The effect of precision parameters will also be studied in this subsection for PS-SC. Then we study the MCMC samples and try to understand their behavior from the point of view of explaining the discrepancy between different methods for estimating Bayes Factors and why PS-SC does better than PS in Subsection 2.3.4.

Our simulated data are similar to those of Ghosh and Dunson [2009] but have different parameters. *We use a 2-factor model and 1-factor model as our complex model M_1 and simpler model M_0 correspondingly to demonstrate the underlying issues.* The loading parameters and the diagonal entries of the Σ matrix are given in Table 2.2 & 2.3. In simulation we take model M_0 or M_1 as true but Σ is not changed. Of course if the one factor model M_0 is true, then since it is nested in M_1, M_1 is also true.

Table 2.2
Loading Factors used for simulation

Factor 1	.89	0	.25	0	.8	0	.5
Factor 2	0	.9	.25	.4	0	.5	0

Table 2.3
Diagonal Entries of Σ

.2079	.19	.15	.2	.36	.1875	.1875

2.3.1 Issues in Complex (2-factor) model

We will study the effect of grid size, prior and the behavior of MCMC, keeping in mind Theorem 3 and the Remarks in Section 2.2. For Path Sampling with PAM path, we now discuss the effect of prior and the two tuning parameters, namely, the effect of the grid size and MCMC size, on the estimated value of the BF and their standard deviation. Following the discussion in Remarks 3 & 4, we know that $\lim_{t\to 0} E_t(U)$ is finite and path sampling converges under some finite moment assumption for the prior. The prior considered in PS by Ghosh and Dunson [2008] are Cauchy and half-Cauchy which do not have any finite moments and so U is not integrable. We therefore choose a relatively diffuse prior, but with enough finite moments for U. For finite mean and variance one needs a t with at least four degrees of freedom. Our favorites are t-distributions with 5 to 10 degrees of freedom. We show results for 5 and 10 d.f. only. But we first explore the sensitivity of the estimate to changes in d.f. of the t-distribution as prior, over a range of 1 through 90. The BF values change considerably until we reach about 40 d.f. and then it stabilizes. In Table 2.4 we report the $\log BF$ values estimated for 5 datasets simulated from 2-factor model using different priors. The behavior of the estimated $\log BF$ with the change of d.f. continuously from 1 to 100 is shown in the figure 1 for 3rd dataset.

Table 2.4
PAM-PS: Dependance of logBF_{21} over prior, 2-factor model true.

PS using grid size .01				
t_1	t_5	t_{10}	t_{90}	normal
2.62	14.42	22.45	70.20	70.25
3.67	11.90	21.39	68.70	68.72
3.00	13.43	21.31	47.06	47.21
4.29	13.17	18.49	48.03	48.13
4.20	13.11	18.48	47.70	47.74

We can see the estimate of the BF changing with the change in the pattern of the tail of the prior. The effect of the grid size and MCMC size on MCMC-standard deviation of the estimate are studied, using priors t_{10} and N(0,1) and reported in Table 2.5. We report mean of the estimates found from 25 different MCMC runs and the corresponding standard deviation as MCMC-standard deviation. The study has been done on the 2nd of the 5 datasets simulated from Model 1 earlier.

As expected Table 2.5 shows a major increase of MCMC-size and finer grid-size reduces the MCMC-standard deviation of the estimator. The difference between the

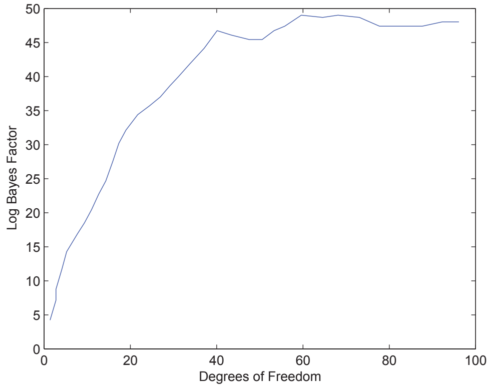


Figure 2.1. Dependance of logBF_{21} over prior for 3rd Data set.

Table 2.5
PAM-PS: Dependence of $\log BF_{21}$ (MCMC-standard deviation) estimates over grid size and MCMC size, 2-factor model true

Grid Size		.01	.001
MCMC-size	Prior	Data 2	Data 2
5000	t_{10}	21.26 (1.39)	21.26 (1.29)
	N(0,1)	66.89 (4.15)	67.21 (3.28)
50000	t_{10}	23.71 (1.21)	23.57 (.52)
	N(0,1)	68.21 (3.62)	68.23 (3.11)

mean values of BF estimated by t_{10} and N(0,1) differ by an order of magnitude. We will study these issues as well as special patterns exhibiting MCMC in Subsection 2.3.4. Though the different variants of PS compared here differ in their estimated value of BF, they still choose the correct model 100% of the time.

2.3.2 Issues in Simpler (1-factor) model

Now we study the scenario when the 1-factor model is true focusing on the effect of prior, grid-size and MCMC-size on the estimated Bayes Factor (Table 2.6). In this scenario the estimates don't change much with the change of prior, so we will report the estimates for prior t_{10} and N(0,1) with different values of MCMC-size and grid-size.

Table 2.6
PAM-PS: Dependence of $\log BF_{21}$ (MCMC-standard deviation) estimates over grid size and MCMC size, while 1-factor model true

Grid Size		.01	.001
MCMC-size	Prior	Data 1	Data 1
5000	t_{10}	-4.26 (.054)	-4.27 (.044)
	N(0,1)	-4.62 (.052)	-4.60 (.051)
50000	t_{10}	-4.24 (.012)	-4.24 (.007)
	N(0,1)	-4.60 (.006)	-4.62 (.005)

This table shows us that the MCMC-standard deviation improves with the finer grid-size and large MCMC-size as expected, but the estimated values of BF_{21} remain mostly same. As noted earlier, PS chooses the correct model 100% of the time when M_0 is true.

We explain tentatively why the calculation of BF is relatively stable when the lower dim model M_0 is true. Since M_0 is nested in M_1, M_1 is also true in this case, which in turn implies both max likelihoods (under M_0 & M_1) are similar and smaller than for data coming from M_1 true (but not M_0). This tends to reduce or at least is associated with the reduction of the conflict between the two models or prior and likelihood along the path mentioned in the Remark 5.

Moreover, the score function for small t causes less problem since for data under M_0, λ_2' is relatively small compared with that for data generated under M_1.

So we see when two models are close in some sense, we expect their likelihood ratio will not fluctuate widely provided the parameters from the two parameter spaces are properly aligned, for example, if found by minimizing a K-L divergence between the corresponding densities or taking a simple projection from the bigger space to the smaller spce. This is likely to make Importance Sampling more stable than if the two models were very different. It seems plausible that this stability or its lack in the calculation of BF will also show up in methods like PS that are derived from Importance Sampling in some way. Ingenious modifications of Importance Sampling seems to mitigate but not completely solve the problem. Following this idea of closer models in some sense, we modify PS in a similar manner below.

2.3.3 Path Sampling with Small Changes : Proposed Solution

In Remark 5, Subsection 2.3.1, a prior-likelihood conflict was identified as a cause of poor mixing. This will be re-examined in the next subsection. In the present subsection we propose a modification of PS which tries to solve or at least reduce the magnitude of this problem.

To solve this problem without having to give up our diffuse prior, we try to reduce the problem to a series of one-dimensional problems so that the competeting models are close to each other. We calculate the Bayes Factor by using the pathsampling step for every single parameter that may be zero, keeping others fixed. It is easily seen that the original log Bayes Factor is the sum of all the log Bayes Factors estimated in these smaller steps. We denote this procedure as PS-SC (Path Sampling with Small Change) and implement with parametric arithmetic mean path (PAMP). (As pointed out by a Referee, there is scope for exploring other paths, including a search for an optimal one, to reduce the MCMC-variance.) More formally, if we consider λ_2 as a p-dimensional vector, then M_0 and M_1 differ only in the last $p-1$ parameters, as λ_{21} is always zero due to upper-triangular condition. We consider p models $M_i' : i = 1, \ldots, p$, where for model M_i' we have first i parameters of λ_2 being zero correspondingly. If we define $BF'_{i,i+1} = \frac{m_i(x)}{m_{i+1}(x)}$, when $m_i(x)$ is the marginal for the model M_i' then,

$$\log BF_{21} = \sum_{i=1}^{p-1} \log BF'_{i,i+1}.$$

So we perform $p-1$ pathsampling computations to estimate $\log BF'_{i,i+1}, \forall i = 1, \ldots, p-1$. And for each of the steps the score function will be of the following form,

$$U'_i(\Lambda, \Sigma, \eta, Y, t) = \sum_{j=1}^{n} (y_j - \Lambda_t \eta_i)' \Sigma^{-1} (0^{p \times (h-1)}, [0_i; \lambda_{h,i+1}; 0_{p-i-1}]) \eta_i,$$

where $\Lambda_t = (\lambda_1, [0_i; t\lambda_{2,i+1}; \lambda_{2,(i+2,\ldots,p)}])$.

As in the case of small model true, the max likelihood under both models are close, and generally the two models are close, suggesting fluctuations are less likely and true BF isn't very large. This seems generally to lead to stability of computation of BF.

Also the parameter λ'_2 is now one dimensional. So the score function is more likely to be small than when λ'_2 is a vector as under PS. We also notice that in each step the score function is not anymore proportional to $\frac{\lambda'_2}{t}$ but rather to $\frac{\lambda'_{2i}}{t}$ which will be much smaller in value, hence reducing the fluctuation and loss of mass.

Computational implementation shows it to be stable for different MCMC-size and grid size regarding MCMC-standard deviation and also produces smooth curve of $E_t(U)$ for every single step. Here we use MCMC-size of 5000/50000 and grid size of .01 for our study and report the corresponding estimated BF values for two datasets from 1-factor and 2-factor model respectively. The MCMC-standard deviation of the estimates along with the mean of the estimated value over 25 MCMC runs are reported in Table 2.7. PS-SC has smaller standard deviation than PS under both M_0 and M_1. In Section 2.2 and Subsection 2.3.4, we argue that, at least under M_1, PS-SC provides a better estimate of BF.

Now we see the effect of changing the precision parameters keeping the factor loadings as before. The diagonal entries of Σ are in Table 8. The precision of these 3 models lie in the ranges of [2.77, 6.55], [1.79, 2.44], [1.36, 1.66] correspondingly.

We study PS-SC for 6 datasets generated from the 3 models (2 datasets with n=100 from each model: Data 1 from 1-factor and Data 2 from 2-factor model) and report the estimated Bayes Factor value in Table 2.9.

Table 2.7
$\log BF_{21}$ (MCMC-standard deviation) estimated by PAM-PS-SC and PAM-PS

True Model	MCMC Size	PS-SC	PS (t_{10})	PS (N(0,1))
1-factor	5000	-8.09 (.013)	-4.26 (.054)	-4.62 (.052)
1-factor	50000	-8.08 (.0067)	-4.24 (.012)	-4.60 (.0065)
2-factor	5000	80.14 (.66)	21.26 (1.39)	66.89 (4.15)
2-factor	50000	80.75 (.54)	23.71 (1.21)	68.21 (3.62)

Table 2.8
Diagonal Entries of Σ in the 3 different models: the first one is modified from Ghosh and Dunson (2008)

Model 1	.2079	.19	.15	.2	.36	.1875	.1875
Model 2	.553	.52	.48	.54	.409	.55	.54
Model 3	.73	.71	.67	.7	.599	.67	.72

Table 2.9
$\log BF_{21}$ (MCMC-standard deviation) estimation by PS-SC : Effect of Precision Parameter

Model	True Model	Data	PS-SC	PS (t_{10})
Model 1	1-factor	Data 1	-8.09 (.012)	-3.84 (.055)
	2-factor	Data 2	71.59 (.66)	19.81 (1.38)
Model 2	1-factor	Data 1	-11.01 (.0066)	-3.09 (.0277)
	2-factor	Data 2	51.41 (.3658)	2.8 (1.9104)
Model 3	1-factor	Data 1	-5.13 (.0153)	-2.6 (.0419)
	2-factor	Data 2	3.975 (.0130)	2.2 (.3588)

The effect of precision parameters are seen on the estimated value of the Bayes Factor (BF), more prominently when the 2-factor model is true. Generally the absolute value of the BF decreases with the decrease in the value of the precision parameters.

For the smaller value of precision parameters, we expect the model selection to be less conclusive, explaining the pattern shown in the estimated BF values.

Under M_1, PS is often bad in estimating the Bayes Factor (BF_{21}) but since the true Bayes Factor is large, it usually chooses the true model as often as PS-SC. When M_0 is true, PS is much better in estimating the Bayes Factor but since the Bayes Factor is usually not that large, it doesn't choose M_0 all the time. The probability of choosing M_0 correctly depends on the data in addition to the true values of the parameters. PS-SC does better than PS in all these cases, it estimates BF_{21} better and chooses the correct model equally or more often. The sense in which PS-SC estimates BF_{21} better has been discussed in detail earlier in this Section. Under M_0 PS-SC estimates BF_{21} better by having a smaller, i.e., more negative value than PS.

2.3.4 Issues regarding MCMC Sampling

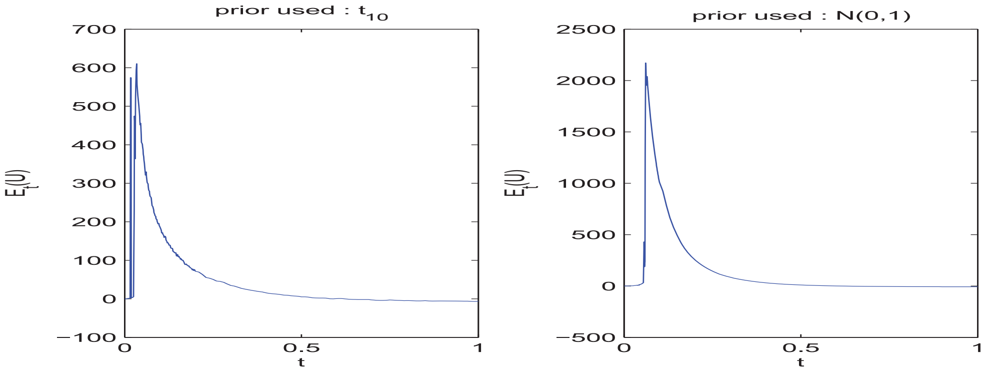


Figure 2.2. $E_t(U)$ for prior t_{10} and $N(0, 1)$, 2-factor model is true.

This subsection is best read along with the Remarks in Section 2.2. We first study the graph of $E_t(U)$ and the likelhood values for the MCMC samples at t for both the t_{10} and N(0,1) prior (figure 2.2 & 2.3). We will plot the likelihood as a scalar proxy since we can not show fluctuations of the vector of factor loadings in the MCMC output. The clusters of the latter can be inferred from the clusters of the former. *We*

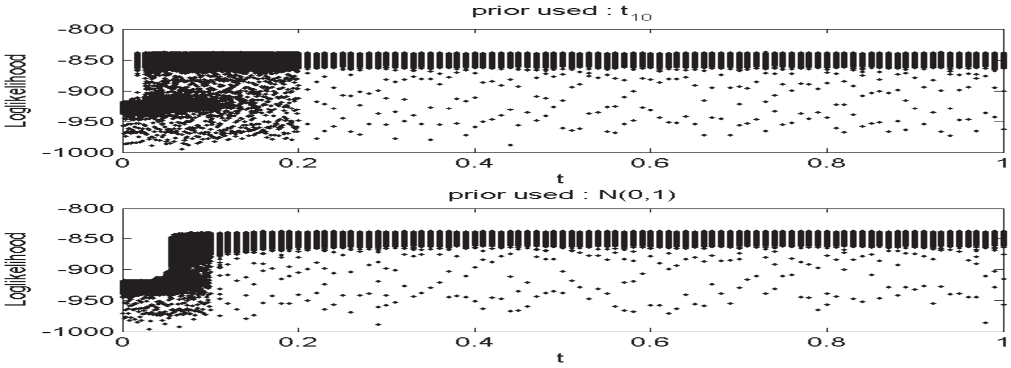


Figure 2.3. *Loglikelihood* for prior t_{10} and $N(0,1)$, 2-factor model is true.

will argue that there are two clusters at each grid point and the mixing proportion of the two clusters has a definite pattern.

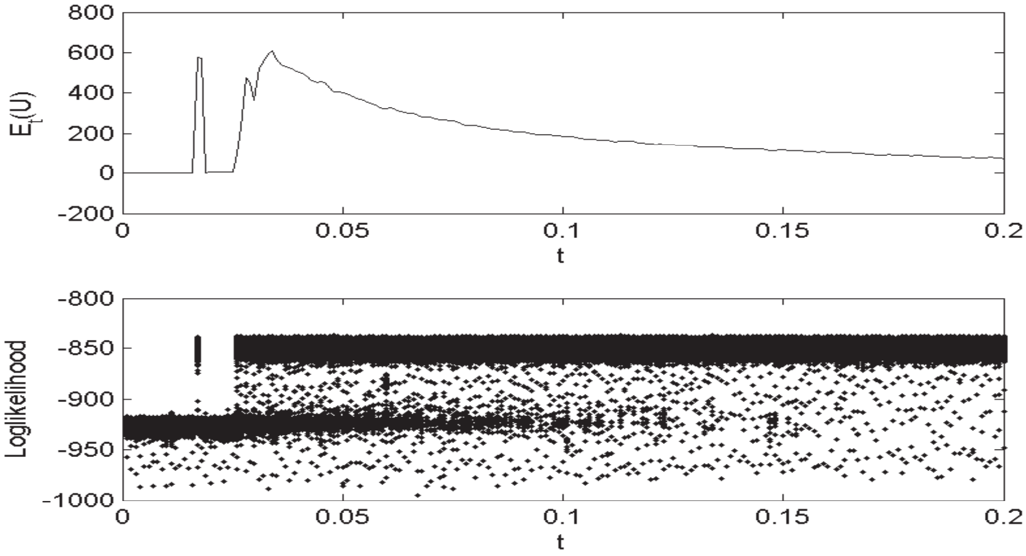


Figure 2.4. $E_t(U)$ and *Loglikelihood* for prior t_{10} in the range $t \in [0, .2]$, 2-factor model is true.

Under the true 2-factor model M_1, denote $\lambda' = [\lambda'_1, \lambda'_2]$ where λ'_i is the loading for the corresponding latent factor under M_t. Here λ'_2 is a 7×1 vector and becomes zero, as it approaches M_0 from M_1 (as $t \to 0$). The posterior distribution at each M_t can be viewed roughly as a sort of mixture model with two components representing M_0 and M_1, the form of the likelihood as given in Theorem 3. In the diagram (Fig 2.4) of log-likelihood of MCMC samples, we see two clear clusters around log-likelihood values -850 and -925, representing MCMC outputs with nonzero λ_2' and zero λ_2' values respectively. We may think of them as coming from the component corresponding to M_1 (cluster 2) and the component corresponding to M_0 (cluster 1). Samples of both clusters are present in the range [.03,.2], while samples appear to be predominantly from cluster 2 until t=.1. A good representation of samples from cluster 1 are only present in the range [0,.1]. In the range [.03,.2], both clusters occur with proportions varying a lot. Moreover here the magnitude of the score function is proportional to $\frac{\lambda_2'}{t}$. We see these fluctuations in Fig (2.4) in the region [.03,.2]. This is also brought out by the MCMC standard deviation of $E_t(U)$ *which are of order of 30-50 in log scale.*

We notice the absence of any samples from M_1 for $t < .03$, except some chaotic representation for a few random values of t (notice in the figure, a spike representing samples from M_1 at t=0.016), clearly representing poor mixing of MCMC samples near the model M_0.

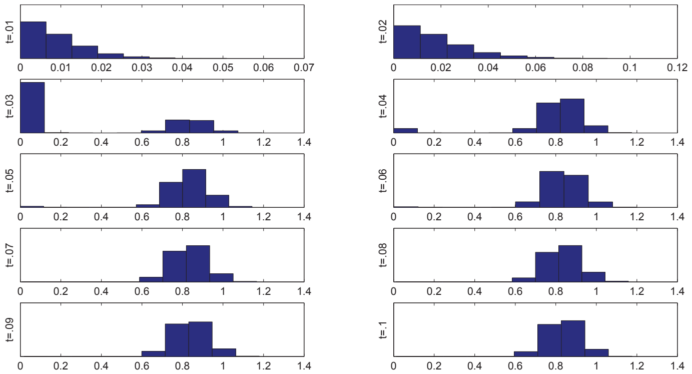


Figure 2.5. Histograms for λ_{22}' for different values of t near t=0 (MCMC size used 50,000), using PS.

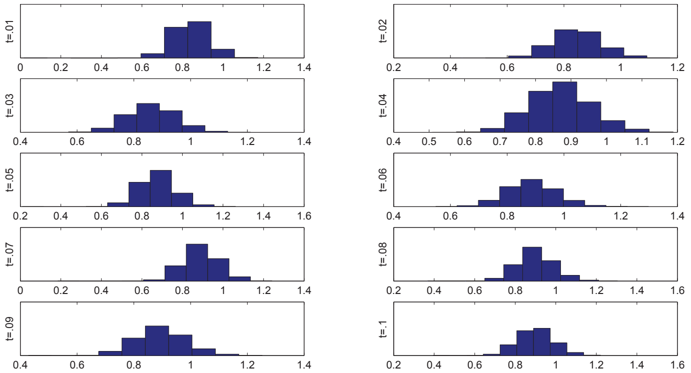


Figure 2.6. Histograms for λ'_{22} for different values of t near t=0 (MCMC size used 50,000), using PS-SC.

The new method PS-SC stabilizes the estimated Bayes Factor value with a very small MCMC-standard deviation. Here we check through figures 2.5 & 2.6 that it avoids prior-likelihood conflict and the problem about mixing for MCMC samples seen for the standard PS. We concentrate our study for the first step of PS-SC. In this step only the first component of λ'_2, λ'_{22} converges to zero as $t \to 0$. So here we consider the spread of MCMC sample of λ'_{22} for different values of t near t=0, from both PS and PS-SC in figure 5 & 6 by considering the histogram of MCMC sample of λ'_{22}. We can easily notice that the spread of the MCMC sample fluctuates in between the two modes in a chaotic manner showing poor or unstable mixing for PS, whereas PS-SC samples come from both the clusters and slowly shift towards the prior mode as $t \to 0$. We have also studied but don't report similar nice behavior regarding mixing of MCMC of PS-SC for the data simulated from 1-factor model.

The poor mixing discussed above for MCMC outputs for PS will now be illustrated with plots of auto-correlation for λ'_{22} for different lags (Figure 2.7). For the sake of comparison we do the same for PS-SC (Figure 2.8). Clearly except very near t=0, i.e., in what we have called the chaotic zone, the auto-correlations for PS are much bigger than those for PS-SC. However, near t=0, though in plots in both Figure 2.7

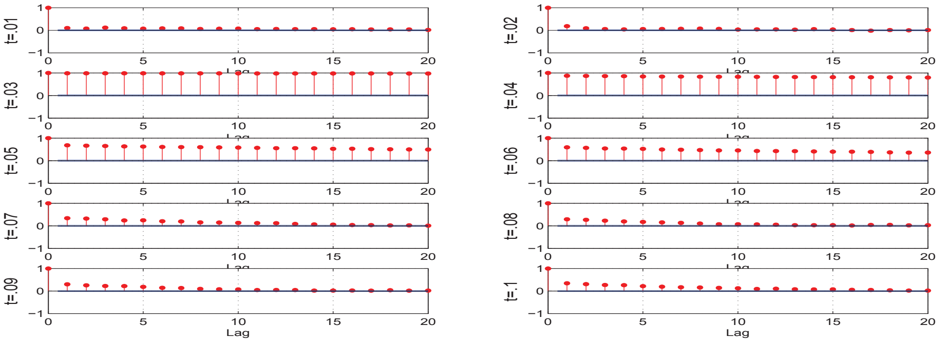


Figure 2.7. Autocorrelation for λ'_{22} for different values of t near t=0 (MCMC size used 50,000), using PS.

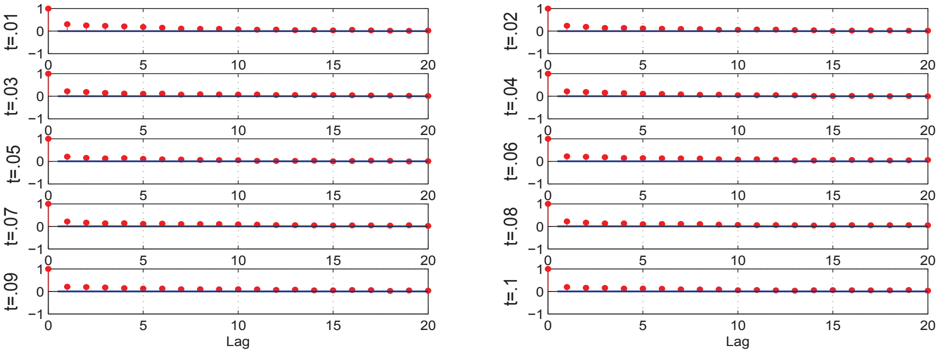


Figure 2.8. Autocorrelation for λ'_{22} for different values of t near t=0 (MCMC size used 50,000), using PS-SC.

& 2.8 are small, those for PS are slightly smaller. We have no simple explanation for this.

Poor mixing seems to lead to missing mass and random fluctuations for calculations for $E_t(U)$. This probably explains the discrepancy we have noticed in the estimation of BF by PS as compared with PS-SC. We now look at autocorrelations for a first factor loading in Figure 2.7 and second factor loading in Figure 2.8. The top rows in each of the two figures show zero autocorrelation as they are very close to $t = 0$. On the other hand, high autocorrelations are shown in the next two rows. We believe they correspond to what we called a chaotic region. The bottom two rows of Figure 2.8 show small autocorrelation. They correspond to the second factor loading which comes in only Model 2, and they also depict the zone dominated by Model 2. The other figure is in the same zone as in the previous line, but the variable considered is a 1-factor loading. Here also autocorrelation eventually tends to 0, but its values are bigger than in Figure 2.8. We don't have any simple explanation for this higher autocorrelation.

The above discussion covers the case when the more complex model is true. If the simpler model (M_0) is true, as noted in Subsection 2.3.3 both PS and PS-SC perform well in estimating the Bayes Factor as well as choosing the correct model. The Bayes Factor based on PS-SC provides stronger support for the true model than the Bayes Factor based on PS.

To check whether PSSC works well in other Examples as in Factor Model, we try to explore its impact on our earlier toy-example. In this case, we were unable to implement Path Sampling with Small Changes, but rather used a pseudo-PSSC scheme. Going back to our example where we have taken m=7 and p=10, we define a sequence of models as following :

$$M_i : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}\right) \text{ when } A_{11} \text{ is } (i \times i) \text{ matrix for } i = 7, 8, 9, 10.$$

We can see our previously defined M_0 and M_1 are now M_7 and M_{10} correspondingly. For our pseudo-PSSC, we estimate $logBF_{i,i+1}$ by $logBF$ between the models M'_0 and M'_1, with $m = i$ and $p = i + 1$:

$$M'_t : y_i \sim N\left(0, \Sigma = \begin{pmatrix} A_{11} & tA_{12} \\ t(A_{12})' & A_{22} \end{pmatrix}\right).$$

Still being underestimates on each step, this method improves on the standard Path Sampling in terms of Bayes Factor estimation as we can see in the following table.

Table 2.10
Performance of PS and pseudo-PS-SC in Toy Example modelling Covariance : Log Bayes Factor (MCMC standard deviation)

Method	Data 1	Data 2
True BF value	258.38	-132.87
PS estimate of BF	184.59 (.012)	-20.11 (.008)
pseudo-PSSC estimate of BF	195.35 (.011)	-25.21 (.007)

2.4 Implementation of other Methods

We have explored several methods of estimating the ratio of normalizing constants, for example the methods of Nielsen [2004], DiCiccio et al. [1997], Rue et al. [2009] and Chib [1995]. The method of Rue et al. [2009] models a link function of means but here we are concerned with models for the variance-covariance matrix. We could not use Chib's method here since for our parameter expanded prior the full conditionals of the original model parameters are not available. But we were able to implement the deterministic variational Bayes method of Nielsen [2004] and the Laplace approximation with a correction due to DiCiccio et al. [1997]. Since the results were not satisfactory, we do not report them in this paper. In the variational Bayes approach, the method selected the correct model approximately 80% of the time but the estimated logBF values were considerably over (or under) estimated. The variational Bayes method is worth further study, possibly with suitable modifications. It appears to us it is still not understood when Belief Propagation provides a good approximation to a marginal or not, e.g., Gamarnik and Shah [2010] commented. *Only recently we have witnessed an explosion of research for theoretical understanding of the performance of the BP algorithm in the context of various combinatorial optimization problems, both tractable and intractable (NP-hard) versions.*

Following the discussion in the Subsection 2.2.4, we have implemented the GMP-PS. Here the marginal for both models is estimated by constructing a path between the prior distribution to the posterior distribution of the model. Due to very high-dimensionality of the model, the mode of prior and posterior distribution are far apart. So as discussed before, the MCMC sampling along the path fails to sample smoothly and fluctuates between the two modes in a chaotic way near the prior mode. Hence the estimate of marginal of both the models become very unstable. Due to the poor estimation of BF this method also fails to choose the correct model very often. As in the case of GMP-PS, AIS with GM path did not also work well. Hence, we implemented AIS with PAM-path. Implementation of PAM-AIS is also very time intensive, so we have only implemented PAMP-AIS with MCMC sample size 5000. PAM-AIS not only shows very high MCMC-standard deviation, but it also fails to choose correct model many a time, when the 2-factor model is correct. The last methods we looked at are,

1. Importance Sampling (IS).

2. Newton-Raftery approximation (BICM).

3. Laplace/BIC type approximation (BICIM).

IS is the most easy to implement and shows moderately good results in choosing the correct model (Ghosh and Dunson [2008]). We study the stability of Bayes Factor values estimated by IS with the change of the MCMC size in Table 2.11.

Similarly we also study the stability of the estimates of Bayes Factor by BICM and BICIM (explained in A.3 in the Appendix) using MCMC sample size 10,000, where both of these methods show significantly less amount of MCMC-standard deviation than other methods considered. Hence we will only consider PS-SC, BICM and BICIM to explore model selection for dimension much higher than previously considered.

Table 2.11
Study of IS, BICM and BICIM for different MCMC size: Estimated Bayes Factor (MCMC Standard Deviation)

Method(MCMC-size) / True Model	2-factor Model	1-factor Model
IS (10,000)	109.78 (168.72)	.0749 (.1063)
IS (50,000)	97.12 (61.25)	-5.39 (84.35)
IS (100,000)	86.92 (110.35)	-3.07 (10.41)
IS (200,000)	83.66 (58.53)	-2.69 (2.96)
BICM (10,000)	68.66 (.93)	-5.72 (.62)
BICIM (10,000)	67.9 (.11)	-5.3 (.57)
PS-SC (5,000)	80.75 (.63)	-8.08 (.0013)

2.5 Effect of Precision parameters and High Dimensional (Simulated and Real) Dataset

Our goal is to explore if PS-SC may be made more efficient by combining with BICM and BICIM and also to explore number of dimension much higher than before and real life examples.

In the examples in this section, p varies from 6 to 26. We have 2 examples of real life examples with p=6 and 26 and simulated example with p=20. As expected PS-SC still takes long time, even with a parallel processing for high dimensional examples. We explore whether PS-SC can be combined with BICM and BICIM to substantially reduce time. Since their performance seems much faster than PS-SC.

We compare the behavior of these methods for higher dimensional model and for some real datasets taken from Ghosh and Dunson [2009] and Akaike [1974b]. We first consider one 3-factor model with p=20 and n=100.

We notice that all the methods are selecting correct models for all the 3 datasets, but based on our earlier discussion of PS-SC, we believe only this method provides a reliable estimate of BF. Now we will compare the methods for some real datasets.

Table 2.12
Simulated model (p=20, n=100) and (k=the number of true factors)
: Comparison of log Bayes Factor

Data	BF	PS-SC	BICM	BICIM
Data1 (k=1)	BF_{21}	-25.91 (.0233)	-32.68	-38.01
	BF_{32}	-24.84 (.0594)	-21.18	-38.24
	BF_{43}	-22.79 (.0483)	-19.81	-43.77
Data2 (k=2)	BF_{21}	225.81 (4.2099)	248.09	219.87
	BF_{32}	-23.61 (.0160)	-23.59	-46.17
	BF_{43}	-19.18 (.0297)	-20.3	-47.98
Data3 (k=3)	BF_{21}	152.07 (1.7422)	185.45	162.3
	BF_{32}	104.17 (2.5468)	198.1	168.54
	BF_{43}	-17.35 (.0276)	-29.73	-48.24

We choose two datasets: "Rodent Organ Data" from Ghosh and Dunson [2009] and "26-variable Psychological Data" from Akaike [1974b]. These datasets have been normalized first before analyzing them further. We not only study the estimated Bayes Factor but also the model chosen by them.

Table 2.13
Rodant Organ Weight Data (p=6, n=60) : Comparison of log Bayes factor

Bayes factor	PS-SC	BICM	BICIM
$logBF_{21}$	4.8	26.34	21.57
$logBF_{32}$	10.52	-3.14	-10.01
$logBF_{43}$	-3.28		

In the "Rodant Organ Data" the model chosen by PS-SC and other methods are correspondingly 3-factor model and 2-factor model. For the "26-variable Psychological Data", where PS-SC, and BICM/BICIM chooses the model with 3 factors and 4

Table 2.14
26-variable Psychological data (p=26, n=300) : Comparison of log Bayes factor

Bayes factor	PS-SC	BICM	BICIM
$logBF_{21}$	122.82	205.27	188.19
$logBF_{32}$	35.27	71.05	35.5
$logBF_{43}$	-10.7	23.16	7.55
$logBF_{54}$	-33.32	-4.63	-25.51
$logBF_{65}$	-16.7	-17.32	-43.21

factors respectively. The models chosen by PS-SC and the other methods are close, but as expected differ a lot in their estimate of BFs.

There is still no rigorously proved Laplace approximation for relatively high dimensional cases because of analytical difficulties. Problems of determining sample size in hierarchical modeling, pointed out by Clyde and George (2004) is avoided by both versions of our approximations (A.3). These two methods seem to be good as a preliminary searching method to narrow the field of plausible models before using PS-SC. This saves time relative to PS-SC for model search as seen in the previous examples.

2.6 Conclusion

We have studied PS for Factor Models (and one other toy example) and have identified the component of PS that is most likely to go wrong and where. This is partly based on the fact that we have a relatively simple sufficient condition for factor models (Theorem 3). Typically for the higher dimensional model the MCMC output for finding the integral along grid points in the path may become quite unreliable at some parts of the path. Some insight about why it happens and how it can be rectified has been suggested. MCMC seems to be unreliable for PS when the higher dimensional model is true. The problem is worse the more the two models differ as when a very high dimensional model is being compared to a low dimensional model.

The suggestion for rectification was based on the intuition that PS, like Importance Sampling itself, seems more reliable when the two marginal densities in the Bayes Factor are relatively similar, as is the case when the smaller of two nested models is true Based on this intuition we suggested PS-SC and justified PS-SC by comparing MCMC output and MCMC standard deviation of both PS-SC and PS.

It is our belief that the above insights as to when things will tend to go wrong and when not, will also be valid for the other general strategy for selection from among nested models namely, RJMCMC.

Our work has focused on model selection by Bayes Factors, which seems very natural since it provides posterior probability for each model. However, model selection is a complicated business and one of its major purposes is also to find a model that fits the data well. Several model selecting statisticians feel this should also be done along with calculation of Bayes Factors.

However, there has not been a good discussion on how one should put together the findings from the two different approaches. We hope to return to these issues in a future communication.

A natural future direction of our study of Factor Models is to add to the model an unknown mean vector with a regression setup. The problem now would be to simultaneously determine a parsimonious model for both the variance-covariance matrix and the mean vector. There are natural priors for these problem but computation of the Bayes Factor seems to be a challenging problem.

3. CLASSICAL MODEL SELECTION : INFERENCE ABOUT SNPS

We first consider the prolems regarding to SNP's. We quote from Thain et al. [2004], single nucleotide polymorphisms are "single DNA base alterations between human individuals", which "are being analysed" as part of association studies between genes (or markers) and diseases.

The statistical problem for identifying significant SNP's from among a huge number, easily a few thousands, is like choosing a few markers in Quantitative Trait Loci (QTL) studies from among many, see for example, Bogdan et al. [2004]. This is a regression problem involving variable selection. If the design matrix were orthogonal, so that the least squares estimates based on the full model are basically independent of model (i.e., of all models that include the particular variable under study), the methods that have been successful for microarrays can be used. Without orthogonality, the high dimensional regression problem is much more difficult. Theoretical study has just begun. We discuss this a bit below.

We illustrate the difficulties by explaining why the theory of optimality of the Benjamini-Hochberg rule will not apply at all. Without independence the notion of the Bayes oracle of Bogdan et al. [2011] is not available. Even more fundamentally, without independence, the Benjamini-Hochberg multiple test isn't easy to define, nor is the theorem of Benjamini and Hochberg [1995] applicable. On the other hand in new work on multiple regression, Bickel et al. [2009] study optimality of popular procedures like Lasso or its relatively recent competitor, the Dantzig selector of Candes and Tao [2007], by studying suitable oracle attaining properties in the sparse case. Oracles are lower bounds to measure of risk of a decision rule.

In the case of SNP's the present study owes a lot to Frommlet et al. [2011] for several basic ideas. Like them we study the following variable selection procedures for a simulated example, namely mBIC of Bogdan et al. [2004, 2008a,b] and Lasso due to Tibshirani [1996]. However, we also study Lasso with a different penalty that is suggested in Bickel et al. [2009] as suitable in the context of their Oracle, i.e., a lower bound that Lasso attains in the sparse case.

Finally we evaluate each procedure by its predictive performance as given by the ratio of Residual Sum of Squares and Total Sum of Squares from the ANOVA table and the accuracy of estimating the number of SNP's in the data. Associating significant β's and true non-zero β's which represent significant SNP's is much more tricky and requires some form of bootstrap sampling and clustering of covariates. It appears one can associate significant SNP's only with such clusters of covariates, which are our proxy clusters of markers in real experiments. Our evalution of Lasso and mBIC is quite different from Frommlet et al. [2011]. We are indebted to Frommlet et al. [2011] for all these insights about SNP's and in the way we generate simulated data. We strongly advise interested readers to read their paper carefully.

3.1 Model assumptions

The SNP based Genome Wise Association Study (GWAS), can be easily seen as a multiple linear regression problem with variable selection as the key issue. Ideally each SNP corresponds to a covariate, so identifying SNP's is equivalent to identifying significant regression coefficients. Actually, the problem is much more delicate because of correlation between covariates. We discuss the more realistic version towards the end of this paper. Let us treat the quantitative trait of n observations as the response varible $y_i : \imath \in \{1, \ldots, n\}$ and the corresponding genotype of person i and SNP j as $x_{ij} \in \{-1, 0, 1\} : \imath \in \{1, \ldots, n\}, \jmath \in \{1, \ldots, p\}$. Now if the subset j^* of the p SNPs, having $k << p$ SNPs $1 \leq j_1^* \leq \ldots \leq j_k^* \leq p$, are causal for the trait y, then we can assume the additive true model as

$$M_j^* : \; y_i = \sum_{l=1}^{k} \beta_{j_l^*} x_{ij_l^*} + \epsilon, \; when \; \epsilon \sim N(0, \sigma^2 I_k) \tag{3.1}$$

But we can also create 2^p similar models using all p SNPs. A model not having any SNP will be denoted as M_0 or the null model and all other models will be denoted by M_j, where j is an ordered subset of elements of the set $\{1, \ldots, p\}$. Following the standard conventions, we write $q = q_j$ for the number of SNPs in a model.We define for each model M_j the matrix X^j containing the genotype of the SNPs in the model. Then the model becomes:

$$M_j : \; y = X^j \beta_j + \epsilon^j \tag{3.2}$$

Now as in multiple regression problem with variable selection, we want to choose the best model containing all the causal variables here. So we will discuss some variable selection methods in the next subsection under sparsity assumption, as our $k << p$.

3.2 Lasso, Lars and Stepwise selection

For variable selection in linear regression problems different skrinkage estimators like ridge regression are very common in use. Following the idea of shrinkage as in ridge regression, the Lasso method introduced by Tibshirani [1996], tries to minimize the least square error of the regression with an upperbound on the L_1 norm of the parameter vector. So the estimate is defined by,

$$\widehat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{3.3}$$

$$subject \; to \; \sum_{j=1}^{p} |\beta_j| \leq t$$

Here the upper bound t for the L_1-penalty of the parameter-vector controls the amount of shrinkage. Lasso chooses subsets of variables depending on the tuning-parameter t. When t is very small, then almost all the parameters are zero, similarly for large enough t value, the parameter estimate $\widehat{\beta}$ is same as the least square estimate. The shrinkage constraint makes the solution nonlinear in y_i, needing a quadratic programming algorithm to compute the estimate. Efron et al. [2004] have shown that Lasso is closely related to another novel shrinkage estimation scheme Lars, introduced by them. In a general setup, they show that the subset selected by Lars, Lasso and Stepwise Selection are similar. Here we will concentrate on the most popular of them, namely, Lasso, its solution path obtained from Lars algorithm for fast variable selection. This has become the standard method for Lasso. The tuning parameter t is chosen by minimizing the cross-validation error.

Suppose we have a statistical inference problem, e.g., testing or estimation. An oracle depending on unkown parameters, is a lower bound for the risk or loss function of all decision functions we consider, and which is asymptotically attained by our chosen rule. If one can construct such an oracle for a particular decision rule, it immediately proves the asymptotic optimality of the chosen decision function. Such oracles were first proposed for AIC, Shibata [1994], Li [1987] and Shao [1997]. An early oracle (not stated as such) is the Cramer-Rao inequality, which is an oracle for the mle. The results in Bogdan et al. [2008a, 2011] are based on a Bayes oracle for all multiple tests.

Recently Bickel et al. [2009] have derived an Oracle property for Lasso under sparsity assumption. The Lasso constraint $\sum |\beta_j| \leq t$ is equivalent to the addition of a penalty term $r \sum |\beta_j|$ to the residual sum of squares [Murray et al. [1981]]. While an explicit mathematical relation between t and r isn't available, the basic idea of convex optimization makes it easy to move from the one to the other. We will use both versions of Lasso. This can be written as following,

$$\widehat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + r \sum_{j=1}^{p} |\beta_j| \tag{3.4}$$

According to Bickel et al. [2009], when the errors ϵ_i are independent $N(0, \sigma^2)$ random variables with $\sigma^2 > 0$, all the diagonal elements of the matrix $X'X/n$ be equal to 1, then under some additional conditions on the Gram matrix, with $r = A\sigma\sqrt{\frac{log(p)}{n}}$ and $A > 2\sqrt{2}$, with probability $1 - M^{1-\frac{A^2}{8}}$, we have

$$|\widehat{\beta}^{lasso} - \beta_0|_1 \leq \frac{16A}{c(s)}\sigma s\sqrt{\frac{log(p)}{n}} \tag{3.5}$$

$$|X(\widehat{\beta}^{lasso} - \beta_0)|_2^2 \leq \frac{16A^2}{c(s)}\sigma^2 slog(p) \tag{3.6}$$

$$\#\{\widehat{\beta}_j^{lasso} \neq 0\} \leq \frac{64}{c_1(s)}s \tag{3.7}$$

when s is the no. of non-zero components in β_0, and c(s), $c_1(s)$ are constants depending on s and the Gram matrix. They have very similar oracles for the Dantzig selector of Candes and Tao [2007], suggesting both methods achieve similar goals. The oracle also makes clear that the penalty should change with sparsity.

The penalized model selection schemes like Lasso have also been implemented in the Bayesian set up by Park and Casella [2008] and Kyung et al. [2010]. They have shown that with the proper choice of prior distribution, the posterior distribution introduces a penalty term, e.g. Laplace (double-exponential) distribution as prior associates a penalty term that is same as Lasso, thus replicating penalized model selection schemes by Bayesian methods. They also demonstrated that the generalized Lasso schemes including Fused Lasso, Elastic net, Group Lasso and others can be implemented in Bayesian set up by proper choice of a prior in the Bayesian Lasso. The discussions about the consistency, standard error, performance and comparison of Bayesian Lasso with standard penalized schemes are also illuminating (Kyung et al. [2010]).

While on the subject of consistency in the context of linear models, we like to mention a few papers chosen from the emerging literature on consistency used in the sense of approximation from a given dictionary of functions. Much of the new vocabulary has come from Machine Learning, but almost all the papers we cite have appeared in statistical journals, some have a Bayesian flavor, Bunea et al. [2006, 2007], Zhao and Yu [2007]. Bunea et al. [2007] contains many references on sparsity, oracles, information theoretic limits. Several of these relate to the Lasso.

3.3 Modified Bayesian Information Criterion

For linear regression under assumption of normal error term $\epsilon \sim N(0, \sigma^2)$ the likelihood function of each model M_j is given by

$$L_j(y|\beta_j, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{(y - X^j\beta_j)'(y - X^j\beta_j)}{2\sigma^2}\right). \tag{3.8}$$

The maximum likelihood estimator of β_j is same as the least square estimate $\widehat{\beta_j}$. So we know for fixed σ using BIC is then equivalent to minimizing a standardized residual sum of squares RSS_j/σ^2 under model M_j with penalty $q_j log(n)$ for model dimension q_j.

$$\frac{RSS_j}{\sigma^2} + q_j log(n). \tag{3.9}$$

But it is known (Bogdan et al. [2004]) that under sparsity, BIC chooses too many regressors. As a remedy, Bogdan et al. [2004] introduced a modification of BIC, as:

$$mBIC : -2logL_j(\widehat{\beta_j}) + q_j(log(n) - 2log(w)) \tag{3.10}$$

where, w can be interpreted as a probability of a particular covariate being relevant. When such prior information isn't available there is also a default choice of w. In recent unpublished work , following a similar idea, Frommlet et al. [2011] introduced a different criterion, namely,

$$mBIC2 : -2logL_j(\widehat{\beta_j}) + q_j(log(n) - 2log(w)) - 2log(q_j!) \tag{3.11}$$

which is suitable for multiple regression and works similarly to the Benjamini-Hochberg correction for multiple testing. We study the performance of this new criterion with Lasso for a simulation based GWAS study. In GWAS study, the number of parameters is too big, so to apply mBIC2, they used a pre-screening scheme, which picks a smaller but more relevant subset of parameters to explore further. They conducted 1-variable β-significance test for each variables and then created a subset consisting only of the variables having a p-value below a pre-specified threshold. They chose the threshold by using their (assumed) rough prior knowledge about the expected number of significant variables. Following Frommlet et al. [2011], we also take the threshold here as 1.5. After the first stage of screening, we choose the final model minimizing mBIC2 criterion with a forward selection procedure. This prior information is used only for mBIC2 but not for Lasso.

3.4 Simulation Study

Like Frommlet et al. [2011] we generated a dataset to mimic a SNP dataset in real life, but still having control over the parameters. We assumed the sample size n=100, and the total number of SNP's under study to be p=30,000. Eeach covariate has been set to be a $\{-1, 0, 1\}$ valued r.v. as in the case of SNP, with the minor allele frequency always lying below 0.5. For our study the minor allele was taken as "1". We also checked the covariance between any two covariates in most of the cases to be in between -.1 to .1, signifying weak but not negligible covariance between the covariates. Our data set is somewhat smaller than the simulated data set of Frommlet et. al. We denote the covariate matrix as X, which is an $n \times p$ matrix. Forty causal SNP's have been selected randomly from the set of all "p" SNPs. Then we have chosen a vector β_0 having non-zero coefficient value lying between .5-1 for those 40 SNPs and we simulate the response as following:

$$y = \beta_0' X + \epsilon, \quad when\ \epsilon \sim N(0, \sigma^2 I) \tag{3.12}$$

Three variable selection procudres, namely Lasso with Cross validation, Lasso with Oracle-penalty and mBIC2, have been compared in this setup. The Lasso with Oracle property is new, not considered in Frommlet et al. [2011]. Since detection of non-zero β's is a much more tricky task, we compare their predictive performance by the ratio of Residual Sum of Squares(RSS) and Total Sum of Squares(TSS) in Table 1, when we define RSS and TSS as following,

$$\begin{aligned} RSS &= \sum (Y_i - X'\widehat{\beta})^2 \\ TSS &= \sum (Y_i - \bar{Y})^2. \end{aligned} \tag{3.13}$$

Also in the definition of mBIC2, we have used the prior knowledge w about the sparsity. So to compare it with Lasso, where we don't use any prior information, we consider mBIC2 for 3 different values of $w = \frac{30}{30000}, \frac{60}{30000}, \frac{120}{30000}$.

Table 3.1
$\frac{RSS}{TSS}$ for different Model Selection Rules

Different Model Selection Rules	$\frac{RSS}{TSS}$
Lasso-CV	.227
Lasso-Oracle	.092
mBIC2($w = \frac{60}{30000}$)	.02
mBIC2($w = \frac{30}{30000}$)	.018
mBIC2($w = \frac{120}{30000}$)	.033

mBIC2 does best but also chooses more variables depending on the information of the sparsity. It chooses 15, 56 and 99 variables correspondingly for $w = \frac{30}{30000}, \frac{60}{30000}, \frac{120}{30000}$. But both the versions of Lasso choose much smaller number of variables, 23 and 31 in Lasso-CV and Lasso-Oracle correspondingly, i.e., they were much more parsimonious. Lasso-Oracle comes nearest to estimating the correct number of SNP's. Unfortunately our simulation would need to be strengthened with Bootstrap before we can identify clusters of co-variates as causal variables, as mentioned in Introduction.

To check if the estimate of RSS/TSS will increase substantially under cross-validation, we simulated another data set and calculated RSS/TSS with the same estimates obtained earlier. We get this time the value in a $\pm.01$ interval of the value found earlier. So our earlier conclusions do not change substantially, as we expected since the total data size is much bigger than the number of unknown parameters.

3.5 Discussion

Even under sparsity there may be approximate colinearity between covariates, as in our simulated example. In this case the correct model may not be identifiable. This will often be the case for studies involving SNP's. Such problems require a thorough simulation and theoretical study.

4. MODEL SELECTION FOR MONOTONICITY AND CHOICE OF VARIABLES IN NONPARAMETRIC SETTINGS

Variable selection for high-dimensional saprse additive nonparametric regression under monotonicity constraints has been first discussed in Fang and Meinshausen [2011]. They suggested a backfitting algorithm called LASSO Isotone (LISO). Total variation of function f is defined as $\Delta(f) = \sup_{x \in R} f(x) - \inf_{x \in R} f(x)$, where R is the range considered for the function f. When M_j is the class of one-dimensional increasing functions, a LISO solution for a particular value of tuning parameter $\lambda \geq 0$ is defined as the minimser $\widehat{m}_\lambda = (\widehat{m}_{j,\lambda})_{j=1}^J$, with $\widehat{m}_{j,\lambda} \in M_j \; \forall j = 1, \ldots, J$; of the LISO loss

$$L_\lambda(m_1, \ldots, m_J) := \frac{1}{2}||Y - \sum_{j=1}^{J} m_j(W_j)||^2 + \lambda \sum_{j=1}^{J} \Delta(m_j). \tag{4.1}$$

An backfitting algorithm for the above minimization problem utilizes the additive structure of the model. At each step all the components are fixed, except the one we want to estimate, to their estimates from the previous step. Then the components are updated using the residual calculated by subtracting estimates of the fixed components from the response. Each component is updated by fixing similarly. This cyclic procedure is continued until desired convergence is reached. As an initial estimate zero functions for all the components and the parameter λ has been chosen by cross-validation scheme.

Hence in each step of the algorithm, it becomes an one-dimensional nonparametric regressional problem under monotonicity constraint. In the absence of the penalty term the PAVA algorithm gave a consistent solution (Mammen and Yu [2007]). For this penalized problem, Fang and Meinshausen [2011] has shown that the LISO solution is a specififc winsorized PAVA estimate. Further details about the optimal thresholding levels for this Winsorazied estimates can be found in [Fang and Meinshausen, 2011, Theorem 1, Theorem 2]. Their simulation studies have shown that the significant variables have been chosen correctly and the estimates of the monotone components were also consistent. But further theoretical studies regarding the consistency of the model selection will be an interesting future direction of work.

In this Section we will suggest an algorithm where we will automatically decide about the monotoniciy of the nonparametric components while doing variable selection. Hence we will not have to assume the nonparametric components to be increasing, decreasing or nonmonotone. But we know given the monotonicity information of the components, we can get a better fit of the nonparamettric components. Hence we will try to adapt LISO algorithm in such a way that we will not loose the benefit of knowing the monotonicity of the components in this new setup. To facilitate our scheme, we assume that the nonparametric components $(h_j)_{j=1}^J$ have bounded total variation. We consider the following nonparametric regression model :

$$Y_i = \sum_{j=1}^{J} h_j(W_{ij}) + \epsilon_i \tag{4.2}$$

for $i = 1, \dots, n$, where every h_j is a function with bounded total-variaton. It is a well-known fact (Ito [1993]) that functions of bounded variation have a unique Jordan decomposition

$$h_j = m_j^+ + m_j^-$$

into monotonically increasing and decreasing functions $m_j^+(x)$ and $m_j^-(x)$, such that $\Delta(h_j) = \Delta(m_j^+) + \Delta(m_j^-)$. Hence we will use the original and the reversed covariate to estimate m_j^+ and m_j^- correspondingly. We define $\tilde{m}_j(-x) = m_j^-(x)$, where $m_j^+(x), \tilde{m}_j(-x) \in M_j$. Then we can estimate the function h_j by combining these two estimates. So we consider an equivalent model as of (4.2) :

$$Y = \sum_{j=1}^{J} m_j^+(W_j) + \sum_{j=1}^{J} \tilde{m}_j(-W_j) + \epsilon \tag{4.3}$$

where m_j^+ and $\tilde{m}_j$ are assumed to be monotone increasing functions along the covariate X and $-X$ correspondingly. We can see that our final estimate of the function h_j

will be $\widehat{h}_j(x) = \widehat{m}_j^+(x) + \widehat{\tilde{m}}_j(-x)$. Hence we have essentially reduced the problem to a nonparametric regression problem where all the nonparametric components are assumed to be monotone increasing. Now we will put a penalty term for each of the components as in LISO loss of equation (4.1). So our solution for a particular value of tuning parameter $\lambda \geq 0$ is defined as the minimser $(\widehat{h}_\lambda(W_j) = \widehat{m}^+_{j,\lambda}(W_j) + \widehat{\tilde{m}}_{j,\lambda}(-W_j))_{j=1}^J$, with $\widehat{m}^+_{j,\lambda}(W_j), \widehat{\tilde{m}}_{j,\lambda}(-W_j) \in M_j \; \forall j$; of the loss

$$\begin{aligned} L_\lambda(h_1, \ldots, h_J) : &= \frac{1}{2}||Y - (\sum_{j=1}^J m_j^+(W_j) + \sum_{j=1}^J \tilde{m}_j(-W_j))||^2 \\ &+ \lambda(\sum_{j=1}^J \Delta(m_j^+(W_j)) + \sum_{j=1}^J \Delta(\tilde{m}_j(-W_j))). \end{aligned} \tag{4.4}$$

We notice that for each of the components we have penalties for both the increasing and decreasing part of the nonparametric component. So when $h(x)$ is a decreasing or increasing function the penalty term will put the corresponding increasing or decreasing part to zero. When $h(x)$ is a non-monotone function both the components will be kept and the final estimate will be found by combining them. Hence minimizing the loss function in equation (4.4) we can simulatneously decide about the monotonic behavior of the significant nonparametric components. To find the minimizer we will be using the similar backfitting algorithm for LISO, as all the components are still monotone increasing functions. We will use the original and the reversed covariates as our set of covariates to estimate the monotone increasing functions and then use them to estimate the original functions. In the next subsection we will study performance of this algorithm for a simulated dataset.

4.1 Simulation Study for Nonparametric Additive Regression : "Variable Selection" and "Decision of monotonicity"

As noted earlier in this Section, we have decomposed every nonparametric components into two increasing functions and then will be using LISO type penalized regression scheme (Fang and Meinshausen [2011]), which simultaneously choose the significant components and decide about the monotonicity of the components. Now we want to try these algorithm on a simulated data set where we have increasing, decreasing and non-monotone nonparametric components as siginificant nonparametric components. We consider here a high-dimensional problem with sample size n=100 and number of covariates p=50.

While simulating the data, we choose three increasing components ($2(X_i^{(1)})^3$, $X_i^{(2)}$, and $sign(X_i^{(3)})|X_i^{(3)}|^{1/5}$), one decreasing components ($-(X_i^{(5)})^3$) and one non-monotone component ($(X_i^{(4)})^3 - 4X_i^{(4)}$). We generate $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^n$ following the setup,

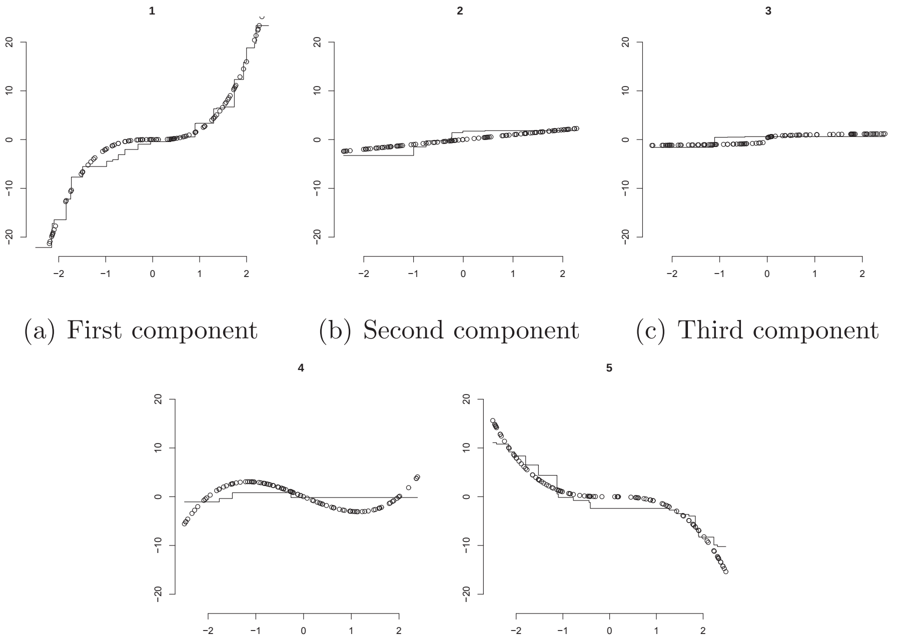
$$\begin{aligned} X_i^{(j)} &\sim Uniform(-2.5, 2.5) \\ \epsilon_i &\sim N(0,1) \\ Y_i &= 2(X_i^{(1)})^3 + X_i^{(2)} + sign(X_i^{(3)})|X_i^{(3)}|^{1/5} + ((X_i^{(4)})^3 - 4X_i^{(4)}) - (X_i^{(5)})^3 + \epsilon_i \end{aligned}$$

Here we choose λ by cross-validation and use the LISO algorithm on this data set. Studying results of the estimated nonparametric components, we see it had chosen only the five significant components correctly and all other components are zero. We plot the estiamted and true nonparametric components of those five significant components in Figure (4.1(a)-4.1(e)).

4.2 Discussion

Consideration of theoretical work regarding model selection consistency for the algorithm will be a future direction of our work. We can also extend these algorithms for median regression. In Section 4.3, we give an weighted version of PAVA-scheme with backfitting algorithm which gives us the solution for the additive median regression under monotonicity constraints. And then the similar Winsorized PAVA estimate with this new estimate for median regression can give us the solution for $\lambda \geq 0$, $\widehat{m_\lambda} = (\widehat{m}_{j,\lambda})_{j=1}^J$, minimiser of

$$L_\lambda(m_1, \ldots, m_J) := \frac{1}{2}|Y - \sum_{j=1}^J m_j(W_j)| + \lambda \sum_{j=1}^J \Delta(m_j) \tag{4.5}$$



(a) First component (b) Second component (c) Third component

(d) Fourth component (e) Fifth component

Figure 4.1. Estimated (dotted-line) and True (continuous-line) values of five significant Nonprametric Components - (a), (b), (c), (d) and (e)

where $m_j \in M_j$. Now as in the suggested algorithm, we can extend the algorithm for the nonparametric components with bounded total variation for median regression also.

4.3 Nonparametric Additive Median Regression under Monotonicity Constraints

In this Section, we consider nonparametric additive median regression models under monotonicity constraints. In practice, the median regression is very common. For example, the growth curve of children as a function of age is usually described using the median. In some practical cases, these median curves have the deterministic monotonic trend for physical reasons, e.g., the growth curve model. Another example is found in pharmacocynetics: the medians of the medicine eliminated by urine as a function of time are also increasing functions of time.

We first consider the d-variate nonparametric regression model without the additive structure in Section 4.3.1. Under the assumption that the error distribution has median zero, we show that the convergence rate of the nonparametric estimate decreases rapidly as the dimension d of the covariate vector increases. This conjectured "curse of dimensionality" motivates us to consider the following nonparametric additive regression model:

$$Y = \sum_{j=1}^{d} m_{j0}(W_j) + \epsilon, \tag{4.6}$$

where $W_j \in \mathbb{R}^1$ and each function m_{j0} is assumed to be monotone. We observe i.i.d. data $(Y_i, W_{i1}, \dots, W_{id})$ for $i = 1, \dots, n$. For simplicity, we assume that the error term ϵ is independent of the covariate vector $\boldsymbol{W}$, where $\boldsymbol{W} = (W_1, \dots, W_d)'$. The model (4.6) has found wide applications in econometrics and epidemiology areas and also covers the possibility of using a (known) link function; see Bacchetti [1989], Morton-Jones et al. [2000]. We will apply the least absolute deviation criterion to estimate the model (4.6) under the minimal smoothness assumption on m_j. When $d = 1$, Cryer et al. [1972] showed that the resulting nonparametric estimate is a monotone step function, i.e., the isotonic median estimate, and can be explicitly expressed using the min-max formula. One major advantage of this isotonic approach stems from the fact that shape constraints automatically "regularize" the estimation problem *without penalization or kernel smoothing*. In fact, we can efficiently compute the isotonic median estimate by the pool adjacent violators algorithm without tuning any smoothing parameter; see Robertson et al. [1988]. The point-wise limit distribution of the isotonic median estimator for $d = 1$ has been derived by Wright [1984], Wang and Huang [2002].

The purpose of this paper is to investigate the asymptotic behaviors of the backfitting estimate that is based on the iterative application of the pool adjacent violator algorithm to the additive components of the model (4.6). A comprehensive introduction to the backfitting estimation procedure can be found in Hastie and Tibshirani [1990]. In Section 4.3.2, we state our main results that each additive component is estimated as well as it would be (by the isotonic median estimate) as if the other components were known, so called *oracle property*. Specifically, our backfitting estimate is shown to have the cubic rate of convergence and non-Gaussian limit distribution, which does not depend on the number of components. In the same model (4.6), the above oracle property has also been shown by Mammen and Yu [2007] who apply the least square estimation criterion, and by Lee et al. [2010] who apply the kernel estimation to each additive component. Note that the smooth kernel estimate in the latter paper has distinct asymptotic behaviors, i.e., root-n consistent and asymptotically normal. Simulations for the comparisons of the backfitting estimator with the oracle estimator are presented in Section 4.3.3. In the end, our Section 4.3.4 briefly discusses the possible extensions to the semiparametric additive quantile regression. All the proofs are postponed to the Appendix. As far as we are aware, our work is the first one that considers the nonparametric additive regression models without regularization under the *non-smooth* criterion function.

4.3.1 Non-additive Estimation

In this section, we consider the nonparametric median regression without the additive structure:

$$Y = M_0(\boldsymbol{W}) + \epsilon, \tag{4.7}$$

where M_0 belongs to a class of uniformly bounded multivariate functions non-decreasing in each coordinate of $\boldsymbol{W} \in [0,1]^d$, denoted as $\mathcal{M}_d$. Note that M_0 is an isotonic (order preserving) function with respect to the partial order " $\ll$ " in $\mathbb{R}^d$ defined as follows: $U \ll V$ if and only if $U_j \leq V_j$, where U_j and V_j are the j^{th} element of U and V, respectively. The requirement that M_0 has the same monotonic direction in each coordinate can be easily satisfied by the change of sign. Under the monotonicity constraint, the non-additive median estimator is defined as

$$\widehat{M} = \operatorname*{arg\,min}_{M \in \mathcal{M}_d} \sum_{i=1}^{n} |Y_i - M(\boldsymbol{W}_i)|. \tag{4.8}$$

The solution of (4.8) is well defined and uniquely determined since the class of isotonic functions forms a closed convex cone. Our Theorem 4 below derives the convergence rate of $\widehat{M}$ for arbitrarily high dimension d and discovers that its rate of convergence decreases rapidly as d increases. This result leads us to conjecture that the curse of dimensionality also exists for the median regression.

Two primary Conditions A1 & A2 on the error distribution are assumed.

A1. Conditional on $\boldsymbol{W}_i$, the error ϵ_i has median zero and a sub-exponential tail, i.e.,

$$E[\exp(|\epsilon_i|/C_0)|\boldsymbol{W}_i] \leq C_0 \text{ a.s.}$$

for some $C_0 > 0$. In addition, we require the error distribution to be symmetric around 0.

A2. Assume that $\mu \mapsto E|\epsilon + \mu|$ is twice differentiable around its point of minimum $\mu = 0$.

The sub-exponential tail condition requires that the tail of ϵ is lighter than that of exponential distribution but may be heavier than that of Gaussian distribution. The above Conditions A1 and A2 hold for various error distributions. A typical example is the family of generalized error distributions with location parameter zero and shape parameter between 1 and 2.

Theorem 4 *Assuming Conditions A1 and A2, we have*

$$\left\|\widehat{M} - M_0\right\|_2 = O_P(n^{-1/3}\log n) \quad \text{for } d = 1, \tag{4.9}$$

$$\left\|\widehat{M} - M_0\right\|_2 = O_P(n^{-1/4}(\log n)^2) \quad \text{for } d = 2, \tag{4.10}$$

$$\left\|\widehat{M} - M_0\right\|_2 = O_P(n^{-1/(4d-4)}\log n) \quad \text{for } d \geq 3, \tag{4.11}$$

where $\|\cdot\|_2$ denotes the L_2 norm.

The convergence rates given in (4.9) – (4.11) are just the upper bounds. We next conjecture that those rates are essentially sharp. It is well known that the convergence rate of nonparametric estimate depends on the entropy number of the corresponding function classes. Therefore, our conjecture follows from the tightness of the entropy bounds for $\mathcal{M}_d$ given in Lemma 6 together with the phase change for the entropy between $d = 1$ and 2; see Remark 5.2 of Gao and Wellner [2007]. The above conjectured curse of dimensionality will be avoided by imposing the additive structure for the d-variate function M. This will require implementing the back-fitting estimation procedure as described in Section 4.3.2.

4.3.2 Additive Monotone Median Regression

Back-fitting Estimation Procedure

The results in Section 4.3.1 motivate us to consider the nonparametric additive regression model (4.6). We will perform the back-fitting procedure to obtain the monotone median estimate for each additive component. To guarantee the uniqueness of the backfitting estimate, we assume that $\int_0^1 m_j(w_j)dw_j = 0$ for any $j = 1, \ldots, d$.

Below, we first introduce a L-statistic type estimate for the population median, and then modify the isotonic median estimate introduced in Cryer et al. [1972] accordingly. Let $J(u)$ be a smooth weight function defined on $[0, 1]$ with $\int_0^1 J(u)du = 1$ and define $\theta(F) = \int_0^1 J(u)F^{-1}(u)du$, where F is any continuous distribution function such that the above integral is well defined. Note that $\theta(F)$ corresponds to the median of F when $J(u)$ is symmetric by 1/2 and the distribution determined by F is symmetric; see Chapter 22 in van der Vaart [2000]. Given that Y follows the distribution F, we can estimate the median of F by

$$\widehat{\theta}(F) = \sum_{i=1}^{n} H_{i,n} Y_{(i)}, \quad \text{where } H_{i,\cdot} = J(i/\cdot)/\cdot$$

rather than by the sample median. In other words, we express $\widehat{\theta}(F)$ as a linear combination of all the order statistic, so called L-statistics. The L-statistic $\widehat{\theta}(F)$ is known to be consistent and has smaller variance than the sample median; see Harrell and Davis [1982]. We next modify the min-max formula given in Cryer et al. [1972] for the univariate isotonic median estimate as follows:

$$\widehat{M}(w) = \max_{0\leq r\leq w} \min_{w\leq s\leq 1} \sum_{i=1}^{N(A)} H_{i,N(A)} Y_{(i),A}, \tag{4.12}$$

where $Y_{(i),A}$ denotes the i-th order statistic in the set $A \equiv \{i' : r \leq W_{i'} \leq s\}$ and $N(A)$ is the cardinality of A. The above formulation will facilitate our asymptotic analysis for the backfitting estimate; see Section 4.3.2.

In the backfitting procedure, we estimate each nonparametric additive component iteratively. For $l = 1, 2, \ldots$, the l-th iterative estimate $\widehat{m}_j^l$ is obtained by minimizing,

$$\sum_{i=1}^{n} |\widehat{Y}_{ij}^l - m_j(W_{ij})|, \tag{4.13}$$

where $\widehat{Y}_{ij}^l = Y_i - \sum_{k\neq j} \widehat{m}_k^{l-1}(W_{ik})$, under the monotonicity constraint. The initial estimate $\widehat{m}_j^0$ is pre-determined. As seen in the simulation Section 4.3.3, the final estimate is very robust to the choice of the initial estimate. According to (4.12), we can compute each backfitting estimate as

$$\widehat{m}_j^l(w_j) = \max_{0\leq r\leq w} \min_{w\leq s\leq 1} \sum_{i=1}^{N(A)} H_{i,N(A)} \widehat{Y}_{(i)j,A}^l, \tag{4.14}$$

where $\widehat{Y}_{(i)j,A}^l$ denotes the i-th order statistic of $\{\widehat{Y}_{i'j}^l : i' \in A\}$.

Asymptotic Analysis for the Back-fitting Estimate

In theory, we define the back-fitting estimate as

$$\widehat{m}_j(w_j) = \max_{0\leq r\leq w_j} \min_{w_j\leq s\leq 1} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widehat{Y}_{(i)j,A_j}, \tag{4.15}$$

where $A_j \equiv \{i : r \leq w_{ij} \leq s\}$ and $\widehat{Y}_{(i)j,A_j}$ denotes the i-th order statistic of $\{\widehat{Y}_{i'j} \equiv Y_{i'} - \sum_{k\neq j} \widehat{m}_k(W_{i'k}) : i' \in A_j\}$. Obviously, $\widehat{m}_j$ can be viewed as $\lim_{l\rightarrow\infty} \widehat{m}_j^l$.

In this section, we will derive the point-wise asymptotic distribution of the backfitting estimate by showing its asymptotic equivalence to the oracle estimate defined as follows

$$\widehat{m}_j^{OR}(w_j) = \max_{0 \le r \le w_j} \min_{w_j \le s \le 1} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widetilde{Y}_{(i)j,A_j}, \tag{4.16}$$

where $\widetilde{Y}_{(i)j,A_j}$ denotes the i-th order statistic of $\{\widetilde{Y}_{i'j} \equiv Y_{i'} - \sum_{k \ne j} m_{k0}(W_{i'k}) : i' \in A_j\}$. Leurgans [1982] showed that $\widehat{m}_j^{OR}$ is actually the slope of the Greatest Convex Minorant (GCM) of the following weighted cumulative sum process $Z_n(w)$ for $w \in (0,1)$:

$$\begin{aligned} Z_n(w + l/n) &= \sum_{j=1}^{l} J(j/(l+1)) \widetilde{Y}_{(j),A_l}/n \ \text{ for } l > 0, \\ Z_n(w + l/n) &= -\sum_{j=1}^{-l} J(j/(-l+1)) \widetilde{Y}_{(j),A'_l}/n \ \text{ for } l < 0, \\ Z_n(w) &= 0, \end{aligned}$$

where $[nw]$ is the smallest integer greater than or equal to nw, $A_l = \{[nw] + 1 \le i \le [nw] + l\}$ and $A'_l = \{[nw] + l + 1 \le i \le [nw]\}$. Define

$$\sigma_j^2(w_j) = \int\int J(F(x|w_j)) J(F(y|w_j)) F(\min(x,y)|w_j)(1 - F(\max(x,y)|w_j)) dx dy,$$

where $F(\cdot|w_j)$ is the c.d.f. of Y conditional on $W_j = w_j$. Given that $\sigma_j(w_j) > 0$ and the first derivative $\dot{m}_{j0}(w_j) > 0$, Leurgans [1982] further gave the following limit distribution for $w_j \in (0,1)$:

$$\left(\frac{2n}{\sigma_j^2(w_j) \dot{m}_{j0}(w_j)} \right)^{\frac{1}{3}} \left(\widehat{m}_j^{OR}(w_j) - m_{j0}(w_j) \right)$$

weakly converges to the slope of GCM of $B(t) + t^2$ at origin, where $B(t)$ is a two-sided Brownian motion, also known as Chernoff's distribution (Chernoff [1964]). The isotonic median estimate based on the sample median was shown to have the same cubic rate of convergence but with different scaling constant in Wright [1984]. Hence, the large-sample relative efficiencies are determined by the multiplicative constants.

These comparisons are not discussed here, since they are the same as those for the ordinary one-sample location problem.

Our main Theorem 5 below will show that the backfitting estimate $\widehat{m}_j(w_j)$ shares the same limit distribution as the oracle estimate $\widehat{m}_j^{OR}(w_j)$ for $w_j \in (0,1)$. We assume that the density function of $\boldsymbol{W}_i$ is continuous and bounded away from 0 and ∞ and that the positive weight function $J(\cdot)$ belongs to some Hölder ball of the order $3/2 < \eta \leq 2$. For example, we can take $J(u)$ as the density for the truncated normal with mean 1/2 and variance 1, i.e.,

$$J(u) = \frac{1}{\sqrt{2\pi(.3829)}} \exp\left\{-\frac{1}{2}\left(u - \frac{1}{2}\right)^2\right\} 1\{0 \leq u \leq 1\}.$$

We also assume the following regularity conditions.

A3. The true functions m_{j0}'s are differentiable with bounded derivatives and satisfy

$$\inf_{|u-v|\geq\delta, 1\leq j\leq d} |m_{j0}(v) - m_{j0}(u)| \geq C_1\delta^\gamma$$

for some constants C_1, $\gamma > 0$ and any $\delta > 0$.

A4. The density function p_{W_k,W_j} of (W_{ik}, W_{ij}) fulfils the following Lipschitz condition

$$\sup_{0\leq u_j, u_k, v_k \leq 1} |p_{W_k,W_j}(u_k, u_j) - p_{W_k,W_j}(v_k, u_j)| \leq C_2|u_k - v_k|^\rho$$

for some constants $C_2, \rho > 0$.

Conditions A3 and A4 imply the strict monotonicity of each m_{j0} and the upper bound of $|p_{W_k|W_j}(u_k|u_j) - p_{W_k|W_j}(v_k|u_j)|$ as $O(|u_k - v_k|^\rho)$, respectively.

Theorem 5 *Suppose that Conditions A1-A4 hold. We show the following asymptotic equivalence relation:*

$$\sup_{n^{-\frac{2}{9}} \leq u_j \leq 1 - n^{-\frac{2}{9}}} |\widehat{m}_j(w_j) - \widehat{m}_j^{OR}(w_j)| = o_P(n^{-\frac{1}{3}}),$$

which directly implies that, for $w_j \in (0,1)$,

$$\left(\frac{2n}{\sigma_j^2(w_j)\dot{m}_{j0}(w_j)}\right)^{\frac{1}{3}} (\widehat{m}_j(w_j) - m_{j0}(w_j)) \tag{4.17}$$

weakly converges to the well known Chernoff's distribution.

To make point-wise inferences on m_{j0}, we still need to estimate the unknown normalization constant in (4.17), which is very challenging especially for the estimation of $\dot{m}_{j0}$, even though Chernoff's distribution is well tabulated in Groeneboom and Wellner [2001]. Inspired by Banerjee and Wellner [2001], we may form a likelihood ratio test statistic and find its asymptotic distribution. By doing so, we can obtain the confidence interval by inverting the constructed likelihood ratio which avoids the need to estimate messy nuisance parameters. An alternative approach will be the subsampling approach (Politis and Romano [1994]).

4.3.3 Simulation Studies

We conduct Monte Carlo simulations to evaluate the finite sample performance of the proposed backfitting estimate. We consider the model:

$$Y = m_{10}(W_1) + m_{20}(W_2) + \epsilon, \tag{4.18}$$

where (W_1, W_2) has the truncated bivariate normal distribution on $[-1, 1]^2$ with correlation parameter ρ and ϵ follows the generalized error distribution with the shape parameter 1.5. The non-decreasing functions are assumed to be $m_{10}(w_1) = \sin(\pi w_1/2)$ and $m_{20}(w_2) = w_2^3$.

We implement the back-fitting algorithm described in subsection 4.3.2 for various sample sizes ranging from 200 to 800. For each sample size, 100 dataset were analyzed. In each setting, we iterate 500 times for the backfitting algorithm. According to our simulation experiences, our results are very robust to the choice of initial estimate. In Table 1, we calculate the empirical mean integrated squared error (MISE) of the back-fitting estimate $\widehat{m}_j$ and the oracle estimate $\widehat{m}_j^{OR}$, and give the ratio of the above two empirical MISEs. From Table 1, we notice that the backfitting and oracle estimator both have very small MISE and the ratio converges to one as sample size increases even under strong correlation between W_1 and W_2. In Figure 1, we plot the true function, oracle estimate and backfitting estimate for m_1 when $n = 800$, and the estimators achieving 25%, 50% and 75% quantiles of the L_2-distance between the back-fitting and oracle estimate. We observe that the backfitting and the oracle estimator produce almost identical curves.

Table 4.1
Comparison between the backfitting and the oracle estimator. Model (4.18) with $m_1(x) = \sin(\pi x/2)$, $m_2(x) = x^3$, sample size 200, 400, 800 and different values of ρ for covariate distribution

		m_1		
n	ρ	Backfitting	Oracle	B/O
200	0	0.0657	0.0692	0.950
	0.5	0.0631	0.0641	0.979
	-0.5	0.0589	0.0594	0.992
	0.7	0.0728	0.0630	1.153
	-0.7	0.0597	0.0620	0.952
400	0	0.0356	0.0352	1.012
	0.5	0.0393	0.0384	1.026
	-0.5	0.0366	0.0360	1.008
	0.7	0.0454	0.0401	1.134
	-0.7	0.0389	0.0401	0.972
800	0	0.0236	0.0237	0.995
	0.5	0.0234	0.0231	1.013
	-0.5	0.0220	0.0221	0.996
	0.7	0.0262	0.0233	1.125
	-0.7	0.0238	0.0234	1.020

In Table 2, we consider the case that one of the nonparametric components is not smooth. Here, $m_{10}(x) = \sin(\pi x/2)$ and $m_{20}(x) = x$ for $|x| > 0.5$; 0.5 for $0 \leq x \leq 0.5$; -0.5 for $-0.5 \leq x \leq 0$. Even in this case the backfitting estimator shows a quite good

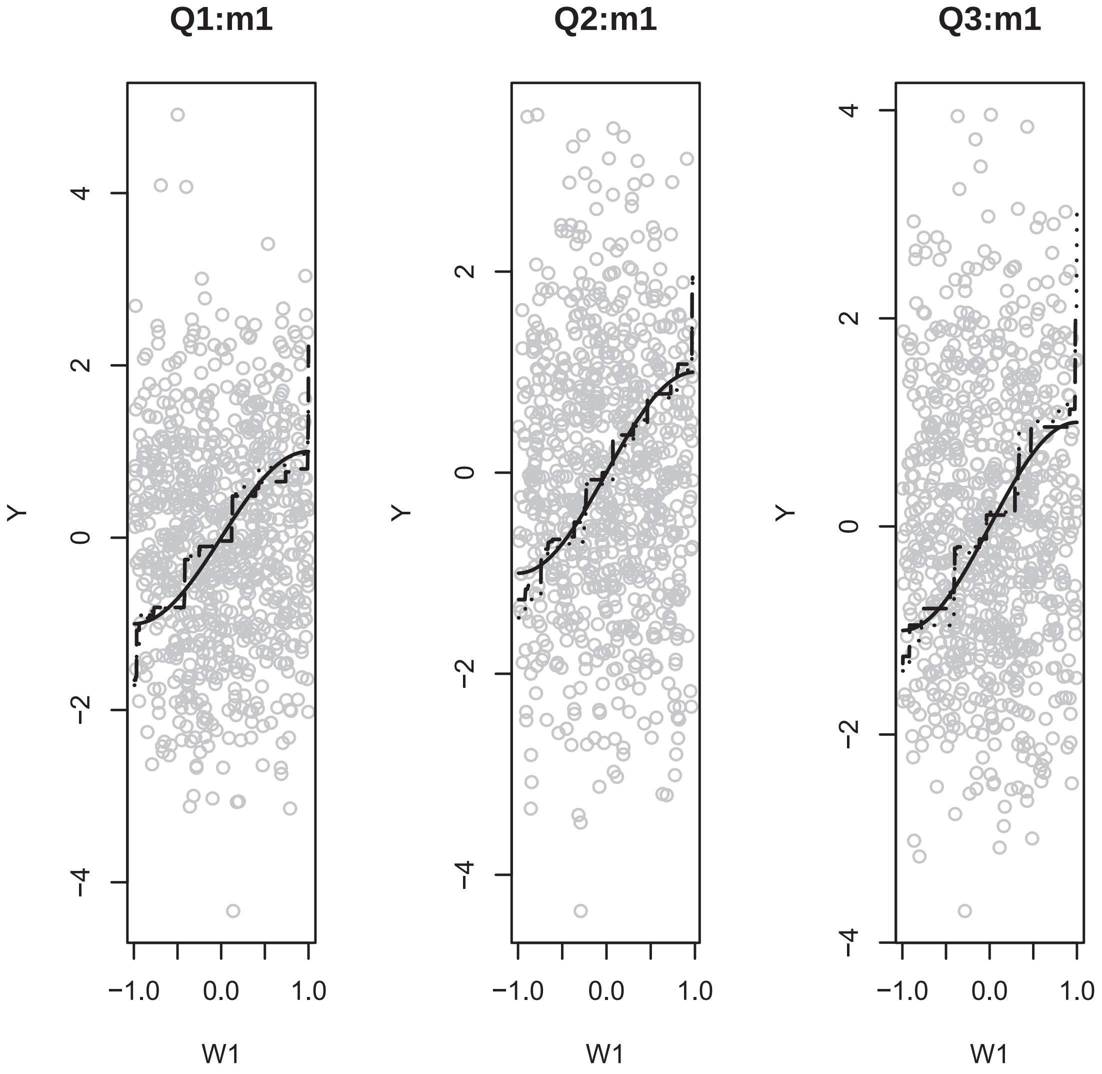


Figure 4.2. The real lines, dashed lines and dotted lines show the true curve, backfitting estimates and oracle estimates, respectively. From left to right, fitted curves for the data sets that produce 25%, 50% and 75% quantiles for the distance between the backfitting and the oracle estimator in Monte Carlo simulations with $\rho = 0.5$ and 800 observations.

performance. Hence, the oracle property of the additive isotonic median regression is strongly supported by the simulations.

Table 4.2
Comparison between the backfitting and the oracle estimator. Model (12) with $m_1(x) = \sin(\pi x/2)$, $m_2(x) = x, |x| > 0.5; 0.5, 0 \leq x \leq 0.5; -0.5, -0.5 \leq x \leq 0$, sample size 200, 400, 800 and different values of ρ for covariate distribution

		m_2		
n	ρ	Backfitting	Oracle	B/O
200	0	0.0678	0.0633	1.072
	0.5	0.0739	0.0664	1.114
	-0.5	0.0684	0.0613	1.117
	0.7	0.0747	0.0625	1.196
	-0.7	0.0679	0.0661	1.028
400	0	0.0421	0.0401	1.050
	0.5	0.0418	0.0383	1.093
	-0.5	0.0417	0.0402	1.038
	0.7	0.0472	0.0362	1.305
	-0.7	0.0389	0.0391	0.994
800	0	0.0240	0.0231	1.039
	0.5	0.0251	0.022	1.14
	-0.5	0.0246	0.0245	1.006
	0.7	0.0300	0.0231	1.301
	-0.7	0.0252	0.0231	1.080

4.3.4 Discussions

It is a natural idea to extend the current median regression to the quantile regression. However, in this case, the practical choice of the weight function $J(\cdot)$ is nontrivial; see the discussions in Harrell and Davis [1982]. Furthermore, it is also meaningful to extend the current model to the semiparametric additive models by incorporating a parametric term:

$$Y = \mathbf{X}'\beta + \sum_{j=1}^{d} m_{j0}(W_j) + \epsilon.$$

The above semiparametric modelling is particularly useful when $\mathbf{X}$ is a dummy variable. A similar backfitting algorithm can be developed by iterating between a cyclic pool adjacent violators procedure and solving a linear quantile regression; see Cheng [2009] for similar discussions.

LIST OF REFERENCES

LIST OF REFERENCES

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transaction of Automatic Control*, 19(6):716–723, 1974a.

Hirotugu Akaike. Factor analysis and aic. *Psychometrika*, 52(3):317–332, 1974b.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley, 2nd edition, 1984.

C. Andrieu, A. Doucet, and C. P. Robert. Computational advances for and from bayesian analysis. *Statistical Science*, 19(1):118–127, 2004.

P. Bacchetti. Additive isotonic models. *Journal of American Statistical Association*, 84:289–294, 1989.

M. Banerjee and J.A. Wellner. Likelihood ratio tests for monotone functions. *Annals of Statistics*, 29:1699–1731, 2001.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897, 2004.

D. J. Bartholomew, F. Steele, I. Moustaki, and J. I. Gabbrith. *The analysis and interpretation of multivariate data for social scientists.* Chapman and Hall, 2002.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B*, 57:289–300, 1995.

J. O. Berger and L. Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of American Statistical Association*, 91:109–122, 1996.

J. O. Berger and L. Pericchi. Training samples in objective bayesian model selection. *Annals of Statistics*, 32(3):841–869, 2004.

J. O. Berger, J. K. Ghosh, and N. Mukhopadhyay. Approximations and consistency of bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, 112(1):241–258, 2003.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

M.S. Birman and M.Z. Solomjak. Piecewise-polynomial approximations of functions of the class w_p^α. *Mathematics of the USSR Sbornik*, 73:295–317, 1967.

M. Bogdan, J. K. Ghosh, and R. W. Doerge. Modifying the schwarz bayesian information criterion to locate multiple interacting quantitive trait loci. *Genetics*, 167:989–999, 2004.

M. Bogdan, P. Biecek, R. Cheng, F. Frommlet, J.K. Ghosh, and R. W. Doerge. Extending the modified bayesian information criterion (mbic) to dense markers and multiple interval mapping. *Biometrics*, 64(4):1162–1169, 2008a.

M. Bogdan, J.K. Ghosh, and M. Żak-Szatkowska. Selecting explanatory variables with the modified version of bayesian information criterion. *Quality and Reliability Engineering International*, 24:627–641, 2008b.

M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. The bayes oracle and asymptotic optimality of multiple testing procedures under sparsity. *Annals of Statistics*, 39:1551–1579, 2011.

F. Bunea, M. H. Wegkamp, and A. Auguste. Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, 136:4349–4364, 2006.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

O. Bunke and X. Milhaud. Asymptotic behavior of bayes estimates under possibly incorrect models. *Annals of Statistics*, 26(2):617–644, 1998.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

A. Chakrabarti. *Model Selection for High Dimensional Problems with Application to Function Estimation*. PhD thesis, Purdue University, 2004.

A. Chakrabarti. Some aspects of bayesian model selection for prediction. In *Bayesian Statistics 8*, pages 83–84. Oxford University Press, 2007.

A. Chakrabarti and J. K. Ghosh. Aic, bic and recent advances in model selection, philosophy of statistics. In *Handbook of the philosophy of Science*. Elsevier, 2011.

M. H. Chen, Q. M. Shao, and J. G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer, 2000.

G. Cheng. Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference*, 139:1980–1991, 2009.

H. Chernoff. General estimation of the mode. *Annals of Statistics*, 136:31–41, 1964.

S. Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

T. Choi and R.V. Ramamoorthi. Remarks on consistency of posterior distributions. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, pages 170–186. Institute of Mathematical Statistics, Beachwood, Ohio, USA, bertrand clarke and subhashis ghosal, eds. edition, 2008.

M. Clyde and E. I. George. Model uncertainty. *Statistical Science*, 19(1):81–94, 2004.

D. R. Cox. Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symp.*, pages 105–123, 1961.

D. R. Cox. Further results on tests of separate families of hypothesis. *Journal of Royal Statistical Society, Series B*, 24:406–424, 1962.

J.D. Cryer, T. Robertson, F.T. Wright, and R.J. Casady. Monotone median regression. *Annals of Mathematical Statistics*, 43:1459–1469, 1972.

S. C. Dass, , and J. Lee. A note on the consistency of bayes factors for testing point null versus non-parametric alternatives. *Journal of Statistical Planning and Inference*, 119(1):143–152, 2004.

T. J. DiCiccio, R. E. Kass, A. Raftery, and L. Wasserman. Computing bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92(439):903–915, 1997.

D. Draper and M. Krnjajic. Calibration results for bayesian model specification. *Bayesian Analysis*, 1(1):1–43, 2010.

M. Drton. Likelihood ratio tests and singularities. *Annals of Statistics*, 37(2):979–1012, 2009.

R. Dutta and J. K. Ghosh. Bayes model selection with path sampling: Factor models and other examples. *Statistical Science*, 2011a. Accepted pending minor correction.

R. Dutta and J. K. Ghosh. Some observations on novel statistical issues in analysis of high dimensional problems of inference about genes. *Journal of the Indian Society of Agricultural Statistics*, 65(2):205–212, 2011b.

R. Dutta, M. Bogdan, and J. K. Ghosh. Model selection and multiple testing - a bayes and empirical bayes overview and some new results. *Journal of Indian Statistical Association*, 50, 2011. Accepted for publication.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Y. Fan, R. Wu, M. Chen, L. Kuo, and P.O. Lewis. Choosing among partition models in bayesian phylogenetics. *Molecular Biological Evolution*, 28(1):523–532, 2011.

Z. Fang and N. Meinshausen. Lasso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, Accepted for publication:1–20, 2011.

C. Fernandez, E. Ley, and M. F. Steel. Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100:381–427, 2001.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of Royal Statistical Society, Series B*, 70:589–607, 2008.

F. Frommlet, F. Ruhaltinger, P. Twarog, and M. Bogdan. A model selection approach to genome wide association studies. *Computational Statistics and Data Analysis*, 2011. doi:10.1016/j.csda.

D. Gamarnik and Y. Shah, D. andWei. Belief propagation for min-cost network flow: Convergence & correctness. In *Proceedings of SODA'2010*, pages 279–292, 2010.

D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC Texts in Statistical Science, 2nd edition, 2006.

F. Gao and J.A. Wellner. Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98:1751–1764, 2007.

S. Geisser. The predictive sample reuse method with applications. *Journal of American Statistical Association*, 70:320–328, 1975.

A. E. Gelfand and D. Dey. Bayesian model choice: asymptotic and exact calculations. *Journal of Royal Statistical Society, Series B*, 56:501–514, 1994.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3:515–534, 2006.

A. Gelman and X. L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 2003.

J. Ghosh and D. B. Dunson. Bayesian model selection in factor analytic models. In *Random Effect and Latent Variable Model Selection*. John Wiley and Sons, d.b. dunson edition, 2008.

J. Ghosh and D. B. Dunson. Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320, 2009.

J. K. Ghosh and T. Samanta. Nonsubjective bayes testing - an overview. *Journal of Statistical Planning and Inference*, 103:205–223, 2002.

J. K. Ghosh and K. Subramanyam. Inference about separated families in large samples. *Sankhya, Series-A*, 37(4):502–513, 1975.

J. K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis - Theory and Methods*. Springer, 2006.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. Groeneboom and J.A. Wellner. Computing chernoff's distribution. *Journal of Computational and Graphical Statistics*, 10:388–400, 2001.

F. E. Harrell and C. E. Davis. A new distribution-free quantile estimator. *Biometrika*, 69:635–640, 1982.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

K. Ito. Functions of bounded variation. In *Encyclopedic Dictionary of Mathematics*, chapter 166, pages 642–643. Cambridge, MA:MIT Press, 1993.

H. Jeffreys. *Theory of Probability.* Oxford University Press, 1961.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32:1594–1649, 2004.

M. R. Kosorok. *Introduction to Empirical Process and Semiparametric Inference.* Springer, 2008.

C. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publications of Statistics*, 2:125–141, 1955.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–412, 2010.

N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *System Biology*, 55:195–207, 2006.

S. Y. Lee and X. Y. Song. Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*, 29:23–40, 2002.

Y. K. Lee, E. Mammen, and B. U. Park. Backfitting and smooth backfitting for additive quantile models. *Annals of statistics*, 38(5):2857–2883, 2010.

G. Lefebvre, R. Steele, A. C. Vandal, S. Narayanan, and D. L. Arnold. Path sampling to compute integrated likelihoods: An adaptive approach. *Journal of Computational and Graphical Statistics*, 18(2):415–437, 2009.

S. Leurgans. Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Annals of Statistics*, 10(1):287–296, 1982.

K.C. Li. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixture of g priors for bayesian variable selection. *Journal of American Statistical Association*, 103(481): 410–423, 2008.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing.* Springer, corrected edition, 2008.

H. F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67, 2004.

S. M. Lynch. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, 2007.

E. Mammen and K. Yu. Additive isotone regression. *IMS Lecture NotesMonograph Series Asymptotics: Particles, Processes and Inverse Problems*, 55:179–195, 2007.

X. L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

E. Moreno, F. J. Giron, and G. Casella. Consistency of objective bayes factors as the model dimension grows. *Annals of Statistics*, 38:1937–1952, 2010.

T. Morton-Jones, P. Diggle, L. Parker, H.O. Dickinson, and K. Binks. Additive isotonic regression models in epidemiology. *Statistics in Medicine*, 19:849–859, 2000.

N. Mukhopadhyay. *Bayesian model selection for high dimensional models with prediction error loss and 0-1 loss.* PhD thesis, Department of Statistics, Purdue Univeristy, 2000.

N. Mukhopadhyay and J. K. Ghosh. Parametric empirical bayes model selection - some theory, methods and simulation. *IMS Lecture Notes - Krishna Atreya et. al. eds. Probability, statistics and their applications: papers in honor of Rabi Bhattacharya*, pages 229–245, 2003.

N. Mukhopadhyay, J. K. Ghosh, and J. O. Berger. Some bayesian predictive approaches to model selection. *Statistics and Probability Letters*, 73:369–379, 2005.

W. Murray, P. E. Gill, and M. H. Wright. *Practical Optimization.* Academic Press, London, 1981.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

F. B. Nielsen. Variational approach to factor analysis and related models. Master's thesis, The Institute of Informatics and Mathematical Modelling, Technical University of Denmark, 2004.

A. O'Hagan. Fractional bayes factors and model comparison. *Journal of Royal Statistical Society, Series B*, 57:99–138, 1995.

T. Park and G. Casella. The bayesian lasso. *Journal of American Statistical Association*, 103(482):681–687, 2008.

D. N. Politis and J. P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22:2031–2050, 1994.

A. E. Raftery, M. A. Newton, J. Satagopan, , and P. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In *Bayesian Statistics*, volume 8, pages 1–45. Oxford University Press, j.m. bernardo, et al. edition, 2007.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, 2nd edition, 2004.

T. Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference.* Wiley, NewYork, 1988.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619, 2010.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

R. Shibata. Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, 71:43–49, 1994.

X. Y. Song and S. Y. Lee. Model comparison of generalized linear mixed models. *Statistics in Medicine*, 25:1685–1698, 2006.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society : Series B*, 36(2):111–147, 1974.

K. Subramanyam. *Some asymptotic properties of maximum likelihood procedures.* PhD thesis, Indian Statistical Institute, 1979.

M. Thain, M. Hickman, M. Abercrombie, C. J. J. Hickman, N. I. Johnson, and R. Turvey. *The Penguin Dictionary of Biology.* Penguin, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Series B*, 58(1):267–288, 1996.

S. T. Tokdar, A. Chakrabarti, and J. K. Ghosh. Bayesian nonparametric goodness of fit. In *Frontier of Statistical Decision making and Bayesian Analysis.* M.h. chen, d.k. dey, p. muller, d. sun and k. ye edition, 2010.

S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, 2000.

S. Van de Geer. A bound for the empirical risk minimizer. In *Oberwolfach reports*, volume 49, pages 2989–2992. 2006.

A. W. van der Vaart. Bracketing smooth functions. *Stochastic Processes and their Aplications*, 52:93–105, 1994.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, 2000.

Y. Wang and J. Huang. Limiting distribution for monotone median regression. *Journal of Statistical Planning and Inference*, 107:281–287, 2002.

F. T. Wright. The asymptotic behavior of monotone percentile regression estimates. *The Canadian Journal of Statistics*, 12(3):229–236, 1984.

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and Chen M. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Syst. Biol.*, 60(2):150–160, 2011.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2007.

APPENDICES

A. OTHER METHODS

A.1 Importance Sampling

Suppose we have two densities proportional to two functions $f(x)$ and $g(x)$, which are feasible to evaluate at every x, but one of the distributions, say the one induced by f(x), is not easy to sample. Then the importance sampling (IS) estimate of the ratio of normalizing constants is based on m independent draws $x_1, \ldots, x_m$ generated from the distribution defined by g(x). We first compute the importance weights $w_i = \frac{f(x_i)}{g(x_i)}$ and then define the IS estimate:

$$\frac{1}{m}\sum_{i=1}^{m} w_i. \tag{A.1}$$

Under the assumption that $g(x) \neq 0$ when $f(x) \neq 0$, $\frac{1}{m}\sum_{i=1}^{m} w_i$ converges as $m \to \infty$ to Z_f/Z_g, when $Z_f = \int f(x)dx$ and $Z_g = \int g(x)dx$ are the normalizing constants for $f(x)$ and $g(x)$. The variability of the IS estimate depends heavily on the variability of the weight functions. So to have a good IS estimate we need to have $g(x)$ as a good approximation to $f(x)$, which is difficult to achieve in problems with high or moderately high dimensional, possibly multimodal density.

Analysis of Bayesian factor models using IS has been introduced by Ghosh and Dunson [2008]. The IS estimator of BF for factor models is based on m samples $\theta_i^{(h)}$ from the posterior distribution, under $M^{(h)}$

$$\widehat{BF}_{h-1,h} = \frac{1}{m}\sum_{i=1}^{m} \frac{p(y|\theta_i^{(h)}, k = h-1)}{p(y|\theta_i^{(h)}, k = h)} \tag{A.2}$$

which in turn is based on the following identity:

$$\begin{aligned}\int \frac{p(y|\theta^{(h)}, k = h-1)}{p(y|\theta^{(h)}, k = h)} p(\theta^{(h)}|y, k = h) d\theta^{(h)} \\ = \int p(y|\theta^{(h)}, k = h-1) \frac{p(\theta^{(h)})}{p(y|k = h)} d\theta^{(h)} \\ = \frac{p(y|k = h-1)}{p(y|k = h)}.\end{aligned} \tag{A.3}$$

Ghosh and Dunson [2008] implemented IS with a parameter expanded prior. They also have noted that IS is fast and often(90%) chooses the correct model in simulation. In our simulation IS chooses a true bigger model correctly, but a 20% error rate was observed when the smaller model is true.

A.2 Annealed Importance Sampling

Following Neal [2001] we consider densities $p_t : t \in [0, 1]$ joining the densities p_0 and p_1. We choose densities by discretising the path $p_{t_{(i)}}$ where $0 = t_{(1)} < \ldots < t_{(k)} = 1$ and then simulate a Markov chain designed to converge to $p_{t_{(k)}}$. Starting from the final states of the previous simulation we simulate some number of iterations of a Markov chain designed to converge to $p_{t_{(k-1)}}$. Similarly we simulate some iterations starting from the final steps of $p_{t_{(j)}}$ designed to converge to $p_{t_{(j-1)}}$ until we simulate some iterations converging to $p_{t_{(1)}}$. This sampling scheme produces a sample of points $x_1, \ldots, x_m$ and then we compute the weights $w_i = \frac{p_1(x_i)}{p_0(x_i)}$. Then the estimate of the ratio of normalizing constant becomes as follows:

$$\frac{1}{m} \sum_{i=1}^{m} w_i. \tag{A.4}$$

Notice that while both AIS and PS are based on MCMC runs along a path from one model to another while the MCMC'S are drawn at each point, but the details are very different. Due to the better spread of MCMC samples, the estimates in AIS seem to be better than those calculated by IS when the smaller model is true, helping in correct model selection and also improving the estimation of Bayes Factors. However, simulations show that AIS has the same problem as IS in estimating the Bayes Factor when the bigger model is true.

A.3 BIC type methods : Raftery-Newton and our method using Information Matrix

In contrast to the methods previously discusssed, we try to directly estimate the marginal under each model and then use these marginals to find the Bayes Factor. We know that BIC is an approximation to the log-marginal based on a Laplace-type approximation of the log-marginal Ghosh et al. [2006], under the assumption of i.i.d. observations. Thus

$$log(m(x)) \approx log(f(x|\hat{\theta})\pi(\hat{\theta})) + (p/2)log(2\pi) + (p/2)log(n) + log(|H_{1,\hat{\theta}}^{-1}|^{1/2}) \tag{A.5}$$

where $H_{1,\hat{\theta}}$ is the observed Fisher Information matrix evaluated at the maximum likelihood estimator using a single observation. For BIC we just use

$$\begin{aligned} log(m(x)) &\approx log(f(x|\hat{\theta})\pi(\hat{\theta})) + (p/2)log(n) \\ &\approx log(f(x|\hat{\theta}))) + (p/2)log(n) \end{aligned} \tag{A.6}$$

ignoring other terms as they are O(1).

It is known BIC may be a poor approximation to the log-marginal in high dimensions (Berger et al. [2003]). To take care of this problem, Raftery et al. [2007] suggest the following. Simulate iid MCMC samples from the posterior distributions, evaluate independent sequence of log(prior×likelihood)s (log-p.l.) $\{l_t : t = 1, \ldots, m\}$, and then an estimate for the marginal is

$$log(m(x)) \approx \bar{l} - s_l^2(log(n) - 1) \tag{A.7}$$

where $\bar{l}$ and s_l^2 will be sample mean and variance of l_t's. We call this method BICM, following the convention of Raftery et al. [2007].

In order to apply A.5, we do not need to evaluate n since it cancels by combining the last two terms. This suggests the approximation A.5 take care of the point raised by Clyde and George [2004]. However A.7 does use n, but we do not know the impact on the approximation.

We have also used the Laplace approximation (A.5) without any change as likely to work better than the usual BIC. We compute the Information Matrix at the maximum prior×likelihood (mpl) value under the model and impute its value in the computation of marginal. To find the mpl estimate we use the MCMC sample from the posterior distribution and pick the maxima in that sample. Then we search for the mple in its neighbourhood, using it as the starting point for the optimization algorithm. In our simulation study, it has been seen to give very good results similar to the computationaly intensive numerical algorithms used to find the maximum of a function over the whole parameter space seen by taking repeatation of MCMC runs and large MCMC samples. In the spirit of Raftery et al. [2007], we call this method BICIM, indicating the use of Information Matrix based Laplace Approximation. We also used several other modifications that did not give good results, so are not reported.

B. A THEORETICAL REMARK ON THE LIKELIHOOD FUNCTION

It appears that the behavior of the likelihood, e.g. its maximum plays an important role in model selection, specifically in the kind of conflict we see between PS and the Laplace approximations (BICM, BICIM) when the bigger model is true (and the prior is a t with a relatively small d.f.). The behavior seems to be different from the asymptotic behavior of maximum likelihood under the following standard assumptions. Assume dimension of the parameter space is fixed and usual regularity conditions hold. Moreover, when the big model is true but the small model is assumed (so that it is a misspecified model) Kullback-Liebler projection of the true parameter space to the parameter space of the small model exists (Bunke and Milhaud [1998]).

Fact Assume the big model is true, and the small model is false. Then, as may be verified easily by Taylor expansion,

1. logL($\hat{\theta}_{big}$)-logL($\theta_{true(big)}$)=$O_P(1)$
2. logL($\hat{\theta}_{small}$)-logL(KL projection of $\theta_{true(big)}$ to Θ_{small})=$O_P(1)$
3. logL($\theta_{true(big)}$)-logL(KL projection of $\theta_{true(big)}$ to Θ_{small})=$O_P(n)$

   and
4. $$\begin{aligned} &logL(\hat{\theta}_{big}) - logL(\hat{\theta}_{small}) \\ &\quad = logL(\theta_{true(big)}) - logL(KL\, projection\, of\, \theta_{true(big)}\, to\, \Theta_{small}) + O_P(1) \\ &\quad = O_P(n) \end{aligned}$$

The maximized likelihood for Factor models substantially over estimates the true likelihood unlike relation (1) above. Unfortunately, as pointed out in Drton [2009] the asympyotics of mle for Factor Models is still not fully worked out.

C. MATRIX USED FOR THE TOY EXAMPLE

$$\Sigma^0 = \begin{pmatrix} 128.35 & 52.69 & -19.25 & -11.86 & 24.34 & 8.80 & 10.63 & 13.75 & -7.40 & -29.80 \\ 52.69 & 73.37 & -21.04 & -37.85 & 12.29 & 8.74 & 15.60 & 12.09 & -14.08 & -17.27 \\ -19.25 & -21.04 & 30.86 & 8.63 & -1.41 & -13.58 & -3.03 & -11.64 & 21.28 & 22.05 \\ -11.86 & -37.85 & 8.63 & 80.49 & 4.66 & 3.26 & -49.24 & -9.68 & 22.18 & 8.52 \\ 24.34 & 12.29 & -1.41 & 4.66 & 15.45 & 2.58 & 2.05 & 3.72 & -1.31 & -7.87 \\ 8.80 & 8.74 & -13.58 & 3.26 & 2.58 & 31.37 & 11.62 & -4.85 & -16.89 & -20.10 \\ 10.63 & 15.60 & -3.03 & -49.24 & 2.05 & 11.62 & 58.09 & 7.00 & -19.58 & 5.16 \\ 13.75 & 12.09 & -11.64 & -9.68 & 3.72 & -4.85 & 7.00 & 26.59 & -3.04 & 11.17 \\ -7.40 & -14.08 & 21.28 & 22.18 & -1.31 & -16.89 & -19.58 & -3.04 & 31.81 & 22.86 \\ -29.80 & -17.27 & 22.05 & 8.52 & -7.87 & -20.10 & 5.16 & 11.17 & 22.86 & 64.68 \end{pmatrix}$$

D. CHOICE OF PRIOR UNDER M_0

A referee has asked whether under M_0, the prior for extra parameter can be chosen in same optimal or a philosophically compelling manner. This has been a long-standing problem but the method followed for Factor Models is one of the standard procedures, apparently first suggested by Edwards et. al. (1963).

This prior is mentioned by Edwards et. al. (1963) and may be justified as follows. One tries to ensure the extra parameter has similar roles under both the models. If the joint prior of (θ_1, θ_2) under M_1 is $\pi(\theta_1, \theta_2)$, then the natural prior for $(\theta_2|\theta_1)$ is the usual conditional density of $\pi(\theta_2|\theta_1)$. In our case $\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$. So $\pi(\theta_2|\theta_1)$ is as we have chosen. This is one of the standard default choices. Another default choice is due to Jeffreys (1961), but when θ_1, θ_2 are independent both lead to the same choice. If we introduce a prior (e.g., minimizing MCMC-variance) it may not be acceptable to Bayesian philosophy.

E. PROOFS FOR CHAPTER 4

Lemma 6 is a simple adaptation of Corollary 1.3 in Gao and Wellner [2007]. Lemmas 7 and 8 will be used in the proof of Theorem 2. Denote $\log N_{[]}(\epsilon, \mathcal{F}, d)$ as the ϵ-bracketing entropy number (measured by d) for the function class $\mathcal{F}$. Let $L_p(Q)$ be the L_p norm w.r.t. the probability measure Q.

Lemma 6 *Suppose Q is a probability measure on $[0,1]^d$ with Lebesgue density q satisfying*

$$1/M \le \inf_{x\in[0,1]^d} q(x) \le \sup_{x\in[0,1]^d} q(x) \le M$$

for some $M > 0$. Then, if $(d-1)p \neq d$, we have

$$K_2\epsilon^{-\alpha} \le \log N_{[]}(\epsilon, \mathcal{M}_d, L_p(Q)) \le K_1\epsilon^{-\alpha}, \tag{E.1}$$

where $\log N_{[]}(\epsilon, \mathcal{M}_d, L_p(Q))$ is the ϵ-bracketing entropy number and $\alpha = \max\{d, (d-1)p\}$. If $(d-1)p = d$, then

$$K_2\epsilon^{-d} \le \log N_{[]}(\epsilon, \mathcal{M}_d, L_p(Q)) \le K_1\epsilon^{-d}(\log(1/\epsilon))^d \tag{E.2}$$

for constants K_1 and K_2 depending only on d, p and M.

Lemma 7 *Under Conditions A1 – A4, we show that*

$$\sup_{n^{-2/9}\le u_j\le 1-n^{-2/9}} |\widehat{m}_j(u_j) - m_{j0}(u_j)| = O_P(n^{-2/9}(\log n)^{2/3}).$$

Proof Theorem 1 implies that

$$\left\| \sum_{j=1}^d \widehat{m}_j - \sum_{j=1}^d m_{j0} \right\|_2 = O_P\left(n^{-1/3}\log n\right).$$

Since we assume the density of $\boldsymbol{W}_i$ to be bounded from zero, we have

$$\int_0^1 [(\widehat{m}_1(w_1) - m_{10}(w_1) + \ldots + \widehat{m}_d(w_d) - m_{d0}(w_d)]^2 dw = O_P[(\log n)^2 n^{-\frac{2}{3}}].$$

Now, under the identifiability condition

$$\int_0^1 m_j(w_j)dw_j = 0$$

we can conclude

$$\max_{1\leq j\leq d}\int_0^1 [\widehat{m}_j(w_j) - m_{j0}(w_j)]^2 dw_j = O_P[(\log n)^2 n^{-\frac{2}{3}}]. \tag{E.3}$$

Without loss of generality, we consider some interval of $[0,1]$, denoted by $\mathcal{I}$, in which the step function $\widehat{m}_j$ is flat. For any $u \in \mathcal{I}$, we define

$$e(u) = \widehat{m}_j(u) - m_{j0}(u).$$

Denote the upper bound of the first derivative of m_{j0} as c. When $e(u) > 0$, we can establish the inequality

$$0 \geq (m_{j0}(v) - \widehat{m}_j(v)) \geq (m_{j0}(u) - \widehat{m}_j(u)) - c(u - v) \tag{E.4}$$

for any $v \in \mathcal{I}$ smaller than u by the mean value theorem. Similarly, when $e(u) < 0$, for any $v \in \mathcal{I}$ larger than u, we have

$$0 \geq (\widehat{m}_j(v) - m_{j0}(v)) \geq (\widehat{m}_j(u) - m_{j0}(u)) - c(v - u). \tag{E.5}$$

Considering (E.4) & (E.5), we have

$$|\widehat{m}_j(v) - m_{j0}(v)| \geq |\widehat{m}_j(u) - m_{j0}(u)| - c|v - u| = |e(u)| - c|v - u|.$$

For any v satisfying $|v - u| < |e(u)|/(2c)$, we have $|\widehat{m}_j(v) - m_{j0}(v)| > c|v - u|$, which implies

$$(\widehat{m}_j(v) - m_{j0}(v))^2 \geq c^2(v - u)^2. \tag{E.6}$$

Based on (E.6), we have

$$O_P[(\log n)^2 n^{-\frac{2}{3}}] = \int_0^1 [\widehat{m}_{j0}(w_j) - m_{j0}(w_j)]^2 dw_j \quad \geq \quad c^2 \int_{-\frac{|e(u)|}{2c}}^{\frac{|e(u)|}{2c}} t^2 dt = \frac{|e(u)|^3}{12c^3},$$

which implies the statement of Lemma 7. ■

□

Lemma 8 *Define $\Delta(s,t) = J(s)g(t)$ over $[0,1]^2$, where the positive function $J(\cdot)$ belongs to some Hölder ball of the order $\eta \geq 1$ and $g(\cdot)$ is the difference of two uniformly bounded monotone functions. Let $Z_1, \ldots, Z_k$ be a triangular array of independent random vectors with values in $[0,1]^2$. Then it holds that uniformly over all functions Δ*

$$\sum_{i=1}^{k} (\Delta(Z_i) - E[\Delta(Z_i)]) = O_P(k^{\frac{2}{3}}), \tag{E.7}$$

$$\sum_{i=1}^{l} (\Delta(Z_i) - E[\Delta(Z_i)]) = O_P(k^{\frac{2}{3}}) \quad \textit{uniformly for } l \leq k. \tag{E.8}$$

Proof We will apply Lemma 5.13 of van de Geer [2000] to prove (E.7) which trivially implies (E.8). We first calculate the δ-bracketing entropy in terms of $L_2(P)$ norm for the class of functions $\{\Delta\}$. Following van der Vaart [1994], we know that the δ-bracketing entropy (in terms of L_2 norm) for the class of smooth functions $\{J(\cdot)\}$ is $O(\delta^{-1/\eta})$. For the difference of two monotone functions, the δ-bracketing entropy in terms of infinite norm is $O(\delta^{-1})$; see Birman and Solomjak [1967]. Hence, Lemma 9.25 of Kosorok [2008] implies that the δ-bracketing entropy for $\{\Delta\}$ is of the order $O(\delta^{-1} \vee \delta^{-1/\eta}) = O(\delta^{-1})$. In addition, Δ is uniformly bounded. Then (E.7) trivially follows from Lemma 5.13 of van de Geer [2000] (by taking $\alpha = 1$ and $\beta = 0$). ■

E.0.1 Proof of Theorem 4

Let $\boldsymbol{w}_{(i)}$ be the observation of $\boldsymbol{W}_{(i)}$ where $\boldsymbol{W}_{(i)} = \{\boldsymbol{W} \in [0,1]^d : M(\boldsymbol{W}_{(i)}) = (M(\boldsymbol{W}))_{(i)}\}$. We first show that

$$\max\{|\widehat{M}(\boldsymbol{w}_{(1)})|, |\widehat{M}(\boldsymbol{w}_{(n)})|\} = O_P(\log n). \tag{E.9}$$

Without loss of generality, we only prove $|\widehat{M}(\boldsymbol{w}_{(n)})| = O_P(\log n)$. Let $\mathcal{L}_d$ be the collection of subset L in $\mathbb{R}^d$ having the property that if $U \in L$ and $U \ll V$ then

$V \in L$. Following Robertson et al. [1988] by an extension of equation (4.12) for $d > 1$, we have

$$\widehat{M}(\boldsymbol{w}_{(n)}) = \max_{\{L:\boldsymbol{w}_{(n)} \in L \subset \mathcal{L}_d\}} \left(\sum_{i=1}^{N(A)} H_{i,N(A)} y_{(i),A} \right)$$

where $A = \{i' : \boldsymbol{w}_{i'} \in L, 1 \leq i' \leq n\}$. Hence, we can establish the following set of inequalities,

$$\begin{aligned} |\widehat{M}(\boldsymbol{w}_{(n)})| &\leq \max_{\{L:\boldsymbol{w}_{(n)} \in L \subset \mathcal{L}_d\}} \left(\sum_{i=1}^{N(A)} |H_{i,N(A)}||y_{(i),A}| \right) \\ &\leq \max_{\{L:\boldsymbol{w}_{(n)} \in L \subset \mathcal{L}_d\}} \left(\sum_{i=1}^{N(A)} |H_{i,N(A)}||M_0(\boldsymbol{w}_{(i),A})| \right) + \max_{\{L:\boldsymbol{w}_{(n)} \in L \subset \mathcal{L}_d\}} \left(\sum_{i=1}^{N(A)} |H_{i,N(A)}||\epsilon_{(i),A}| \right) \\ &\leq C + |\epsilon|_{(n)}, \end{aligned}$$

for some constant $C < \infty$, where $\boldsymbol{w}_{(i),A}$ and $\epsilon_{(i),A}$ is the observation corresponding to $y_{(i),A}$. Considering the sub-exponential tail of ϵ, we have proven (E.9).

Assumption A2 implies that

$$P(|\epsilon - (M - M_0)(\boldsymbol{w})| - |\epsilon|) \gtrsim (M - M_0)^2(\boldsymbol{w}), \tag{E.10}$$

where " $\gtrsim$ " means greater than up to an universal constant. Considering (E.10) and the definition of $\widehat{M}$, we obtain the following inequality

$$\begin{aligned} P\left(|Y - \widehat{M}(\boldsymbol{W})| - |Y - M_0(\boldsymbol{W})|\right) &\leq (\mathbb{P}_n - P)\left(|Y - M_0(\boldsymbol{W})| - |Y - \widehat{M}(\boldsymbol{W})|\right) \\ \|\widehat{M} - M_0\|_2^2 &\lesssim (\mathbb{P}_n - P)\left(|Y - M_0(\boldsymbol{W})| - |Y - \widehat{M}(\boldsymbol{W})|\right), \end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of $(Y_i, \boldsymbol{W}_i)$ and " $\lesssim$ " means smaller than up to an universal constant. Define

$$\Im_n = \{r_n|y - M_0(\boldsymbol{w})| - |r_n y - G(w)| : G \in \mathcal{M}_d\}$$

where $r_n = (\log n)^{-1}$. Note that $r_n||\widehat{M}||_\infty = O_P(1)$, where $||\cdot||_\infty$ is the uniform norm, based on (E.9). We next study the δ-bracketing entropy of $\Im_n$ in terms of $L_2(P)$-norm. And we know that δ-bracketing entropy of $\mathcal{M}_d$ is of the order $1/\delta$ for $J = 1$,

$\delta^{-2}(\log(1/\delta))^2$ for $J = 2$, and $\delta^{-(2J-2)}$ for $J > 2$ based on the above lemma A.1. Since the function in $\Im_n$ is Lipschitz continuous in G, δ-bracketing entropy of $\Im_n$ is the same as that of $\mathcal{M}_d$. In view of the discussions in page 326 of Van der Vaart and Wellner (1996), we derive the convergence rate that $\|\widehat{M} - M_0\|_2 = O_P(\delta_n/r_n)$ based on Theorem 3.4.1 and Lemma 3.4.2 in Van der Vaart and Wellner (1996), where δ_n is determined by

$$\sqrt{n}\delta_n^2 \geq \tilde{J}_{[]}(\delta_n, \Im_n, L_2(P)) \left(1 + \frac{\tilde{J}_{[]}(\delta_n, \Im_n, L_2(P))}{\sqrt{n}\delta_n^2}\right),$$

and

$$\tilde{J}_{[]} = \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \Im_n, L_2(P))} d\epsilon \quad \text{if} \quad J = 1,$$
$$\tilde{J}_{[]} = \int_{c\delta^2}^\delta \sqrt{1 + \log N_{[]}(\epsilon, \Im_n, L_2(P))} d\epsilon \quad \text{if} \quad J \geq 2.$$

□

E.0.2 Proof of Theorem 5

Recall that the oracle estimate and backfitting estimate are defined as, respectively,

$$\begin{aligned} \widehat{m}_j^{OR}(w_j) &= \max_{0 \leq r \leq w_j} \min_{w_j \leq s \leq 1} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widetilde{Y}_{(i)j,A_j}, \\ \widehat{m}_j(w_j) &= \max_{0 \leq r \leq w_j} \min_{w_j \leq s \leq 1} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widehat{Y}_{(i)j,A_j}. \end{aligned}$$

We define the following localized version of $\widehat{m}_j$ as

$$\begin{aligned} \widehat{m}_{j,loc}(w_j) &= \max_{w_j - e_n \leq r \leq w_j} \min_{w_j \leq s \leq w_j + e_n} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widehat{Y}_{(i)j,A_j}, \\ \widehat{m}_{j,loc}^{+}(w_j) &= \max_{w_j - e_n \leq r \leq w_j} \min_{w_j + d_n \leq s \leq w_j + e_n} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widehat{Y}_{(i)j,A_j}, \\ \widehat{m}_{j,loc}^{-}(w_j) &= \max_{w_j - e_n \leq r \leq w_j - d_n} \min_{w_j \leq s \leq w_j + e_n} \sum_{i=1}^{N(A_j)} H_{i,N(A_j)} \widehat{Y}_{(i)j,A_j}, \end{aligned}$$

where $e_n = (\log n)^{\frac{1}{\gamma}} n^{-\frac{2}{9\gamma}} c_n$ with $c_n \to \infty$ slowly enough and $d_n = n^{-\delta}$, and $1/3 < \delta < 4/9$. Localized oracle estimate is defined similarly by replacing $\widehat{Y}_{(i)j,A_j}$ with $\widetilde{Y}_{(i)j,A_j}$ in the above equations.

The above definitions imply that, for $n^{-\frac{1}{3}} \leq w_j \leq 1 - n^{-\frac{1}{3}}$,

$$\widehat{m}^{-}_{j,loc}(w_j) \leq \widehat{m}_{j,loc}(w_j) \leq \widehat{m}^{+}_{j,loc}(w_j),$$
$$\widehat{m}^{OR,-}_{j,loc}(w_j) \leq \widehat{m}^{OR}_{j,loc}(w_j) \leq \widehat{m}^{OR,+}_{j,loc}(w_j).$$

According to Lemma 7, Condition A3 and representations of $\widehat{m}_j$ and $\widehat{m}_{j,loc}$, we have $\widehat{m}_j(w_j) = \widehat{m}_{j,loc}(w_j)$ for $0 \leq w_j \leq 1$ with probability tending to one. Similarly, we also have $\widehat{m}^{OR}_j(w_j) = \widehat{m}^{OR}_{j,loc}(w_j)$ for $0 \leq w_j \leq 1$ with probability tending to one. Therefore, we can conclude that

$$\widehat{m}^{-}_{j,loc}(w_j) \leq \widehat{m}_j(w_j) \leq \widehat{m}^{+}_{j,loc}(w_j), \tag{E.11}$$
$$\widehat{m}^{OR,-}_{j,loc}(w_j) \leq \widehat{m}^{OR}_j(w_j) \leq \widehat{m}^{OR,+}_{j,loc}(w_j). \tag{E.12}$$

for $n^{-\frac{1}{3}} \leq w_j \leq 1 - n^{-\frac{1}{3}}$ with probability tending to one. Following the properties of the isotonic quantile estimators (Robertson et al. [1988]), we have

$$\sup_{0 \leq w_j \leq 1} \widehat{m}^{OR,+}_{j,loc}(w_j) - \widehat{m}^{OR,-}_{j,loc}(w_j) = o_P(n^{-\frac{1}{3}}). \tag{E.13}$$

Define the index set $A(u_j)$ as $A(u_j) = \{i : W_{ij} \leq u_j\}$. We now consider

$$\widehat{S}_j(u_j, w_j) = \frac{1}{n} \sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j)))\widehat{Y}_{(i)j,A(u_j)} - \frac{1}{n} \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j)))\widehat{Y}_{(i)j,A(w_j)}$$
$$\widehat{S}^{OR}_j(u_j, w_j) = \frac{1}{n} \sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j)))\widetilde{Y}_{(i)j,A(u_j)} - \frac{1}{n} \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j)))\widetilde{Y}_{(i)j,A(w_j)}.$$

For $w_j - e_n \leq u_j \leq w_j + e_n$, we consider the functions that map $N(A(u_j))$ onto $\widehat{S}_j(u_j, w_j)$ or $\widehat{S}^{OR}_j(u_j, w_j)$ respectively. According to Leurgans [1982], we know that $\widehat{m}_{j,loc}(w_j)$ and $\widehat{m}^{OR}_{j,loc}(w_j)$ are the slopes of the GCM of these functions at $u_j = w_j$. Hence, we have

$$\begin{aligned}
& \widehat{S}_j(u_j, w_j) - \widehat{S}_j^{OR}(u_j, w_j) \\
= \ & \frac{1}{n}(\sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j)))(\widehat{Y}_{(i)j,A(u_j)} - \widetilde{Y}_{(i)j,A(u_j)}) \\
- \ & \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j)))(\widehat{Y}_{(i)j,A(w_j)} - \widetilde{Y}_{(i)j,A(w_j)})) \\
= \ & \frac{1}{n}\sum_{l \neq j}(\sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j)))(m_{l0} - \widehat{m}_l)(W_{(i)l,A(u_j)}) \\
- \ & \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j)))(m_{l0} - \widehat{m}_l)(W_{(i)l,A(w_j)})), \qquad \text{(E.14)}
\end{aligned}$$

where $W_{(i)l,A}$ is the i-th order statistics of the set $\{W_{i'l} : 1 \leq i' \leq n \text{ and } i' \in A\}$. We next apply Lemma 8 to analyze (E.14). Conditional on $W_{1j}, \ldots, W_{nj}$, we set $Z_i = (i/N(A), W_{(i)l,A})'$ and $\Delta(Z_i) = I\{n^{-\frac{2}{9}} \leq u \leq 1 - n^{-\frac{2}{9}}\}J(i/N(A))[m_{l0}(W_{(i)l,A}) - \widehat{m}_l(W_{(i)l,A})]/(n^{-\frac{2}{9}}(\log n)^{2/3})$ based on Lemma 7, where $A = A(u_j)$ or $A(w_j)$. Hence, we can further simplify (E.14) as

$$\begin{aligned}
& \widehat{S}_j(u_j, w_j) - \widehat{S}_j^{OR}(u_j, w_j) \\
= \ & \frac{1}{n}\sum_{l \neq j}\left[\sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j))) - \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j))) \right] \\
& \left(\int_0^1 (m_{l0}(w_l) - \widehat{m}_l(w_l)) p_{W_l|W_j}(w_l|W_{ij}) dw_l \right) + \epsilon_1 \\
= \ & \frac{1}{n}\sum_{l \neq j}\left[\sum_{i=1}^{N(A(u_j))} J(i/N(A(u_j))) - \sum_{i=1}^{N(A(w_j))} J(i/N(A(w_j))) \right] \\
& \left(\int_0^1 (m_{l0}(w_l) - \widehat{m}_l(w_l)) p_{W_l|W_j}(w_l|w_j) dw_l \right) + \epsilon_1 + \epsilon_2,
\end{aligned}$$

where $\epsilon_1 = O_P(|u_j - w_j|^{2/3} n^{-11/9} (\log n)^{2/3})$ and $\epsilon_2 = O_P(|u_j - w_j| n^{-2\rho/9\gamma} (\log n)^{2/3})$ based on Conditions A3 – A4. So now using (E.11), (E.12), the above equation and the conditions on $J(\cdot)$, we can conclude that,

$$\widehat{m}_j^{\pm}(w_j) = \widehat{m}_j^{OR,\pm}(w_j) - \sum_{l \neq j} \int_0^1 (m_{l0}(w_l) - \widehat{m}_l(w_l)) p_{W_l|W_j}(w_l|w_j) dw_l + o_P(n^{-\frac{1}{3}}).$$

Considering (E.13), we prove the following:

$$\widehat{m}_j(w_j) = \widehat{m}_j^{OR}(w_j) - \sum_{l \neq j} \int_0^1 (m_{l0}(w_l) - \widehat{m}_l(w_l)) p_{W_l|W_j}(w_l|w_j) dw_l + o_P(n^{-\frac{1}{3}}). \quad \text{(E.15)}$$

The remaining proof is the same as that of Theorem 1 in Page 191 of Mammen and Yu [2007]. □

VITA

VITA

Ritabrata Dutta was born on September 8, 1983 in Calcutta, WB India. He received his B.Stat. and M.Stat. from the Indian Statistical Institute in 2005 and 2008 respectively. He has been a graduate student at Purdue University since 2008.