Bayesian Regression and Discrimination with Many Variables

by

Kai-Ming Chang

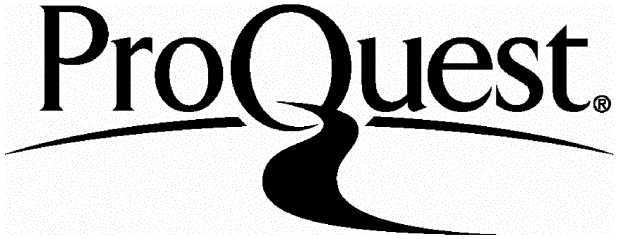
Department of Statistical Science

University College London

Thesis submitted for the degree of Doctor of Philosophy

in the University of London

September 2001



Abstract

This thesis attempts to provide general procedures for Bayesian regression and discriminant analysis with many variables and explore potential problems in the analysis. For regression analysis, a normal random regression model is assumed, i.e. the joint distribution of the response variables and the regressors is multivariate normal given their means and covariance matrix. For the discriminant analysis, we consider the case that each observation is from one of several multivariate normal populations. In classical statistics, the problem in fitting a multivariate model with more variables than the number of observations is that the estimate of the covariance matrix of the multivariate normal distribution is singular and the distribution is degenerate. In Bayesian statistics, this problem can be avoided by using proper prior assumptions for the covariance matrix. We assign an inverse-Wishart distribution (which is a conjugate prior in the case of a non-hierarchical analysis) for the covariance matrix and suppose the prior expected covariance matrix has a simple structure so that the number of hyperparameters required in the model is small. Hierarchical modelling of these hyperparameters is employed.

Although we have managed to keep the model relatively simple with our strong assumptions, the posterior model is still complicated. We found ARMS within Gibbs sampling with multiple chains to be an appropriate MCMC strategy for fitting our models. Convergence checking for multiple chains MCMC is simple. Due to the ill-condition of the sample covariance matrix and the large number of variables, the computational problems are significant. Appropriate matrix manipulating and rescaling techniques are required.

Two practical cases are considered as examples, one for regression and the other for discrimination. Both cases involve NIR spectral data with many variables. The high correlation between variables makes the examples more challenging. We consider three correlation structures including the over-simplified identity structure and two autoregressive correlation functions, which are believed to be

much closer to the real situation than the over-simplified one. However, we found the autoregressive correlation functions do not guarantee better predictions in our examples.

Acknowledgements

Firstly, I would like to thank my supervisour Prof Tom Fearn for his excellent supervision and patient guidance. The experience of being his student is simply invaluable. I would also like to express my appreciation to many staff in my department, especially to Prof A. Philip Dawid and Dr Richard E. Chandler, who have been so helpful whenever I needed opinions. Special acknowledgements go to my examiners Prof Paul Garthwaite and Prof Sylvia Richardson for their useful comments and opinions on my thesis.

I would not forget to thank Dr Suja M. Aboukhamseen and other colleagues in the research student room for their suggestions, encouragement and friendship. In particular, I have to thank Ms Anne Andrew for her help to improve my poor English writing.

My deepest gratitude goes to my parents. Their great love and full support have helped me throughout every difficulty.

Contents

List of Figures

List of Tables

Notation

0.1 Matrices

I: Identity matrix.

I_p: a p by p identity matrix.

Let X be an arbitrary matrix:

X^t: the transpose of matrix X.

$X_{p\times q}$: an alternative notation for X in which $p \times q$ indicates the dimension of X.

Suppose X is a square matrix:

$X > 0$: X is positive definite.

$X \geq 0$: X is positive semi-definite.

tr$X$: the trace of X.

$|X|$: the determinant of X.

X^{-1}: the inverse matrix of X.

0.2 Probability

$|$: conditional on or given.

$\perp\!\!\!\perp$: conditionally independent.

$E(X)$: Expectation of X.

$\mathcal{C}(X)$: Covariance matrix of a column vector X, $\mathcal{C}(X) = E(XX^t) - E(X)E(X)^t$.

$P(D)$: probability measure of event D.

$p(X)$: probability density or mass function of random variable X.

0.3 Matrix Distributions

$\mathcal{N}$: matrix normal distribution.

$\mathcal{W}$: Wishart distribution.

$\mathcal{IW}$: inverse-Wishart distribution.

$\mathcal{T}$: matrix-t distribution.

$\mathcal{F}$: matrix-F distribution.

Chapter 1

Introduction

This thesis deals with regression and discriminant analysis with many variables in a Bayesian framework. Regression and discriminant analysis are important multivariate statistical techniques that have been widely applied in many fields. In regression analysis, one attempts to relate two sets of variables with a model so that one set of the variables can be predicted by the other set. In discriminant analysis, one aims to predict which of two or more groups an object belongs to using a model that has as its input a set of variables we observe for the object. If we take the group membership of an object as a categorical variable, we can link the discriminant model to the regression one. Problems arise in fitting models and making predictions when there are many more variables than the number of observations we use to fit the model. Two major problems are the singularity of sample* covariance matrices and overfitting, which are also two general problems in statistics.

The topic of the thesis is motivated by the analysis of near infrared (NIR) transmission spectroscopy, where chemical analysts take the (possibly transformed) NIR absorbances of a sample , e.g. a chemical compound, at certain wavelengths as predictor variables and use these measurements to predict chemical composition of the sample or to classify the sample to one of several groups. Often, the number

*The word 'sample' has two meanings throughout this thesis: the standard statistical one, as in a sample from a population; and the chemical meaning, which is a quantity of some substance presented for analysis. Which of these meanings is intended should be clear from the context.

of wavelengths observed can be up to one thousand or even more. That is, for each sample there can be more than one thousand predictor variables observed. The linear model has been widely accepted in analysing data from NIR spectroscopy. According to the Beer-Lambert law (see chapter 3), an NIR transmission spectrum of a sample is, under ideal conditions, a linear combination of the NIR transmission spectra of the constituents of the sample, and the weight of each linear component would be proportional to the concentration of the corresponding constituent in a sample. The ideal conditions rarely hold in practice, but linear models have been found to work well in most applications. NIR transmission spectra arise from the absorption of light by organic chemical bonds. A chemical bond has absorption peaks at certain wavelengths, which depend on the atoms at the two ends of the bond and on their relationship with other atoms in the molecule. An individual absorption peak has a smooth bell shape. Unfortunately a typical NIR spectrum will contain many thousands of overlapping peaks, and the chemical information we wish to extract will occur in several (not precisely predictable) places and be seriously mixed up with other information. Thus, the simple approach of selecting a small number of relevant wavelengths is not usually appropriate, and it is common to use models where all the spectral variables are taken as predictors.

Standard statistical inferences for regression and discriminant analysis use least squares estimation (LS) or maximum likelihood estimation (MLE) to estimate the parameters in the models. However, in the cases when we are using more variables than samples and the variables cannot be pre-selected in order to reduce the number of variables so that they are less than the number of samples, LS estimation and MLE of parameters are inappropriate because they require inversion of the sample covariance matrix, which is a singular matrix. Many regularised methods aimed at keeping as much information from the predictor variables as possible whilst avoiding the inversion of a singular matrix have been developed. Existing methods for regression analysis include principal components regression (PCR), partial least squares regression (PLSR), ridge regression (RR), and continuum regression (CR). For discriminant analysis, one can also apply principal components

analysis (PCA) to overcome the problem of singularity of the sample covariance matrix. One can refer to Brown [23], Martens and Næs [94] and Osborne, Fearn, and Hindle [100] for further details of these methods as well as their application in NIR analysis.

When using more variables than observations in a model, there is always a risk of overfitting. In the regression case when we have more variables than observations, we can always (unless collinearity means that the variables lie in a subspace of dimension less than the number of observations) find a set of coefficients which fits the data perfectly. In general, when the number of regressors increases, the model always fits the data better. However, the variance of prediction is not always reduced when the number of regressors goes up (Seber [113]). Models that fit data too well usually predict future observations badly. One way to check the models is to use the cross-validation method (Stone [118]), which has been widely employed in many applications. The purpose of cross-validation is to make sure a fitted model can reasonably predict or classify future observations by fitting the model with a training data set and assessing the performance of the fitted chosen model on validation data with an appropriate scoring rule. We use it as a method to assess our models.

This thesis explores the properties of normal regression and discrimination with many variables in a Bayesian framework. The Bayesian approach makes inference by combining prior knowledge of the model with observed data. The inference on the unknown quantities of interest is summarised by a posterior distribution derived using Bayes' rule. Statistical methodology in a Bayesian framework has developed considerably since the 1960's, not only because it provides an easily understood way of summarising results, but also due to progress in computer technology and developments in Markov chain Monte Carlo simulation (see Gilks *et al.* [73]). Because of these advantages, the Bayesian approach is able to deal with very complex models, which may involve many parameters with complicated relationships between them. It is also known that the Bayesian approach does not have a constraint on the number of variables and the number of training sam-

ples. The insufficient rank of the data is made up by the use of prior information. Therefore, it is natural to consider the Bayesian approach for modelling with many variables.

It is well known that prior assumptions become important when there are more variables than observations. A major focus in this thesis is the effect of prior assumptions about the covariance matrix of the predictor variables. Since there are many variables, the covariance matrix is huge. In order to reduce the number of parameters and make the modelling process tractable, we assume there is a simple structure in the covariance matrix, which requires only a few parameters to describe, chosen so that an analytical inverse matrix and determinant are available for computational efficiency. At the same time we try to use realistic prior assumptions, i.e. ones that would generate data resembling those we have observed. The structure of the expected covariance matrix should satisfy the principle of structural coherence defined by Brown [23], that the structure of the expected covariance matrix of a refinement of a random vector should be generated by the same structural consideration that generates the expected covariance matrix of the random vector. A prior distribution is assigned to the parameters in the expected covariance matrix so our models are hierarchical.

For regression analysis, we consider the non-conjugate multivariate normal random regression model which has been conceptually suggested by Mäkeläinen and Brown [93]. The non-conjugate model was proposed in order to avoid the deterministic property of the natural conjugate regression model. Dawid [43] proved that under the normal-inverse-Wishart prior assumption, the natural conjugate regression model with an infinite number of predictor variables can predict the future precisely when the parameters in prior distributions are known. The property is called determinism by Dawid. Other applications of Bayesian regression with many variables either consider information compression (e.g. West [126]) or focus on variable selection with a computationally convenient prior assumption (e.g. Brown *et al.* [25, 28]). We consider modelling with the entire set of regressor variables with more realistic prior distributions, and investigate whether this leads

to improvement in predictive performance.

For discriminant analysis, we assume a multivariate normal distribution within groups and apply Bayes' formula to obtain the posterior group membership probability of an object. Our main focus is on the use of more realistic prior distributions for the variance parameters of a full model with many variables. Posterior predictive probabilities of a sample belonging to different groups are calculated and taken as a criterion for allocating the sample. Recent applications in discrimination for NIR data can be found in Brown *et al.* [24] and Fearn *et al.* [56].

We use two examples of NIR spectroscopic data of wheat samples, one for regression and one for discrimination. In both examples, measurements of the NIR spectrum of each wheat sample are recorded digitally at a hundred equally spaced wavelengths. In the regression example, there are 50 samples in total, while there are in total 292 samples of 9 varieties (groups) for the discrimination problem.

Bayesian theory and stochastic simulation provide the possibility of handling complicated situations and producing easily understandable summaries of the inference results. Although in our examples the predictor variables are very highly correlated, the model we investigate in this thesis would be applicable to situations with less strongly correlated variables. Moreover, the MCMC simulation scheme (ARMS within Gibbs sampling) we employed for fitting our models is a very general method which is very useful in fitting models with many parameters. In practice, data analysts are facing ever more situations where there are many variables in their models. Examples exist in the fields of molecular biology, econometrics, geostatistics, chemometrics, etc. For instance, genetic data analysis is currently a hot topic that involves models with huge number of variables.

Chapter 2

Bayesian Inference for Linear Regression and Discrimination

2.1 Bayesian Theory

The idea of Bayesian inference first appeared in Thomas Bayes' paper in 1763 [6], where he proposed a uniform distribution for the parameter in the binomial distribution. Later, Laplace independently discovered the general form of Bayes' theorem. The idea of Bayes was then largely ignored for two hundred years. During the second half of the 20th century, scientists started to realise the potential of Bayesian methodology and applied Bayes' theory in many areas.

Bayesian inference for an unknown quantity yields a probability distribution, which is derived by combining the observed data with a probability model for the quantities we observe and the unknown quantities about which we want to learn. Bayesian theory is based on a simple and fundamental probability rule

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}, \tag{2.1}$$

where M and D are two random events, $P(\cdot)$ is the probability measure of events and the symbol '|' represents "conditional on". In Bayesian theory, M represents the hypothesis of interest, and D is the evidence, the data we observe. $P(D|M)$ is the probability measure of the events we observe under hypothesis M, and $P(M)$

represents the prior probability of the hypothesis, while $P(D)$ is the marginal probability of the data over all possible hypotheses. Therefore, in Bayesian modelling, we need a *sampling distribution* which the data we observe are assumed to follow, and a *prior distribution* of all possible hypotheses, which formalises our prior belief about the hypotheses. Rule (2.1) provides the distribution of the hypothesis conditional on the data we observe, which is called the *posterior distribution* of the hypothesis. The subject of a hypothesis can be parameters, predictors, or even a model.

For example, suppose X is a continuous random variable from sample space $\mathcal{X}$. Its density function is $p(X|\theta)$, where θ is from a continuous parameter space Θ, and $p(\theta)$ is the prior density function of θ. We observe $x_1, x_2, \ldots, x_n$ for X as training samples to fit (train) our model. The observations are sampled independently with the same density function as X, hence,

$$p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

We can then derive the posterior density function of θ, which is

$$p(\theta|x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{p(x_1, x_2, \ldots, x_n)}, \tag{2.2}$$

where

$$p(x_1, x_2, \ldots, x_n) = \int_\theta p(x_1, x_2, \ldots, x_n|\theta)p(\theta)d\theta = \int_\theta \prod_{i=1}^{n} p(x_i|\theta)p(\theta)d\theta.$$

Now, if we want to predict future m observations of X, we use the predictive density function of X based on the posterior distribution of θ. Suppose we want to predict $x_{n+1}, x_{n+2}, \ldots, x_{n+m}$, which are conditionally independent given θ. The posterior predictive density function is

$$\begin{aligned} & p(x_{n+1}, x_{n+2}, \ldots, x_{n+m}|x_1, x_2, \ldots, x_n) \\ = & \int_\theta p(x_{n+1}, x_{n+2}, \ldots, x_{n+m}|\theta)p(\theta|x_1, x_2, \ldots, x_n)d\theta \\ = & \int_\theta \prod_{i=1}^{m} p(x_{n+i}|\theta)p(\theta|x_1, x_2, \ldots, x_n)d\theta. \end{aligned}$$

Equation (2.1) is called Bayes' formula or Bayes' rule. With Bayes' formula, we update the prior model that is based on our prior knowledge to a posterior model that is conditional on the evidence we observe.

For a more detailed introduction to Bayesian statistics readers may refer to [12], [65], or other Bayesian textbooks. In this chapter, we introduce concepts that are important in Bayesian regression and discriminant modelling with many variables.

2.2 Prior distributions

The prior distribution plays an important role in Bayesian inference. It introduces the expert's knowledge of the unobservable parameters in a model by formalising the expert's opinion about the parameters as prior distributions. However, specifying a prior distribution is not an easy step. The choice of prior distribution for the parameters is probably the most controversial issue in Bayesian statistics.

2.2.1 Conjugate Prior Distributions

In Bayesian statistics, when the joint prior distribution and the joint posterior distribution of parameters in the sampling distribution belong to the same parametric distribution family, the prior distribution is called a conjugate prior distribution for the sampling distribution, and Bayesian analysis with a conjugate prior is called conjugate analysis.

Consider again the example in the previous section. After the model has been set up, the most complicated part of the inference is the integration required to calculate the denominator in (2.2). Since the denominator in (2.2) is a constant, we have

$$p(\theta|x_1, x_2, \ldots, x_n) \propto p(x_1, x_2, \ldots, x_n|\theta)p(\theta).$$

If a posterior distribution belongs to some known parametric distribution family, the integration is actually unnecessary, and the distribution can be identified simply from the product of the likelihood function and prior distribution.

The advantage of using a conjugate prior distribution is not only that integration can be avoided. Since the prior distribution and posterior distribution belong to the same parametric distribution family, the inferential process is just a matter of updating the parameters in the prior distribution so the cost of computing is greatly reduced. The disadvantage is that the choice of prior distribution for a given sampling distribution is very limited, and available conjugate prior distributions may not be adequate to represent our prior opinion. Modelling with non-conjugate prior distributions that are more realistic is in many cases unavoidable. Even so, conjugate analysis has been applied in many practical cases due to the convenience in computing. Examples of conjugate analysis can be found in all Bayesian textbooks.

2.2.2 Non-informative Prior Distributions

It is often the case that the prior information for a parameter in a model is very limited or uncertain. In this case, one would naturally expect to assign a prior distribution that makes little contribution to the posterior distribution of the parameter and 'let the data speak for themselves', i.e. let the data dominate the posterior distribution of the parameter. Non-informative priors (also called vague priors) have been frequently used in Bayesian applications in order to represent prior ignorance.

The simplest type of prior that may represent prior ignorance is Laplace's rule, or the principle of insufficient reason (see Kass and Wasserman [86]). It assumes every value for the parameter is equally possible, i.e.

$$p(\theta) \propto c.$$

Such a prior was first applied by Bayes and Laplace. When the parameter space is bounded, $p(\theta)$ is a uniform distribution. When the parameter space is not bounded, $p(\theta)$ is not a well-defined distribution because its integral over the parameter space is infinite. One problem with using this flat prior is that it is inconsistent under different parameterisations of the same problem. Suppose $p(\theta) \propto c$ is the prior

density function of θ. Let ϕ be a re-parameterisation of θ with $\phi = \exp(\theta)$. The prior density function of ϕ is a multiple of $1/\phi$, which does not follow Laplace's rule of prior ignorance. Jeffreys [82] proposed a procedure for creating prior density functions which are invariant to re-parameterisation in ways that will be described below. The Jeffreys' prior $p(\theta)$ for the parameter θ in a probability model is given by

$$p(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta)$ is the Fisher information matrix of θ. The Jeffreys' prior has been widely used in one dimensional cases. However, its performance in higher dimensional cases is not always satisfactory.

Invariance has been considered to be important in creating non-informative prior density functions. Dawid [42] concludes that there are three types of invariance:

- Parameter invariance: The prior distributions of two models derived under this principle should be equivalent when one model is a re-parameterised version of the other.

- Data invariance: Suppose Y is a transformation of X and that Y and X have a common parameter θ. The prior distributions of θ derived from the distributions of Y and X under this principle should be the same.

- Context invariance: No features of the structure, meaning, or context of a model other than its distribution model should be taken into account in forming an invariant prior.

Jeffreys' prior is an example that satisfies these three principles. Other variations of these principles exist. Dawid [42] and Kass and Wasserman [86] give further coverage of invariance theory.

Alternatives to Jeffreys' prior are available. Some of them are based on minimising the information in the prior distribution. Berger and Bernardo (see [9, 10, 11]) initiated a method which looks for priors that minimise the Kullback-Leibler

distance between the posterior density and the prior density. This produces invariant non-informative priors. Many authors have investigated maximum entropy priors. Jaynes proposed a maximum entropy method that also produces invariant priors. There are many other methods in the literature that can be used to produce non-informative priors. Kass and Wasserman [86] provide a review of formal rules for selecting non-informative prior distributions.

Although there are many methods for creating non-informative priors, most of these priors have improper density functions. There are many problems in using improper priors, especially in modelling with many variables. The major problem is that improper prior density functions are very likely to yield improper posteriors. The use of improper priors is widely accepted if the resulting posterior density functions are proper. Whether a posterior is proper or not can usually be easily examined if the model is simple. However, it is no longer easy when the model is more complicated. Kass and Wasserman [86] also summarised four other problems: incoherence and strong inconsistencies (see the examples in Stone [119]), the dominating effect of the prior, inadmissibility, and marginalisation paradoxes (see Dawid, Stone and Zidek [46]).

2.3 Hierarchical Modelling

In many applications, there may be more than one parameter in the sampling distributions. Frequently, these parameters are related to each other by some hierarchical structure according to the nature of the application. Suppose we have random quantities X, θ_1 and θ_2 whose joint density function is $p(X, \theta_1, \theta_2)$. The hierarchical structure of these random quantities is based on a prior relationship that X is independent of θ_2 given θ_1, which is denoted by

$$X \perp\!\!\!\perp \theta_2 | \theta_1$$

by Dawid [40]. The relationship of X, θ_1, and θ_2 can also be illustrated in a simple directed graph

$$X \longleftarrow \theta_1 \longleftarrow \theta_2,$$

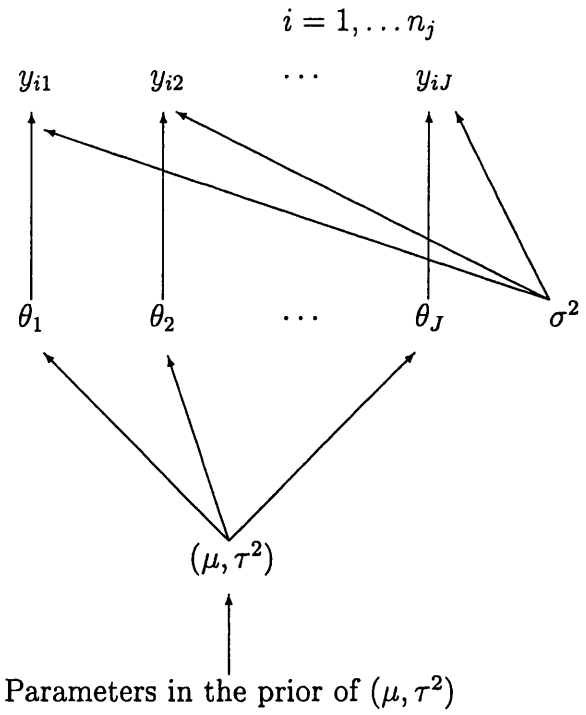


Figure 2.1: One-way normal random effects model when σ^2 is given.

which shows the hierarchy. The conditional distribution of X depends only on the parameter connected to it in the graph, which is θ_1 in this case, and hence, the joint distribution can be expressed as $p(X|\theta_1)p(\theta_1|\theta_2)p(\theta_2)$. A parameter which does not connect to X is called a hyperparameter, whose prior distribution is called a hyper prior distribution.

Consider the one-way normal random effects model as a simple example. Suppose there are J independent experiments. In the j^{th} experiment, there are n_j data points, with unknown mean θ_j and common known variance σ^2 for each observation. Denote the i^{th} observation in the j^{th} experiment as y_{ij}. Therefore,

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$$

independently for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. The conjugate prior distribution of θ_j is $N(\mu, \tau^2)$ for every j, and θ_j are independent of each other given hyperparameters μ and τ^2. The hyperparameters are given prior distributions. The hierarchical structure of the model is presented in the figure 2.1.

Consider a general case of hierarchical model with some parameters θ_1,

..., θ_J. If the joint distribution of these parameters is invariant to permutations of the indices $(1, \ldots, J)$, we say $\theta_1, \ldots, \theta_J$ are exchangeable. In practice, such an assumption is frequently made because there is not enough information to distinguish one parameter from the others. The most simple assumption is that the parameters $\theta_1, \ldots, \theta_J$ are mutually independent with identical prior density functions conditional on some hyperparameters (de Finetti [47]). The one-way normal random effects model is one example. The parameters θ_j's are exchangeable given μ and τ^2, and the observations $(y_{1j}, y_{2j}, \ldots y_{n_j j})$ are exchangeable given θ_j. The observations y_{ij} are referred as partially exchangeable because they are only exchangeable in the subset to which they belong. Exchangeable observations are often referred to as iid samples given a set of parameters.

In many applications, the graph for the random variables and parameters in a model is quite complicated. Models with a hierarchical configuration can usually be displayed systematically even though the number of parameters is large. The joint distribution of random variables and parameters can always be factorised as the product of some conditional prior density functions with some parameters. Statistical inference with the resulting factorised joint distribution is often more computationally efficient (see examples in the next paragraph). However, the hierarchical structure of a model should follow the structure of the application itself.

Hierarchical structure has been exploited in many statistical models, e.g. in the linear model of Lindley and Smith [91], for discrimination in Brown *et al.* [24], in the spatio-temporal model in Zhu and Carlin [132], for experimental design in Tiao and Box [123], longitudinal analysis in Kass and Steffey [85], categorical data in Albert and Chib [1], and many other examples. Due to the progress in computing ability, hierarchical modelling has been applied in more and more practical cases. For example, Johnson [83] proposed a hierarchical specification for image analysis; Cohen *et al.* [34] considered criminal cases relating to drugs and robbery with a hierarchical model; a DNA profile modelling application can be seen in Dawid and Pueschel [45]; Brown *et al.* [24] applied it to spectroscopic data,

Besag and Higdon [15] used it in agricultural experiments. There are also many applications in medical statistics, such as Congdon and Best [35]. Most of the applications have to employ MCMC techniques because of their high-dimensional parameter space and complicated posterior structure.

2.4 Model Assessment

A fundamental problem in modelling is that we can rarely claim we know the true model for the events we observe. Ideally, we would like to explore all possible models in the universe in order to find the best one. However, it is impossible to fit universal models. Instead, we choose the best model among a collection of models, or we gradually modify our original model until we accept one.

Research in model assessment has an extensive literature. Relevant topics include model comparison, outlier detection, sensitivity analysis and others. In this section we briefly review some of these topics. We divide the topics into three categories: model checking, model selection and sensitivity analysis. Model checking concerns whether our model describes or fails to describe the data. Examples are outlier detection, influential observation detection, checking of model assumptions and overall fitting. Model selection means to select the best model from a collection of models or to select the best subset of variables in a model. Sensitivity analysis focuses on the stability of the models we choose. Many of these methods for model assessment are simulation-based. In this section, we briefly introduce some topics in model checking and sensitivity analysis.

2.4.1 Model Checking

In classical analysis, residuals have been an important source of information for model diagnostics. One can simply plot the residuals to check whether there are outliers or whether the model assumption is right. A similar idea has been brought into Bayesian analysis. In classical statistics, the residual is defined as the difference between the observation and the fitted value of the observation using

the fitted model, and each residual is a fixed value because the fitted value is a fixed value. However, the residuals in Bayesian statistics are not fixed values but have a distribution, since the prediction of a future value is summarised by a posterior predictive distribution. Box [16] and Rubin [111] considered more general residual functions for examining model adequacy in a Bayesian context. The development of model checking is generally based on the posterior predictive distribution of samples. Simulation is frequently required because the posterior predictive distributions are usually non-standard.

One method of checking the appropriatness of a model is proposed by Gelman *et al.* [66]. This method is based on the method developed by Rubin [111]. A proper discrepancy variable T (a function of data) is defined for a model. Let y^{rep} (notation of Gelman *et al.*) represent the data generated by the posterior predictive distribution and y^{obs} represent the observed data. The posterior predictive p-value $p(T(y^{\text{rep}}) > T(y^{\text{obs}})|y^{\text{obs}})$ is calculated to evaluate the model. Many examples are shown in Gelman *et al.* [65]. Also see Gelman and Meng in Gilks *et al.*[73]. Gelman *et al.* [66] also emphasise the importance of using a graphical comparison of the histogram of y^{obs} and the histogram of y^{rep}. The graphical display of these histograms usually provides more information than a p-value can provide. One problem with Gelman *et al.* [66]'s method is that y^{obs} have been used to fit the model which produces y^{rep}. As a result, their method is less critical of the model than it might be (see Dey *et al.* [50]). It cannot prevent overfitting.

Some authors focus on the model diagnostic methods for hierarchical modelling. Albert and Chib [1] consider outliers, exchangeability and other properties for conditionally independent hierarchical models. Their approach is in fact a model comparison approach. Dey *et al.* [50] propose a stage-wise checking method to examine the failure in each stage of the structural assumption for the hierarchical models. Their method is also based on the discrepancy measurement as in Gelman *et al.*[66]. Hodges [80] considers the general hierarchical linear models and suggests tools for examining candidate added variables, transformations, collinearity, case influence, and residuals.

Outlier detection is also a topic considered. One example was given by Chaloner and Brant [31], who propose an outlier detection method for Bayesian linear regression when the variance of the random errors is known. The probability of an observation being an outlier is calculated. A graphical diagnostic tool is also proposed. A similar idea is to check whether there is any observation that is very influential to the model. For a review see Pettit [101]). Hodges [80] also considered using some graphical tools in his paper. Some authors suggest the use of cross-validation so that the replicated data generated from the posterior predictive distribution are compared with observations that have not been used to fit the model. Examples of model checking using cross-validation are Pettit and Young [102] and Gelfand *et al.* [62]. Dey *et al.* [50] and Gelman *et al.* [65] provide reviews of many methods for model checking.

2.4.2 Sensitivity Analysis

A good model does not only fit the data well. We also expect the model to be robust. In a robust Bayesian analysis, small changes in a prior model (the sampling distribution and the prior density functions for the parameters in the model) should not cause significant changes in the posterior model (posterior distributions for parameters and the posterior predictive distribution) . The main purpose of sensitivity analysis is to evaluate the stability of the models.

Bayesian sensitivity analysis examines the mapping from prior to posterior across a class of sampling models or prior distributions (see Draper [52]). To test priors only, one fixes the sampling distribution and varies the prior distribution. Usually, the prior distribution is varied by changing the values of the parameters of the prior distribution. In such a case, one selects several parameter settings in the same distribution family as priors and obtains the corresponding posterior distributions or predictive distributions, then compares the posterior means of parameters of models or compares the predictive performance using a suitable scoring rule, for example by cross-validation in Draper [53]. Berger [8] suggests the use of the ϵ-contamination class as a collection of prior distributions. The

ϵ-contamination class is defined as

$$\Gamma = \{\pi : \pi(\theta) = (1-\epsilon)\pi_o(\theta) + \epsilon q(\theta), q \in \mathfrak{L}\},$$

where $\pi_o(\theta)$ is a chosen prior of θ, $0 < \epsilon < 1$ is the weight of $q(\theta)$ and $\mathfrak{L}$ is a class of possible "contaminations." A different approach is to consider the Bayes risk of candidate models with different prior settings (see Berger [8]). Kass and Raftery [84] suggested using sensitivity to examine whether the Bayes factor is sensitive to the prior or not. A theoretical introduction to Bayesian posterior and risk sensitivity analysis is available in Berger [8]. Alternative approaches are suggested in Weiss [125].

When the posterior model can be obtained analytically, sensitivity analysis can be easily achieved since the posterior means of parameters or the scores of predictive performance are simply functions of given parameters and observed quantities. When the analytical posterior model is not available, sensitivity analysis is time-consuming. The same process of numerical approximation or stochastic simulation of parameters has to be done for each prior setting. There seems to be no short cut for doing sensitivity analysis when a model is complicated. When there are many parameters, the amount of computing for sensitivity analysis considering variation for every parameter is tremendous.

2.5 Bayesian Multivariate Analysis for Normal Variables

2.5.1 Matrix-variate Distributions

In this subsection we briefly introduce matrix-variate distributions, including matrix normal, Wishart, inverse-Wishart, matrix-T, and matrix-F distributions, which are involved in the models in this thesis. We follow the notation for these distributions developed by Dawid [41]. The notation is not necessarily the same as the notation for these distributions which has been considered by the other authors. The notation system is designed so that symbolic Bayesian manipulations

for matrix-variate conjugate analysis can easily be carried out. The density functions of these distributions are provided in appendix A.

Matrix Normal

The distribution of an n by p random matrix X with independent standard normal elements is denoted by $X \sim \mathcal{N}(I_n, I_p)$. For constant matrices A with n columns, B with p rows, and M with the same dimension as AXB, the distribution of $M + AXB$ is denoted by $M + \mathcal{N}(\Lambda, \Sigma)$, where $AA^t = \Lambda$ and $B^tB = \Sigma$.

Wishart Distribution

Let $X_{n\times p} \sim \mathcal{N}(I_n, \Sigma_{p\times p})$, and $Z = X^tX$. The distribution of Z is a Wishart distribution with shape parameter n and scale matrix $\Sigma_{p\times p}$, denoted as $\mathcal{W}(n; \Sigma_{p\times p})$. For a general Wishart distribution, the shape parameter can be any positive real number, and the scale matrix needs to be non-negative definite. Suppose $\psi \sim \mathcal{W}(\nu; \Lambda)$, the expectation of ψ is $\nu\Lambda$.

Inverse-Wishart Distribution

Let a p by p matrix Φ be inverse-Wishart distributed with shape parameter $\delta > 0$ and scale matrix $\Sigma \geq 0$, we denote it as $\Phi \sim \mathcal{IW}(\delta; \Sigma)$. The expectation of Φ is $\Sigma/(\delta - 2)$ if $\delta > 2$ and $\Sigma > 0$. The distribution of Φ^{-1} is a $\mathcal{W}(\nu; \Sigma^{-1})$, where $\nu = \delta + p - 1$.

Matrix-t Distribution

Suppose $T_{n\times p} \sim \mathcal{N}(\Lambda, \Phi)$, given Φ and $\Phi_{p\times p} \sim \mathcal{IW}(\delta; \Sigma)$. Then the marginal distribution for $T_{n\times p}$ is a matrix-t distribution, denoted by $T_{n\times p} \sim \mathcal{T}(\delta; \Lambda, \Sigma)$. The distribution $\mathcal{T}(\delta; I_p, I_q)$ is called a standard matrix-t distribution with parameter δ.

Matrix-F distribution

The $p \times p$ random matrix U having a matrix-variate F distribution with parameters ν, δ, and K is denoted as $U \sim \mathcal{F}(\nu, \delta; K)$, with mean $\nu K/(\delta - 2)$. Suppose $U|\Phi \sim \mathcal{W}(\nu; \Phi)$ with $\Phi \sim \mathcal{IW}(\delta; K)$, then marginally $U \sim \mathcal{F}(\nu, \delta; K)$. If $U \sim \mathcal{F}(\nu, \delta; I_p)$, then $U^{-1} \sim \mathcal{F}(\delta + p - 1, \nu - p + 1; I_p)$. If $T \sim \mathcal{T}(\delta; I_p, I_q)$ then $T^t T \sim \mathcal{F}(p, \delta; I_q)$.

2.5.2 Bayesian Models for a Covariance Matrix

In multivariate analysis, we often assume variables are normally distributed. For a multivariate normal distribution, there are two parameters: the mean and the covariance matrix. A covariance matrix is also called a variance matrix, a dispersion matrix, or a variance-covariance matrix. Often, the mean and the covariance matrix are unknown, and we have to assign prior distributions for them. Suppose we assume the mean is again from a multivariate normal distribution, then there is another covariance matrix to be specified. Therefore, the assumption for covariance is inevitably an important issue.

Sometimes, we may have reliable information about the covariance matrix, but frequently we do not have much information about it, and diffuse distributions such as inverse-Wishart with small shape parameter or a flat prior to represent our prior ignorance are frequently in use. Suppose p is the number of variables in our model. For small p with many data, the assumption for the covariance matrix is usually not important because data speak for themselves and the estimation is usually very close to the maximum likelihood solution. However, the prior distribution is increasingly informative when the number of variables increases. Therefore, more consideration has to be given to the prior assumptions. In this section, we introduce some of the frequently used prior assumptions for covariance matrices.

Jeffreys' prior where $p(\Sigma) \propto |\Sigma|^{(p+1)/2}$ and a flat prior $p(\Sigma) \propto 1$ are both commonly used as non-informative prior for the covariance matrix Σ. However, care must be taken when applying them because both of them may lead to improper posterior distributions, and the flat prior can be very informative in the use of

a small data set. There are also other alternatives. For example, Daniels [37] derived a non-informative prior for the covariance matrix as a hyper parameter in a hierarchical model.

A conjugate prior is always an attractive assumption because of the convenience in manipulation. The natural conjugate prior for covariance matrices of normal variables is the inverse-Wishart distribution. Chen [32] for example assumed the natural conjugate prior, which is inverse-Wishart for the covariance matrix and assumed the parameters in the inverse-Wishart known. However, it is known that the inverse-Wishart prior lacks flexibility. Once its mean has been decided, we can only use the scalar shape parameter to determine the distribution of the $p \times (p+1)/2$ parameters in the covariance matrix.

Suppose a p by p matrix $\Sigma \sim \mathcal{IW}(\delta; \Phi)$,where $\Phi > 0$. Let σ_{ij} be the ij th element of Σ and ϕ_{ij} be the ij th element of Φ. According to Theorem 5.2.2 in Press [103],

$$\text{var}(\sigma_{ii}) = \frac{2\phi_{ii}^2}{(\delta-2)^2(\delta-4)}$$

for $\delta > 4$,

$$\text{var}(\sigma_{ij}) = \frac{\phi_{ii}\phi_{jj} + \frac{\delta}{\delta-2}\phi_{ij}^2}{(\delta-1)(\delta-2)(\delta-4)},$$

for $\delta > 4$ and $i \neq j$, and

$$\text{cov}(\sigma_{ij}, \sigma_{kl}) = \frac{\frac{2}{\delta-2}\phi_{ij}\phi_{kl} + \phi_{ik}\phi_{jl} + \phi_{il}\phi_{kj}}{(\delta-1)(\delta-2)(\delta-4)},$$

for $\delta > 4$ (for all i, j, k, l). When the shape parameter is small, the variance is large. When $\delta < 4$, the variance does not even exist. Therefore, an inverse-Wishart distribution with a small shape parameter is usually considered as a diffuse prior, while for one with large δ, the prior is more informative.

Consider the conjugate model for a 1 by p random vector X

$$\begin{aligned} X &\sim \mathcal{N}(1, \Sigma), \\ \Sigma &\sim \mathcal{IW}(\delta; \Phi), \end{aligned}$$

Suppose we observe n independent samples for X, represented as x, which is an n by p matrix (each row represents one observation). The posterior distribution

for Σ is $\mathcal{IW}(\delta + n; x^t x + \Phi)$, with expectation $(x^t x + \Phi)/(\delta + n - 2)$. When the number of samples n is very large, $E(\Sigma|x)$ is almost $x^t x/(\delta + n - 2)$ given the same Φ. When n is small, Φ is more influential for the posterior Σ.

In a non-hierarchical model, δ and Φ are considered as known constants although one may not be so certain about how well the inverse-Wishart distribution represents our prior belief. Further assuming a hyper prior distribution for the hyperparameters extends the flexibility of the prior configuration. Moreover, the marginal distribution of Σ can be more diffusive in a hierarchical model. Therefore, the hierarchical model can be less sensitive than a non-hierarchical one.

An inverse-Wishart distribution with a structured scale matrix has been considered as the prior for Σ by many authors. Such an assumption reduces the number of parameters from $p(p+1)/2$ to a small number so that the computational aspect of model inference is simpler. Ideally, the structure should be consistent with our belief in the data. However, the real structure of a covariance matrix is usually too complicated or simply unknown, especially in a high dimensional case. The most common and simple form is the diagonal matrix with equal diagonal elements. Dickey, Lindley and Press [51] consider an intraclass covariance structure for the scale matrix. Brown [23] considered the structural coherence (see chapter 5) of data and suggested the use of ARMA-type correlation structure. In hierarchical modelling of the covariance matrix, hyper prior distributions are assigned to the parameters in the structured scale matrix.

There are also methods which consider the spectral decomposition of the covariance matrix. Suppose the covariance matrix is Σ. Yang and Berger [129] and Daniels and Kass [38] consider an orthogonal decomposition of Σ to $O^T D O$ where D is a diagonal matrix and O is some orthogonal matrix which is further decomposed into $p \times (p-2)/2$ matrices. Barnard, McCulloch and Meng [5] decompose the covariance Σ as $\Sigma = \text{diag}(S)\, R\, \text{diag}(S)$, where R is the correlation matrix of the normal variables, S is the vector of standard deviations, and $\text{diag}(S)$ is a diagonal matrix with diagonal elements S. Prior distributions are then assigned to S and R. Structures can be considered for the correlation matrix. Leonard and

Hsu [89] also consider the orthogonal decomposition. They do not assign separate priors to the individual components. They consider $A = \log(\Sigma)$ and arrange the elements of the upper triangle of A as a vector α, then they use an approximation for the likelihood function of α from the likelihood function of A and assume α has a multivariate normal prior distribution. They also consider using a hyper prior to express belief about the parameters for the distribution of α, with a structured covariance matrix for α.

2.5.3 Bayesian Regression

In regression analysis, we create a model to predict response variables $Y_1, Y_2, \ldots, Y_q$ using explanatory variables $X_1, X_2, \ldots, X_p$. Let $Y = (Y_1, Y_2, \ldots, Y_q)$ and $X = (X_1, X_2, \ldots, X_p)$, which are 1 by q and 1 by p, respectively. In a regression model, Y is predicted by $X\beta$, where β is a p by q regression coefficient matrix. According to the way we treat the training samples of X, there are generally two types of Bayesian regression model. The most widely applied one considers the training data x for X to be fixed. These x can be designed or observed. When x are designed, the training data for both Y and X cannot represent the population. The model is

$$Y = X\beta + E,$$

where E $(1 \times q)$ is a vector of random errors and the only source of uncertainty. The sampling distribution of this model is the distribution of Y given X. We call this a controlled regression model. The other type of regression is called the random regression model, which considers the random property of X in the model. In this case, training data for (Y, X) are sampled randomly from the population. The sampling distribution of the model is the joint distribution of Y and X. The regression coefficient matrix can be derived from the joint distribution of them.

The central interest in regression analysis is the estimation of β and the prediction of future responses. In this thesis, we apply regression analysis in a Bayesian framework. Bayesian regression has been studied since the mid 20^{th} century. Tiao and Zellner [124] and Geisser [59] independently worked out the

posterior results for a multivariate regression model with the same vague prior assumption for the parameters under a non-hierarchical structure. Early Bayesian books by Box and Tiao [18] and Zellner [131] give a comprehensive introduction to various regression models. The book by Broemeling [20] specialises in the Bayesian linear model.

The development of Bayesian regression is associated with progress in general Bayesian theory. Lindley and Smith [91] first applied de Finetti's idea of exchangeability to the regression coefficients of multiple regression and expanded the model to a three-stage model, with proper priors for parameters, e.g. regression coefficients and the covariance matrix of regression coefficients. In Chen's [32] paper about estimating the covariance matrix he considered a random regression application. Dickey, Lindley and Press [51] also applied their intraclass covariance structure to the joint distribution of the explanatory and response variables in a random regression model.

An interesting problem in regression occurs when the number of variables exceeds the number of samples. This is the main problem we consider in this thesis. More recently, some research has focused on this topic, mainly taking advantage of the fact that the Bayesian approach does not have constraints on the number of variables. Dawid [43] first developed the theory for conjugate Bayesian random regression with an infinite number of regressors. Fang and Dawid [54] continued the study for non-conjugate infinite random regression. Mäkeläinen and Brown [93] considered coherent priors for a partially exchangeable model. They developed a class of inverse-Wishart priors for a finite or countably infinite dimensional normal model with unknown covariance matrix. Later, Brown and Mäkeläinen [27] used a structural coherent prior for the covariance matrix in the model. They assumed the correlation of predictor variables had the structure of the autocorrelation function for an ARMA process. Brown [23] defined a generalised inverse-Wishart distribution for the covariance matrix in the multivariate regression model in an attempt to overcome the natural limitations of the standard inverse-Wishart distribution but still retain the analytic tractability of modelling. Brown *et. al.* [28] considered

variable selection procedures for the natural conjugate random regression model with many variables using simulating annealing, while Brown *et al.* [25] considered Bayesian variable selection based on the model in Fang and Dawid [54]. West [126] proposed a methodology of Bayesian regression analysis which is different from Brown's approach. West's approach is based on latent factor regression models, which are essentially controlled regression models. In his approach, responses are regressed on new explanatory variables that are linear combinations of original regressors. These new regressors are produced through singular-value decompositions. The same approach has been applied in analysing a binary regression model with many variables in West *et al.* [127].

2.5.4 Bayesian Discrimination

Discriminant analysis handles the problem of allocating an observation to one of several groups or populations on the basis of a multivariate observation. The number of populations can be known or unknown. The parameters in the density functions of the populations usually need to be estimated. Frequently it is known which groups the training data come from. However, due to the cost of collecting membership information, we may not be able to distinguish the population identity of some training data. Anderson [2] reviewed classical normal discriminant analysis, indicating that Bayes' procedure is admissible. McLachlan [95] provides a comprehensive introduction to discriminant analysis.

Suppose an item must come from one of g groups, labelled as group 1 to group g. The Bayes' procedure minimises the risk of misclassifying an object. It is based on a loss function $U(\pi)$ and the group membership distribution $p(\pi|z)$, where z is the multivariate observation on the object and π represents the identity of the group we allocate the object to, i.e. $\pi \in \{1, 2, \ldots, g\}$. According to Bayes' formula, the predictive probability of the item being from the i th group is

$$p(\pi = i|z) = \frac{q_i p(z|\, \pi = i)}{\sum_{i=1}^{g} q_i p(z|\, \pi = i)}, \tag{2.3}$$

where q_i is our prior probability that this item should come from the i th group. In order to find the predictive probability, we need to know individual $p(z|\, \pi = i)$.

An alternative approach is logistic discrimination, which models $p(\pi|z)$ directly using a logistic regression model using z as regressors [95].

The Bayesian approach to discriminant analysis for multivariate normal variables has been discussed by Geisser [58] [60] and [61]. He obtains the Bayesian estimation for $\{p(z|\pi = i)|i = 1, \ldots g\}$, then calculates the predictive probability using equation (2.3). Rigby [107] further investigated the posterior density and the credibility interval of the predictive probability of a new observation in one of the two possible populations. Rigby [108] compared classical and Bayesian results and concluded that the Bayesian method can produce a less extreme result when there are many variables. However, the number of variables in his example can only be regarded as small in our context. Bayesian logistic discrimination has also been developed, see for example Fearn *et al.* [56], which is essentially a logistic regression analysis. Bayesian logistic regression has been applied by many authors.

Research in exploring discriminant analysis with training data without knowing their group identity is also considered by some authors. It involves inference for mixture models. Special techniques for handling such problems are being developed because of their natural complexity. Lavine and West [88] applied iterative resampling techniques by Gelfand and Smith [63] for a known number of populations. Sometimes even the number of populations is unknown. Richardson and Green [106] analyse such problems with reverse jump MCMC.

When the number of variables exceeds the number of samples, the sample covariance matrices are not invertible and most classical approaches to discriminant analysis become impossible. As in regression, the Bayesian approach still works, if we have proper prior distributions. Similar to Bayesian regression analysis, Dawid and Fang [44] proved that under conjugate prior assumptions for the parameters of the density functions of two normal populations, the model produces perfect discrimination, something we should not wish to happen in many applications. Brown *et al.* [24] considered a practical case in NIR calibration where there are many variables using the predictive probability approach. A brief review for Bayesian discrimination with many variables can be found in Brown [23].

Chapter 3

Near Infrared Spectroscopical Analysis

3.1 Introduction

In recent years near infrared (NIR) spectroscopy has become a very important tool in analytical chemistry. Traditional laboratory-based methods for analytical chemistry are often time consuming, hazardous and the cost of the space, equipment and personnel for a laboratory is expensive. However, a modern NIR instrument is rapid, low cost and safe, and is able to produce many measurements with similar accuracy to that obtained using laboratory methods. The process of taking NIR spectral measurements and analysing data is usually completed in one small box, and the instrument is easy to operate so that less training is required for staff. Sample preparation is easier for NIR measurements, and sometimes it can even be omitted.

The NIR light band consists of light from wavelength 700nm to 2700nm (nm = nanometres, which is 10^{-9} meter), which is a sub-band of infrared radiation (700-10^6 nm) with longer wavelength than any visible light. The NIR spectra are the measurements of samples' absorption of radiation at wavelengths within the NIR light band. NIR absorbance spectra principally involve the interaction between NIR radiation and C-H, O-H, and N-H chemical bonds. Each kind of

bond only absorbs radiation at particular wavelengths (several for each bond, because it has several modes of vibration) and each absorption corresponds to a peak in the spectrum. For a complex sample such as foodstuff, there are very many absorption peaks so that many peaks overlap. Moreover, the instrumental NIR spectra will have been smoothed both by the hardware, since what is usually measured is the absorption averaged over a narrow band of wavelengths, and often by software as well. The result is typically a smooth looking spectrum that is actually made of up hundreds or even thousands of overlapping peaks. By contrast, the absorption peaks of a chemical bond in the mid infrared band (2700-25000nm) do not overlap so seriously and are generally much more distinguishable. Mid infrared spectroscopy can be used to fingerprint simple chemicals by identifying their absorption peaks, which cannot be achieved by using NIR spectroscopy. However, NIR spectroscopy has been widely used for quantitative analysis and certain qualitative analysis (i.e. discrimination) on complex materials because NIR spectroscopy is cheaper and easier to implement.

The main usage of NIR spectra is to predict the concentration of a constituent in samples or to discriminate between samples. Different instruments have been developed according to the requirements of users. They can be designed for specialised on-site process control, or for flexible laboratory use. Some instruments focus on the spectral measurements at certain important wavelengths, while others are designed to generate spectra at wavelengths spread over the NIR band. Samples can be in liquid or solid state for analysing. Solid samples may need to be ground although some instruments will handle, for example, samples of whole grain.

The recent advances in NIR spectroscopic analysis are not only due to the improvement in the mechanical and optical aspects of the instruments, but also to improvement in the techniques for calibration. Very often measurements at more wavelengths than the number of samples are taken. Entire spectra are often included in a calibration model in order to gather more complete information. It is known that classical regression methods cannot handle problems with a larger

number of variables than the number of samples. Methods for compressing data and dimension reduction (e.g. PCR, PLS) have been developed. These methods are reviewed in section 5.

3.2 Theory of NIR Absorption

Near infrared spectra are the result of light absorption by molecules, especially organic chemicals, mainly consisting of carbon (C), hydrogen (H), and nitrogen (N). The absorption of light by molecules corresponds to a change in the status of the atomic rotation and the vibration between two atoms at the two ends of a chemical bond. The absorption of radiation at NIR bands is the consequence of vibration only, and it is the result of overtones or the combinations of overtones rather than the fundamental changes of vibrating status that appear in the mid-IR region.

The vibration of two atoms at both ends of a chemical bond is an oscillation system. According to classical physics, the total mechanical energy of an oscillation should be a continuous function of the frequency and the maximum amplitude of the oscillation. However, the oscillation of atoms in fact obeys quantum theory and has discrete energy levels, labelled as ground state, the second state, the third state, etc. The transition of energy state can only occur by the absorption or emission of quanta, which are countable energy packages. The frequency of radiation absorbed or emitted is decided by the energy difference of the two states involved in the activity. The energy level system is different for different combinations of atoms in a chemical bond. The structure of the molecules also affects the required energy for transition. Therefore, an absorption band of a particular chemical bond is slightly different in different molecules and indeed may differ from the same bond at different locations in the same molecules. This is the reason why we see so many peaks.

When a transition happens between ground state and the first state, the transition is described as fundamental. When a transition happens between the ground state and a state higher than the second state, it is called an overtone. Ac-

cording to the selection rules for a harmonic oscillating activity in a chemical bond, transition can only occur for one step. Overtones are due to non-harmonic oscillation between two atoms, where the selection rules allow changes between more energy levels. In polyatomic molecules, many chemical bonds interact at the same time and the transition of the energy state is a consequence of the combination of the changes at individual bonds.

3.3 Linear Relationship between NIR Spectrum and Concentration of Constituents

When monochromatic radiation interacts with a sample, it may be absorbed, transmitted, or reflected. According to the law of conservation of energy, the incident radiant power (P_O) is equal to the sum of the radiant power absorbed (P_A), the radiant power transmitted (P_T) and the radiant power reflected (P_R), i.e.

$$P_O = P_A + P_T + P_R.$$

When the experiment is arranged properly, one of P_T and P_R can be zero, and P_A can be deduced by measuring the non-zero power of transmittance or reflection.

The energy absorption of radiation within NIR bands is normally described by the Beer-Lambert law, which states that the fraction dP/P of radiant energy P absorbed by an infinitesimal thickness of sample is proportional to the number of molecules dn which actually absorb the radiant energy in that thickness

$$-dP/P \propto dn,$$

which implies

$$\log(P_O/P_T) = abC$$

where the constant a is called the absorptivity of the molecule, b is thickness through which the radiation passes, and C is the concentration of molecules in the sample. In transmission spectroscopy, the fraction of radiation (P_T/P_O) transmitted by the sample is measured and called transmittance (T). The transmittance is

converted to absorbance (A), which is defined by $A = \log(1/T)$. Consequently, the Beer-Lambert law suggests a linear relation for the absorbance and concentration

$$A = \varepsilon C \tag{3.1}$$

or equivalently

$$C = \beta A \tag{3.2}$$

where $\beta = \varepsilon^{-1} = b^{-1}a^{-1}$. However, the relationship between A and C is rarely found to be perfectly linear in practice. Several reasons (see Osborne *et al.* [100]) cause the deviation from linearity. In addition, the graph of A against C does not always pass through the origin. Background absorbance is one of the reasons for this. This problem can be removed by applying methods of background correction. However, despite the nonlinearity, it is often possible to describe the concentration-absorption relationship as locally linear.

An absorbing peak is rarely caused by a single constituent of a sample but by several of them because the type of chemical bond that causes the absorption may exist in several of the constituents. Define $\varepsilon_i = a_i b$, where a_i is the absorptivity of the i th constituent and b is the thickness through which the radiation passes, which is equal for every constituent. If the law of additivity holds, the absorbance is

$$A = \sum_{i}^{q} (\varepsilon_i C_i),$$

where C_i is the concentration of the ith constituent and q is the number of constituents. Suppose we observe absorbance at p wavelengths. Define $\varepsilon_{ij} = a_{ij} b$, where a_{ij} is the absorptivity of the i th constituent at the j th wavelength. We then have p simultaneous equations

$$A_j = \sum_{i}^{p} (\varepsilon_{i,j} C_i), \qquad \text{for } j = 1, \ldots, p, \tag{3.3}$$

which is the foundation of the use of linear calibration for NIR absorbance spectra.

NIR diffuse reflectance is also widely used when samples are opaque and non-absorbing. Although there is no definitive theory for diffuse reflectance (Shenk, Workman, and Westerhaus [114]), many rules have been developed to describe the

relation between diffuse reflectance spectra and concentration. Linear dependence appears in practice to yield the most successful results (Olinger and Griffiths [99]), and the linear rule is

$$\log(1/R) \propto \frac{aC}{s},$$

where a is the absorptivity, R is the intensity of the reflected radiation of samples, and s is called the scatter constant. This is affected by a number of sample properties such as particle size, refractive index, moisture content, etc. If the law of additivity holds, then equations analogous to (3.3) can be written for diffuse reflectance.

The NIR spectra are greatly affected by some physical properties of samples, such as particle size, packing density, moisture status, and temperature (Osborne *et al.* [100]). The spectral measurements at different wavelengths can be very highly correlated due to these properties. These factors produce error in estimating the features that are independent of them. Techniques for background correction are therefore required. Practical experience has shown that using derivative spectra provides a better model in some cases. It yields new variables with lower correlations. The most frequently used derivative spectrum is the second derivative. Higher order derivative spectra are rarely used in practice. The price of taking the derivative of spectra is that it reduces the signal to noise ratio. Moreover, this approach will not be appropriate if the features in which we are interested are related to those physical properties whose effects are removed by taking the derivative.

3.4 Applications of NIR Spectroscopy

NIR spectroscopy has been successfully applied to predict protein, moisture, fat and carbohydrate content, which are the main ingredients of food and beverages [100]. Since the technique is rapid, requires less sample preparation, and can be non-destructive, NIR spectroscopy is especially useful for on-line quality control. Applications to monitoring other quality attributes such as sensory

tenderness, texture, and flavour have also been reported (e.g. Byrne *et al.* [30], Sørensen *et. al.* [115]). Fungi or parasites can also be detected by NIR spectroscopy (Baker *et al.* [4], Kiskó *et al.* [87]). In addition to the food and beverage industry, applications of NIR spectroscopy extend to areas which are related to organic material, such as tobacco, textiles, petrochemical and pharmaceutical industries, agricultural research and the life sciences (See Handbook of Near-Infrared Analysis [29]).

3.5 Examples

Two examples involving spectra of wheat samples that have been used in this thesis for Bayesian regression and discriminant analysis are introduced in this section. The principle of the instrument that produced the spectra in the examples is explained in the first subsection. The second and the third subsections describe the examples for regression and discriminant analysis respectively.

3.5.1 Generating Spectra

The spectra were measured on samples of unground wheat using a Tecator Infratec Grain Analyzer which measures transmission of NIR radiation through wheat samples. The size of the instrument is about $60 \times 45 \times 45$ cm (see figure 3.1). Wheat samples are collected from a hopper on the top and then go through the transmitting cell, where the absorption and transmission occur, and are taken away from a drawer after they are detected. The transmitting cell is thin (typically 20 mm) so that the NIR path through the sample is quite short (see figure 3.2). The instrument is connected to a computer where the data are analysed. The instrument is specialised for measuring the NIR absorbance spectra of whole grains and offers predictions for the percentage of protein, oil, starch and moisture of the grain samples.

The amount of wheat for each sample is a few hundred grams. The light from a source passes through a wavelength selector, and only selected radiation

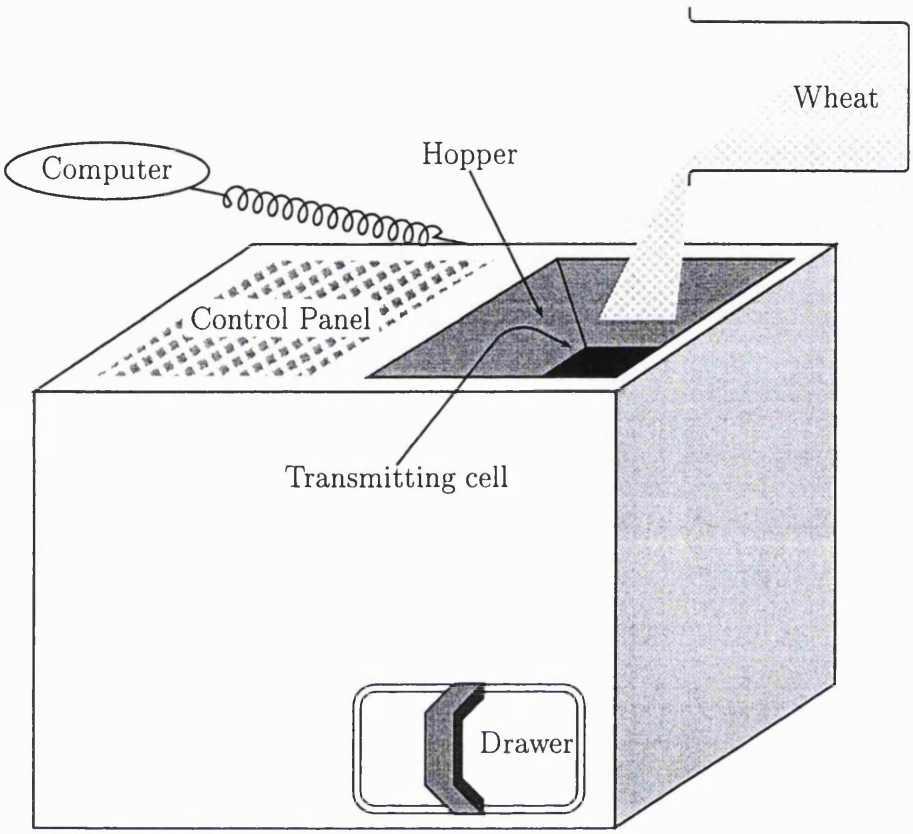


Figure 3.1: Simple diagram of the Tecator Infratec Grain Analyzer

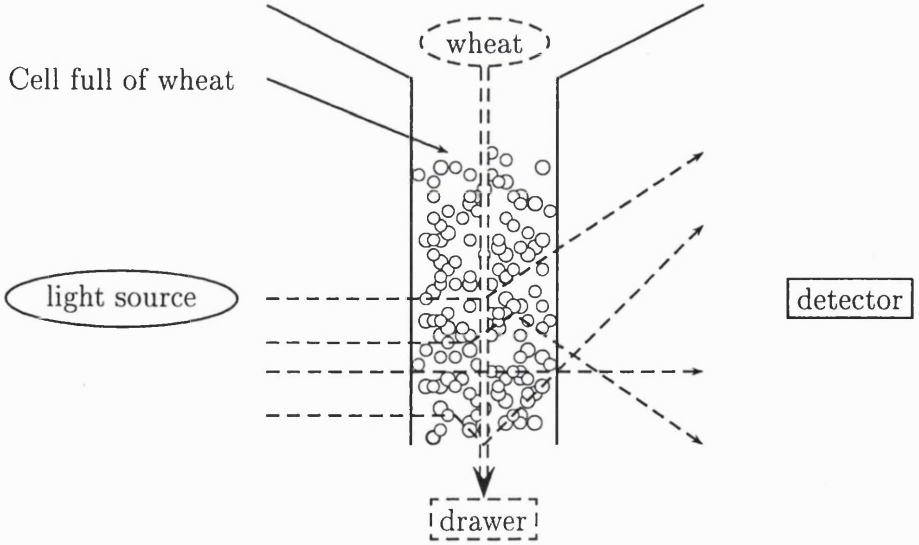
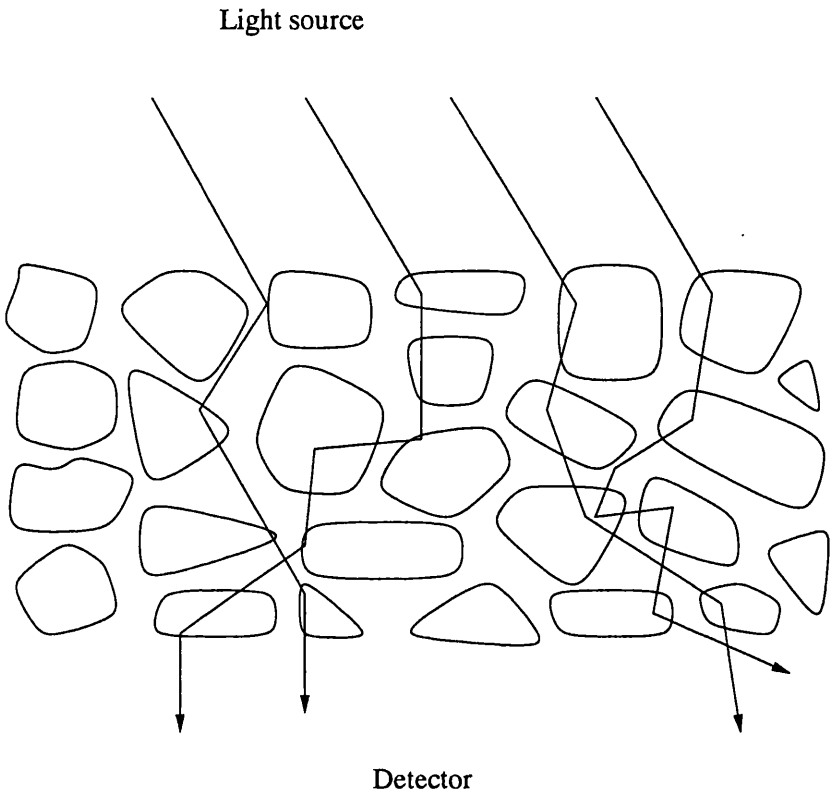


Figure 3.2: Transmitting cell

Figure 3.3: Mixed effect of reflection and absorption.



at certain wavelength goes through the sample. There is a gate between the transmitting cell and the drawer. One sample is put into the transmitting cell. The sample is gradually released from the transmitting cell to the drawer through a gate between the cell and the drawer. When the gate is closed, the wheat sample stops in the cell, and a certain amount of wheat is at the area in the cell where the light passes through and interaction between light and wheat happens. The detector measures the transmitted radiation energy when the sub-sample is stationary in the cell so that the measurement is more accurate. Then, the gate opens to release a certain amount of wheat to the drawer and then closes again. Another sub-sample of wheat is then measured. This procedure is repeated. As a result, there will be several spectra for the entire wheat sample. The final spectral output of the sample is the average of the spectra of the several sub-samples.

The transmission spectra produced by the Tecator Infratec Grain Analyzer is in fact a mixed effect of reflection and absorption (see figure 3.3). Empirically,

Variety	1	2	3	4	5	6	7	8	9	Total
Training	42	11	29	23	54	10	13	30	22	234
Validation	10	3	7	6	14	3	3	7	5	58

Table 3.1: Wheat data: numbers of samples of each variety

the linear model has provided good calibration for this type of spectra. In our examples, the transmission spectra of wheat are measured at 100 wavelengths from 850 nm to 1048 nm, with 2 nm increments at each step.

3.5.2 Two Examples

In the first example, there are 50 samples of wheat. NIR spectra and the protein percentage of each sample have been measured. The protein percentage is measured by the standard laboratory method, Kjeldahl nitrogen analysis on ground wheat. The original spectra of the 50 samples are shown in figure 3.4. The spectra are very smooth and the spectra of the 50 samples are shifted almost parallel to each other mainly due to the packing density and the particle size effect. The correlation between any pair of the 100 measurements in a spectrum is consequently very large (close to one). Second derivative spectra of the original ones are shown in figure 3.5. The correlations between the measurements at different wavelengths due to packing density and particle size effect are greatly reduced by taking the derivative. It also allows some of the other spectral variation to be seen. As we can see figure 3.5, one of the 50 (the 7th in our entire data set) 2nd derivative spectra is very different from the other spectra between 850 and 900 nm. The percentage protein of each sample is shown in figure 3.6. The corresponding point of the 7th observation is the point with largest percentage protein, but this is not the cause of the deviations.

In the second example, the data set consists of NIR transmission spectra on 292 samples of wheat from nine varieties. The identity of each sample is known. The spectra have the same shape properties as the spectra in the first example. The 292 samples were split randomly within groups into training and validation

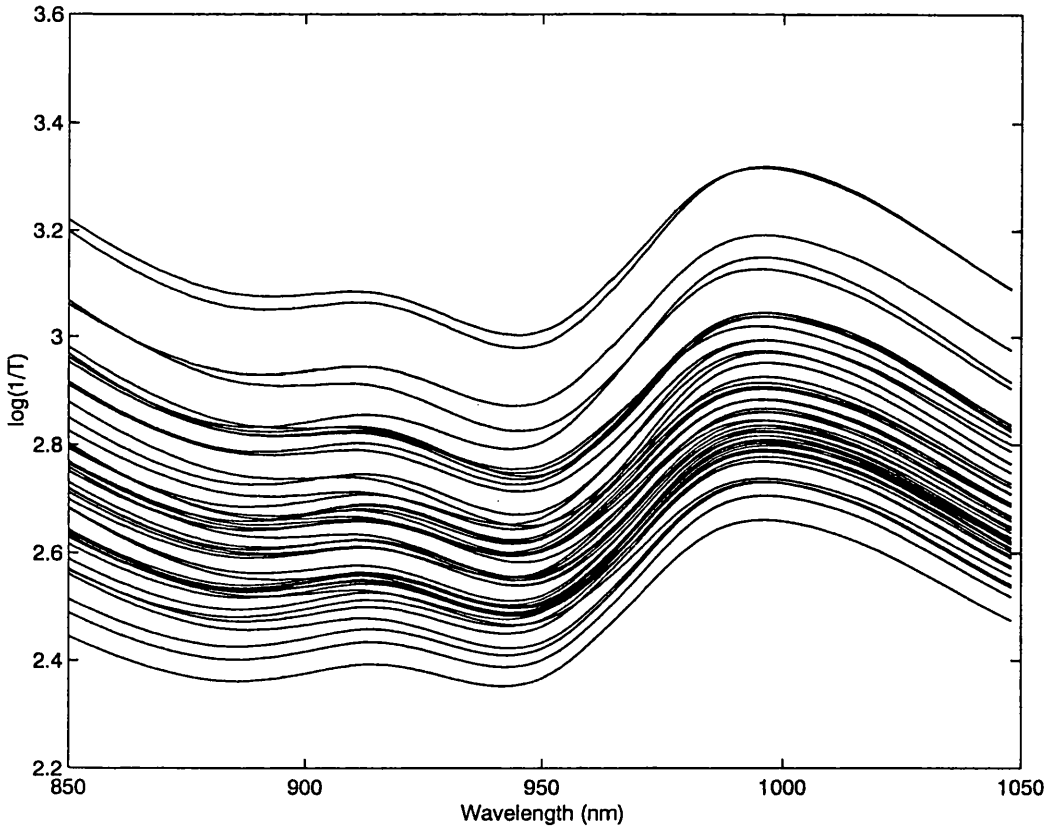


Figure 3.4: Spectra of 50 wheat samples

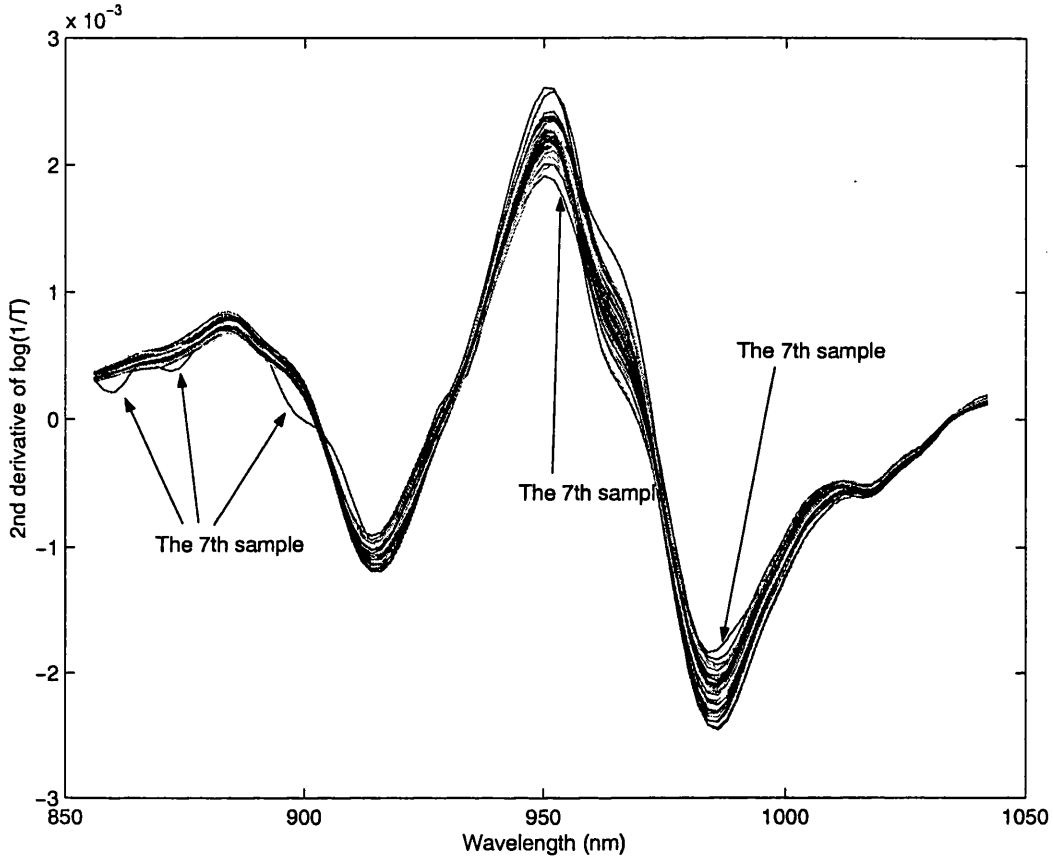


Figure 3.5: Second derivative spectra of 50 wheat samples

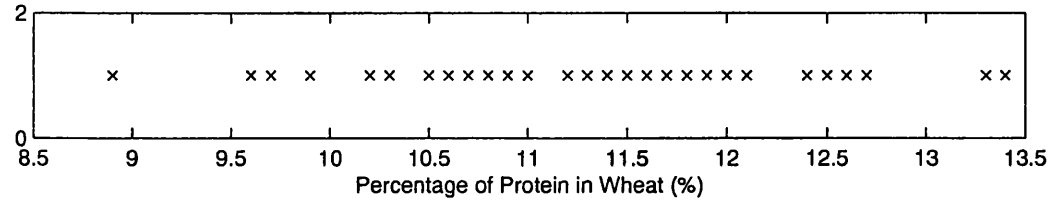


Figure 3.6: Dot plot of the percentage protein in the 50 wheat samples

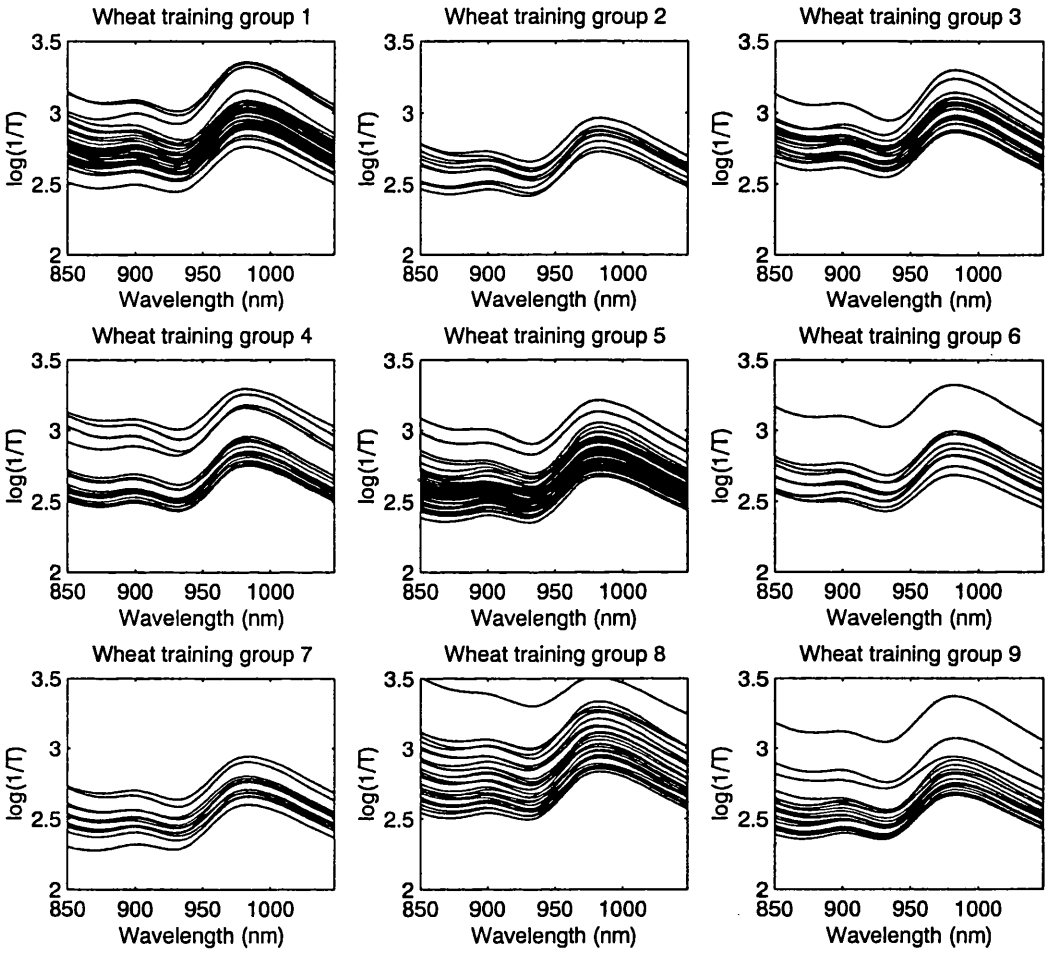


Figure 3.7: Transmission spectra of nine wheat varieties in the training set.

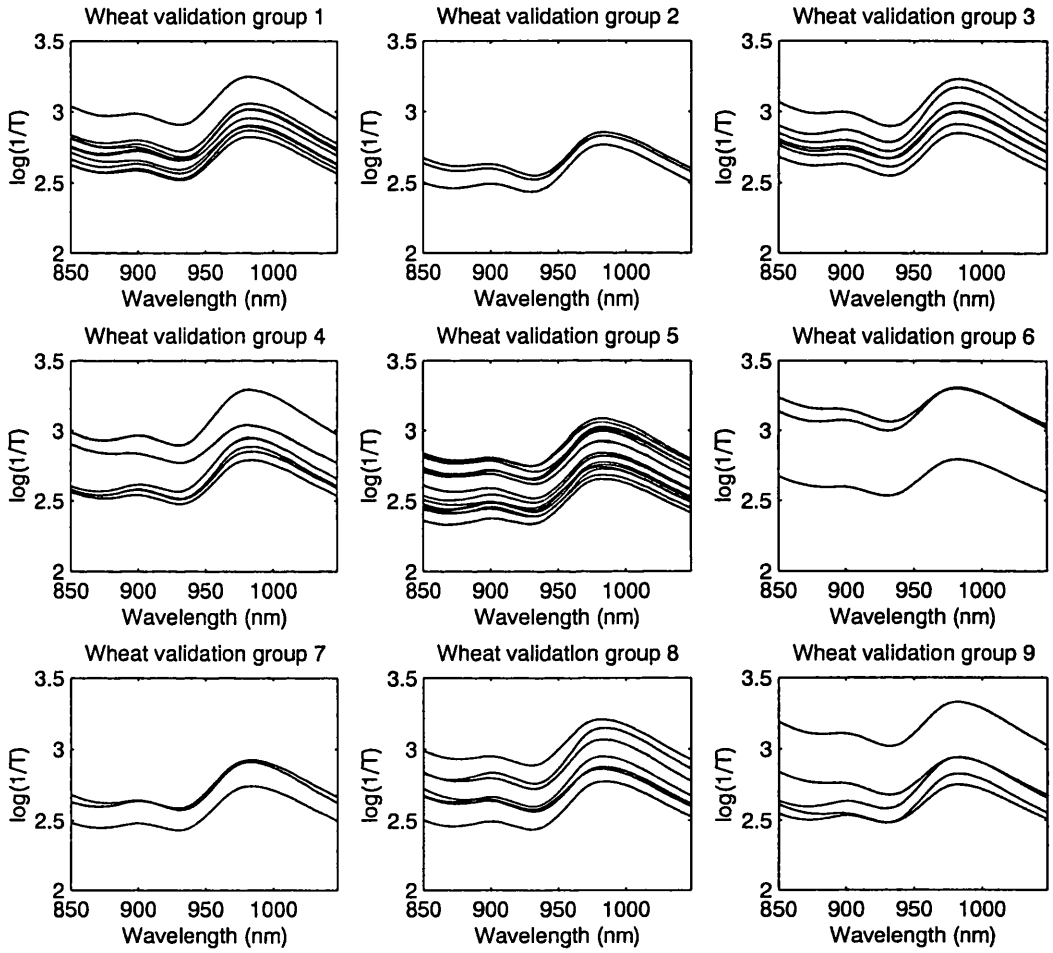


Figure 3.8: Transmission spectra of nine wheat varieties in the validation set.

sets, which contain 80% and 20% of the samples respectively. Table 3.1 shows the number of samples of each variety and in the training and validation sets. Figure 3.7 gives the spectra of the 234 samples in the training set, while figure 3.8 displays the spectra of the samples in the validation set.

In this thesis, a Bayesian regression model is applied to the data in the first example for predicting the percentage of protein using the NIR transmission spectra. The data in the second example are used to evaluate our Bayesian discriminant model, which aims to allocate new samples to one of the nine varieties.

3.6 NIR Calibration

The purpose of NIR calibration is to predict the concentration of a constituent in an unknown sample by the spectral measurements. If the concentration of a constituent in a sample has a linear relationship with the absorbance/reflectance, a simple statistical approach is to create a regression model for the concentration of a constituent and the absorbance/reflectance and fit the model using observed data.

3.6.1 Linear Regression

Suppose we collect n samples of calibration data (Y_1, X_1), (Y_2, X_2), $\ldots$, (Y_n, X_n), where $X_i = (X_{i,1}, \ldots, X_{i,p})$ and $Y_i = (Y_{i,1}, \ldots, Y_{i,q})$ are the vectors of the absorbance at p wavelengths and the vector of the concentrations of q constituents of the i th sample, respectively. Let $Y = (Y_1^t, Y_2^t, \ldots, Y_n^t)^t$ and $X = (X_1^t, X_2^t, \ldots, X_n^t)^t$ so that Y is an n by q matrix and X is an n by p matrix. We will follow the common practice of subtracting sample means from both X and Y and using models with mean zero to simplify results. As a result, we assume Y and X are mean-corrected variables.

There are two ways of thinking about the calibration modelling: do we regress concentration on absorbance or do we regress absorbance on concentration? In considering a causal model, one should regress absorbance on concentration

since the difference in absorbance is due to the difference in the concentration of a constituent. The model is written as

$$X = Y\eta + F, \tag{3.4}$$

where η is a q by p regression coefficient matrix and F is a matrix of random errors. In an ideal calibration case when the spectrum of every constituent in the sample is known, say $\eta_1, \eta_2, \ldots, \eta_q$, with each one a $1 \times p$ vector, let $\eta = (\eta_1^t, \eta_2^t, \ldots, \eta_q^t)^t$, and the covariance of each spectrum X_i be Σ (> 0). Then, if $\eta\Sigma^{-1}\eta^t$ is invertible, Y_f, the concentrations of q constituents of a future sample can be predicted using a weighted least square estimator

$$X_f\Sigma^{-1}\eta(\eta\Sigma^{-1}\eta^t)^{-1}$$

where X_f is the spectrum of the future sample. In practice, one may need to estimate η by its LS estimator

$$\hat{\eta} = (Y^tY)^{-1}Y^tX$$

if $(Y^tY)^{-1}$ exists.

Alternately, since one would like to predict the concentrations using spectra, one might regress Y on X using the model

$$Y = X\beta + E, \tag{3.5}$$

where β is the regression coefficient matrix and E is the error. If X^tX is invertible (which will not usually be the case), the LS predictor for Y_f would be simply $\hat{Y} = X_f\hat{\beta}$ where $\hat{\beta} = (X^tX)^{-1}X^tY$. A detailed discussion of whether one should use the causal model (3.4) or a direct one (model 3.5) is given in Martens and Næs [94]. Practically, the latter model is usually preferred in NIR calibration. The reasons are explained in Osborne *et al.* [100]. In this thesis, we consider the latter model as an application for our Bayesian multiple linear regression.

Normally, the number of calibration samples is larger then the number of concentrations we would like to predict. That is, Y^tY is usually invertible. Hence, the existence of the prediction of Y in both approaches is subject to the invertibility

of $\eta\Sigma\eta^t$ and X^tX. When the number of calibration samples is larger than the number of wavelengths, X^tX is unfortunately singular. When the number of constituents is larger then the number of wavelengths, $\eta\Sigma\eta^t$ is singular. For the direct model, methods for reducing the number of variables are required in order to apply the least-squares approach. Prior knowledge about the spectra becomes important for pre-selection or refinement from the large variable group. However, it is inevitable that deleting variables loses information. In order to keep most of the information in full spectra, many 'regularised methods' have been developed in order to compress information in all the variables into few new variables. These methods will be introduced in the next section.

3.6.2 Regularised Regression

In order to use the information in the full spectra more efficiently, some methods have been developed to compress most of the information in a spectrum $(A_{\omega_1}, A_{\omega_2}, \cdots, A_{\omega_p})$ at wavelengths $(\omega_1, \omega_2, \cdots, \omega_p)$ into fewer variables. The two most popular methods are principal components regression (PCR) and partial least squares regression (PLSR). These two methods use linear combinations of $(A_{\omega_1}, A_{\omega_2}, \cdots, A_{\omega_p})$ as new variables, denoted as $(S_1, S_2, \cdots, S_r)$, which contain most of the information in the original spectrum. The number of new variables r can be controlled so that it does not exceed the number of samples n. Continuum regression (CR) is another method using linear combinations of the original spectrum that links together MLR, PCR and PLSR. PCR, PLSR and CR include two steps: compressing information into fewer new variables and regressing explanatory variables on new variables. An alternative approach is ridge regression (RR), which does not have a compressing step. It uses the data in the original form.

Principal components regression regresses the response variables on the principal components (PC's) of the original explanatory variables, using a number of PC's less and often very much less than the original number of regressors. Suppose X contains the observations of n spectra at p wavelengths as in section 2.5.1. The first PC S_1 is associated with the first loading vector p_1, such that p_1 is a unit

vector ($p_1^t p_1 = 1$) and is chosen to maximise the variance of the score $S_1 = Xp_1$. The second loading vector p_2 is also normalised and is chosen to maximise the sample variance of $S_2 = Xp_2$ under the constraint that p_1 and p_2 are orthogonal. The procedure continues in this way under the constraint that each new PC is orthogonal to the previous PCs. It can be shown that p_i is an eigenvector of X^tX corresponding to the i th largest eigenvalue, and the vector of eigenvalues $(\lambda_1, \lambda_2, \cdots \lambda_r)$ is proportional to the sample variance of $(S_1, S_2, \cdots, S_r)$. Thus, a PC with smaller index number contains more of the variation in X. Usually, the first few PCs include most of the information in X. The maximum number of principal components included in a full model is $n - 1$ so that the LS or MLE estimates of regression coefficients always exist. The best sub-model can then be chosen by variable selection procedures.

The step of creating new regressors in PCR is principal components analysis (PCA). The criterion for creating a new regressor is based on the ability of the new regressor to explain the variation in the spectral data, and does not take into account its ability to predict the dependent variables at all. PLSR follows the different philosophy that a new regressor should be able to explain the dependent variables well. Therefore, the regressors created at an early stage are never less important (for prediction) than the regressors created later in a regression model, while this is not necessary true in PCR. Therefore, PLSR usually ends up with fewer explanatory variables than a PCR model requires (Martens and Næs [94]).

PLSR is developed from Herman Wold's iterative fitting methods. Different algorithms developed later by different authors may create a different set of regressors. For examples, the algorithm developed by Svante Wold creates models with orthogonal scores, while Martens' algorithm creates models with orthogonal loadings. These two algorithms are provided in Martins and Næs [94]. Mathematical interpretation of PLSR was given several years later. The following explanation is based on Stone and Brooks' [120] paper, also described in Brown [23].

Suppose the data matrix of explanatory variables is X with n samples and p explanatory variables, and Y is a column vector containing n observed values for

response variables. The first new factor is $S_1 = Xp_1$ where the loading weight p_1 is normalised and is chosen to maximise $S_1^t Y$. The second new factor $S_2 = Xp_2$ is orthogonal to S_1 and p_2 is chosen to maximise $S_2^t Y$. Following the same procedure, we can obtain r new factors $S_1, S_2, \cdots, S_r$ for the regression model, where r is usually chosen by cross-validation. PLSR can be used for multivariate regression as well.

Stone and Brooks [120] proposed continuum regression, which creates a link between MLR, PCR and PLSR with a continuous parameter γ. The method is also based on using linear combinations of the original variables as new variables. The new variables are constructed by choosing a loading vector p which maximises

$$(p^t X^t Y)^2 (p^t X^t X p)^{\gamma - 1},$$

and the new variable is Xp. When $\gamma = 0$, continuum regression gives MLR by considering only the first linear combination, which maximises the correlation between the new variable and Y. When $\gamma = \infty$, it corresponds to PCR since the procedure is to maximise the variance of the new variables; when $\gamma = 1$, the process maximises the covariance between new variables and Y giving PLSR. Stone and Brooks suggested the use of cross-validation to choose the most appropriate γ. It is not hard to see that continuum regression is computationally intensive.

Ridge regression provides a way to regress Y on the full spectra. Consider regressing Y on X. The LS and ML solution for the regression coefficient β is $\widehat{\beta} = (X^t X)^{-1} X^t Y$. The estimate $\widehat{\beta}$ does not exist when $X^t X$ is singular, i.e. is not invertible. The principle of RR is to adjust the singular $X^t X$ by adding a matrix kI, where k is a scalar constant small in comparison with the diagonal of $X^t X$, and I is the identity matrix. The estimator of β is therefore $\widehat{\beta}_{RR}(k) = (X^t X + kI)^{-1} X^t Y$. The value of k is chosen to stabilise the ridge trace (the curves of regression coefficients as functions of k) whilst not penalising the residual sum of squares (which is also a function of k) too much. One may also apply cross-validation as a method to choose k. A detailed discussion of ridge regression can be found in Brown [23].

3.7 NIR Discriminant Analysis

Another important application of NIR spectroscopy is classification of a sample as belonging to one of several classes. The number of classes is usually known, but the distribution of the measurement of samples within each class needs to be learnt from training data. There are various ways to discriminate between samples. The problem in discriminant analysis with many variables is the same as the problem in regression analysis with many variables, i.e. (X^tX) is not invertible. Similar strategies to those in regression have been considered in discriminant analysis. One may use PC's of X as new variables to reduce the number of variables. One may also use a strategy similar to RR to improve the condition of the matrix that needs to be inverted.

3.7.1 Probability Approach

A standard statistical approach is to allocate a sample according to a criterion decided by the stochastic properties of the groups. Each group has a distribution for the measurements of the samples from the group, and the parameters in the distributions often have to be estimated using the training data. A discrimination procedure is chosen as a decision criterion that should minimise the loss due to misclassification. This procedure is called a Bayes' procedure [2], and has been introduced in section 2.5.4. Suppose a sample $\mathfrak{X}$ must be from one of several groups, labelled as group 1, group 2, ..., group g. The sample is measured quantitatively as X. According to equation (2.3), the probability of $\mathfrak{X}$ being in group i is

$$p(\mathfrak{X} \text{ being in group } i) = \frac{q_i\, p_i(X)}{\sum_{i=1}^{g} q_i\, p_i(X)}, \tag{3.6}$$

where q_i is the prior probability of $\mathfrak{X}$ being in group i, $p_i(X)$ is the probability of observing X for a sample from group i and g is the total number of groups. The risk of misclassification can be derived using equation (3.6) and a loss function of missclassification. Suppose the loss is a constant, then Bayes' procedure allocates a sample to the group with the maximum posterior probability.

3.7.2 Linear and Quadratic Discriminant Functions

Consider the case where there are only two groups and X is normally distributed in both groups. The ratio of the posterior probability of $\mathfrak{X}$ being in group 1 and group 2 is used to allocate $\mathfrak{X}$ to one of the two groups. Following the Bayes' procedure, if $p_1(X)/p_2(X) > k$ then $\mathfrak{X}$ is assigned to group 1, where

$$k = \frac{q_2 C(1|2)}{q_2 C(2|1)}$$

and $C(i|j)$ is the loss of assigning $\mathfrak{X}$ to the i th group when $\mathfrak{X}$ is actually from the j th group. When the two groups have the same covariance matrix, the Bayes' procedure leads to linear discriminant analysis, which allocates $\mathfrak{X}$ according to a linear discriminant function

$$X^t \Sigma^{-1}(\mu_1 - \mu_2), \tag{3.7}$$

where Σ is the common covariance matrix and μ_1 and μ_2 are the means of group 1 and 2 respectively. Observing $X = x$, if

$$x^t \Sigma^{-1}(\mu_1 - \mu_2) \geq \frac{1}{2}(\mu_1 + \mu_2)^t \Sigma^{-1}(\mu_1 - \mu_2) + \log k,$$

$\mathfrak{X}$ is classified to group 1, otherwise to group 2. When the covariance matrices of the two groups are different, the same Bayes' procedure leads to quadratic discriminant analysis, where X is classified according to the value of the quadratic discriminant function

$$(X - \mu_2)^t \Sigma_2 (X - \mu_2) - (X - \mu_1)^t \Sigma_1 (X - \mu_1),$$

where Σ_1 and Σ_2 are the covariance matrices of groups 1 and 2 respectively.

3.7.3 Distance Based Methods

Distance based methods discriminate between samples by comparing the distance of the sample from the centre of each group and picking the nearest group. In order to measure the distance between two points in the space, a distance metric needs to be defined. The Euclidean distance between two points $(x_1, x_2, \ldots, x_m)$

and $(y_1, y_2, \ldots, y_m)$ in $\mathbf{R}^m$ is simply $[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_m - y_m)^2]^{0.5}$. The most frequently used distance in discriminant analysis with many variables is the Mahalanobis distance. The Mahalanobis distance from X to group i is $d_i = \sqrt{(X - \mu_i)^t \Sigma_i^{-1} (X - \mu_i)}$. For other definitions of distance, refer to [95]. Comparison of the Mahalanobis distance of X to different populations is closely related to the methods of linear and quadratic discriminant analysis.

3.7.4 Estimation of Mean and Covariance Matrix

The means and covariance matrices of the groups are usually unknown and have to be estimated. Sample means and sample covariance matrices are the most straightforward estimators of the means and covariance matrices. When the number of variables is larger than the number of samples, the sample covariance matrix is singular and the above methods do not work. Therefore, one might like to compress the data into PCA scores with a small number of PCs', then apply the above discriminant methods. This is the discrimination equivalent of PCR.

Friedman [57] proposed regularised discriminant analysis (RDA) as a compromise between linear and quadratic discrimination. The estimator for a covariance matrix he used is a combination of individual and pooled sample covariance matrices and a specified matrix, such as an identity matrix. The identity matrix has the same effect as in ridge regression in that it improves the condition of the estimated covariance matrix.

3.8 Remark

In section 2.5.2 we introduced continuum regression which integrates MLR, PCR, and PLSR. Other researchers also have linked different methods. Principal covariates regression, proposed by de Jong *et al.* [48] links MLR, PCR, and PLSR via a continuous path which is different from that of Stone and Brooks. Höskuldsson [81] showed that PCR and PLSR are related by the Heisenberg principle of mathematical modelling. Sundberg [121] demonstrated the relation between first-factor CR

and RR and argued that first-factor CR is preferable in principle.

NIR calibration is mainly based on the linear relationship between dependent variables and spectra. However, nonlinearities are sometimes observed. Weighted regression, nonlinear regression and non-parametric regression [94] are possible methods for handling it. The artificial neural network is a framework which provides more solutions for nonlinear problems [100]. We will not consider or explain these alternatives here.

Chapter 4

Markov Chain Monte Carlo

4.1 Introduction

Monte Carlo integration using Markov chains, also called Markov chain Monte Carlo (MCMC), is a very important computing tool in high-dimensional modelling, where we need to integrate over high-dimensional probability distributions to make inference on unknown random quantities, for example to calculate the marginal probability distributions of unknown quantities. Frequently, analytical integration in practical modelling is not possible. Traditional numerical methods such as trapezoidal or Simpson's rule may work well for very low dimensional cases but become very inefficient when the dimension of the model is large. Instead, MCMC is a more efficient method of integration when the dimension of the parameter space is high. The history of the development of MCMC is not long, but MCMC has already been widely applied in many practical Bayesian data analyses. Applications can also be found in non-Bayesian cases.

MCMC consists of a sampling step and an integration step. Suppose we would like to calculate the expectation of $s(X)$ where X is a random quantity with probability density function (pdf) f. In the sampling step, MCMC draws a sequence of samples $X_1, X_2, \ldots, X_n$ from f. In the integration step, it estimates $E(s(X))$ using $n^{-1}\sum_{i=1}^{n} s(X_i)$. Direct generation of independent samples from f can be difficult or even impossible. MCMC draws samples from a cleverly designed

ergodic Markov chain such that it is easier to generate samples from the Markov chain than from the true distribution of the samples. The Markov chain is iterated for a long time so that the chain is eventually in the equilibrium stage. The samples generated at the equilibrium stage are generated from the equilibrium distribution, which is equal to the true density function of the samples. These samples are then used in the Monte Carlo integration step.

Theoretically, a Markov chain reaches the equilibrium stage after an infinite number of iterations. It is not possible to wait for a infinitely long time to collect the samples. Fortunately, adequate convergence may be reached with finite iterations in many cases. Practically, we need to decide a sufficiently long *burn-in*, say the first m iterations, which is the early stage of the Markov chain when the chain has not yet converged to the stationary distribution. The Markov chain approximately converges after the *burn-in* period. Suppose the total length of the chain is n, and $(X_1, X_2, \cdots X_m, X_{m+1}, \cdots X_n)$ is the entire sequence of samples generated by the Markov chain. The *burn-in* samples $X_1, \ldots, X_m$ are discarded and $E(s(X))$ is estimated by $(n-m)^{-1}\sum_{i=m+1}^{n} s(X_i)$, which is called the *ergodic average*. Many rules have been developed to detect a non-convergent chain, but there is no way to guarantee that the chain has definitely converged. With proper checking, one can still make a reasonable estimation using the samples generated after the *burn-in* section even though we may not be able to guarantee the chain has definitely converged. There are many techniques for checking convergence. Cowles and Carlin [36], Brooks and Roberts [22], and Mengersen *et al.* [96] provide summaries of the techniques for monitoring convergence of Markov chain simulations.

There are many ways of constructing a Markov chain. According to the properties of the models, different models require different methods in order to achieve efficient and reliable estimation. Most of these methods are based on the algorithms developed by Metropolis and Hastings [79]. These methods generate samples from a fixed dimensional space. The Gibbs sampler named by Geman and Geman [68] is an example of a Metropolis-Hastings algorithm. Green [76] proposed the reversible-jump MCMC algorithm, which is a generalisation of MCMC using

a Metropolis-Hastings algorithm. It allows proposal and target density functions having different dimensions, i.e. samples from reversible-jump MCMC sampler may have different dimensions. It is especially useful for modelling mixture distributions with an unknown number of components and Bayesian variable selection.

It is known that the convergence assessment of MCMC can never guarantee whether a chain has actually converged or not. However, MCMC strategies that guarantee samples can be generated from their exact target density function within a finite number of iterations have been discovered. Propp and Wilson [105] first proposed an exact sampling algorithm using backward coupled Markov chains. Exact sampling, sometimes called perfect sampling, is still a very new development in MCMC, in comparison with the 'traditional' MCMC approaches. Variations and generalisation of exact sampling are already available. However, they have so far only been successfully applied to low dimensional cases and some high dimensional models with very particular structure (Green and Murdoch [77]).

The development of MCMC strategies is a very active area since models with different properties and difficulties require different special techniques in order to sample from them correctly and efficiently. Gilks *et al.* [73] provide a broad review of MCMC related topics. Gelman *et al.* [65] cover both simulation-based and non simulation-based posterior inference techniques. Besag's [13] paper reviews the most general and up-to-date developments in the research of MCMC. We shall not attempt to review all techniques in this chapter, but only focus on the strategies we apply to the examples in this thesis. In this chapter we first introduce some basic non-iterative samplers and describe the general framework of MCMC for generating from continuous distributions. Then we introduce the method we apply in our examples: the adaptive rejection Metropolis sampling (ARMS) within Gibbs sampling for MCMC introduced by Gilks *et al.* [72]. In our examples we use multiple-chain MCMC as suggested by Gelman and Rubin [67], which uses variance ratio methods for convergence assessment. In hierarchical Bayesian models, some parameters are so highly correlated that it is almost impossible to observe convergence with a short single chain MCMC. Proofs of methods are not given in

this thesis.

4.2 Direct Sampling

Many random generators have been developed for sampling independent random numbers. Many of these random generators are reviewed in Ripley [109]. Almost all of them generate scalar random quantities. When the inverse cdf of a distribution exists, one can sample from this distribution directly. Many random samplers are specialised for particular distributions. General methods also exist. Some MCMC schemes for multivariate cases, for example, the Gibbs sampler, require the use of these random samplers. In this section, we introduce some general methods for generating scalar random numbers.

Inverse CDF: Continuous Cases

Suppose we want to sample a random number X from a continuous pdf f, whose cdf is F, and the inverse function F^{-1} or a good approximation to F^{-1} exists analytically. Let $x = F^{-1}(u)$, where u is sampled from $U(0,1)$. Then x is a sample for X.

Inverse CDF: Discrete Cases

There are two cases in which one would like to generate random numbers from a discrete distribution $P_k = P(X \leq k)$, $k = 1, 2, \ldots$: firstly, when the random quantity is on a discrete space; secondly, when we would like to use a discrete distribution to approximate a continuous distribution.

Suppose X is a discrete scalar random quantity with cdf F. The inverse cdf of X is then $F^{-1}(x) = \min\{x|F(x) > u\} = i$ where $P_{i-1} < u < P_i$. In order to generate a number for X, one first generates u from $U(0,1)$. Then, a proper searching algorithm is applied to search a value r so that $P_r < u < P_r + 1$. Then r is the sample for X.

Rejection Sampling

Suppose we want to sample X from a pdf f whose inverse cdf is not available analytically. Rejection sampling is a simple way to sample from f. In Bayesian modelling, it is very often the case that the distribution of X is known up to an unnormalised density function, say k such that $k \propto f$. Instead of sampling X from f, we sample a candidate Y from a pdf g for which there exists a constant M such that $k \leq Mg$ for every X in the sample space. The candidate Y is accepted as a sample of X with probability

$$\frac{k(Y)}{M \times g(Y)}.$$

The unnormalised density function Mg is called an envelope function of k. The algorithm for generating from f is

Algorithm 4.1 *Rejection Sampling*

Repeat

Generate Y from g;

Generate U from $U(0,1)$;

If $MU \leq k(Y)/g(Y)$ accept Y;

Until a Y is accepted.

Return $X = Y$.

It can be proved that X is drawn from the pdf f (see Ripley [109]).

An ideal envelope function should be a function which is nearly proportional to f. If g is equal to f, every draw will be accepted with probability 1. If M is very large, most of the draws will be rejected. Computing time can be saved by using squeezing functions $a(Y)$ and $b(Y)$, where $a(Y) \geq g(Y) \geq b(Y)$ for all Y such that the squeezing functions are easier to calculate than g. The if-statement in Algorithm 4.1 is modified to

If $MU > a(Y)/g(Y)$ reject Y;

else if $MU \leq b(Y)/g(Y)$ accept Y;

else if $MU \leq f(Y)/g(Y)$ accept Y.

4.3 Basic Markov Chain Simulation

4.3.1 General methods

MCMC has been shown to be a reliable and convenient tool in many applications. In MCMC, the pdf needs only to be known up to a constant of proportionality, just as in rejection sampling. Since very often the joint posterior distribution of the parameters in a Bayesian model is very complicated, and these posterior distributions are known only up to unnormalised density functions. MCMC is an important tool in Bayesian modelling.

Suppose we would like to generate samples of X whose probability density function is $f(X)$. In Markov chain simulation, we generate a sequence of random quantities $X_0, X_1, X_2, \cdots$. At each time t, X_t is generated from density function $P(X_t|X_{t-1})$, which should converge to a unique stationary density function $\pi(X)$ as t goes to infinity, and the chain has been set up so that $\pi(X)$ is equal to $f(X)$. The transition kernel $P(X_t|X_{t-1})$ is constructed from a proposal density function g and a candidate-acceptance probability $\alpha(X, Y)$ so that

$$P(X_t|X_{t-1}) = g(X_t|X_{t-1})\alpha(X_{t-1}, X_t),$$

and they have to satisfy the detailed balance equation

$$\pi(X_{t-1})P(X_t|X_{t-1}) = \pi(X_t)P(X_{t-1}|X_t).$$

The methodology was first developed by Metropolis in 1953 [97], and was generalised by Hastings in 1970 [79].

Metropolis-Hastings Method

The Metropolis-Hastings method was proposed by Hastings. The original method of Metropolis can be seen as a special case of Hastings' method. Let

$$\alpha(X, Y) = \min\left(1, \frac{\pi(Y)g(X|Y)}{\pi(X)g(Y|X)}\right) \tag{4.1}$$

where π is the target density and $g(\cdot|\cdot)$ has to be carefully chosen so that MCMC is efficient, although theoretically it can have any form. To generate X_t, we sample a candidate Y from $g(\cdot|X_{t-1})$. The candidate is accepted with probability

$\alpha(X_{t-1}, Y)$. Once Y is accepted, set $X_t = Y$, otherwise, $X_t = X_{t-1}$. This assumption satisfies the detailed balance equation (see Gilks, Richardson and Spielhalter [73]). The algorithm is

Algorithm 4.2 *Metropolis-Hastings Algorithm*

Initialise X_0; *set* $t = 0$.

Repeat

Generate Y *from* $g(\cdot|X_t)$,

Generate U *from* $U(0,1)$,

If $U \leq \alpha(X_t, Y)$ *set* $X_{t+1} = Y$

otherwise set $X_{t+1} = X_t$.

Increment t.

The Metropolis method was proposed by Metropolis before Hastings generalised it as the Metropolis-Hastings algorithm. It imposes the condition $g(X|Y) = g(Y|X)$. As a result, the candidate-acceptance probability (4.1) is simplified as

$$\alpha(X, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X)}\right),$$

and the algorithm 4.2 becomes the Metropolis algorithm.

4.3.2 Full Conditional Distribution and Gibbs Sampling

In the previous section we sampled X as a whole. When X is a long vector, it is rarely possible to sample efficiently or reliably in this way. Sometimes, it is more computationally efficient to divide X into several smaller components or sub-vectors and update X component by component because it is much easier to generate one small component given all the other components than to generate the entire X. Suppose we divide X into h smaller components $X_1, X_2, \ldots, X_h$ so that $X = (X_1, X_2, \ldots, X_h)$. There will be h steps in each iteration. Denote the i^{th} component of X at the t^{th} iteration as $X_{t,i}$ and let $X_{(t)}^{[i]} = (X_{t+1,1}, X_{t+1,2}, \ldots, X_{t+1,i-1}, X_{t,i+1}, X_{t,i+2}, \ldots, X_{t,h})$. At the i^{th} step in the $(t+1)^{\text{th}}$ iteration, $X_1, X_2, \ldots, X_{i-1}$ have been updated, while $X_{i+1}, \ldots, X_h$ have not yet. Therefore, $X_{t,i}$ and $X_{(t)}^{[i]}$ form the current set of X at this stage. A sample

for X_i is then generated with a distribution conditional on current values of all the other components, which is $X_{(t)}^{[i]}$. Such a procedure is called the *single-component Metropolis Hastings algorithm* in Gilks *et al.* [73].

To generate X_i at the i^{th} step of the t^{th} iteration, a candidate Y_i is sampled from a proposal density function $g_i(X_{t+1,i}|X_{t,i}, X_{(t)}^{[i]})$, then Y_i is accepted as new X_i with probability

$$\alpha(X_{(t)}^{[i]}, X_{t,i}, Y) = \min\left(1, \frac{\pi(Y_i|X_{(t)}^{[i]})g_i(X_{t+1,i}|Y_i, X_{(t)}^{[i]})}{\pi(X_{t+1,i}|X_{(t)}^{[i]})g_i(Y_i|X_{t,i}, X_{(t)}^{[i]})}\right),$$

where $\pi(X_i)$ is the target density function of X_i. Let $X^{[i]} = (X_1, \ldots X_{i-1}, X_{i+1}, X_h)$. The conditional density function $\pi(X_i|X^{[i]})$ is called the full conditional density function of X_i.

Gibbs sampling is a special case of the single-component Metropolis-Hastings method, whose proposal density function $g_i(Y_i|X_i, X^{[i]})$ is equal to $\pi(Y_i|X^{[i]})$, the full conditional density function of X_i. The acceptance probability α is always 1. A direct sampling method is then applied in order to sample from $\pi(Y_i|X^{[i]})$. Gibbs sampling is efficient if it is easy to sample from the full conditional density functions.

4.4 ARS and ARMS

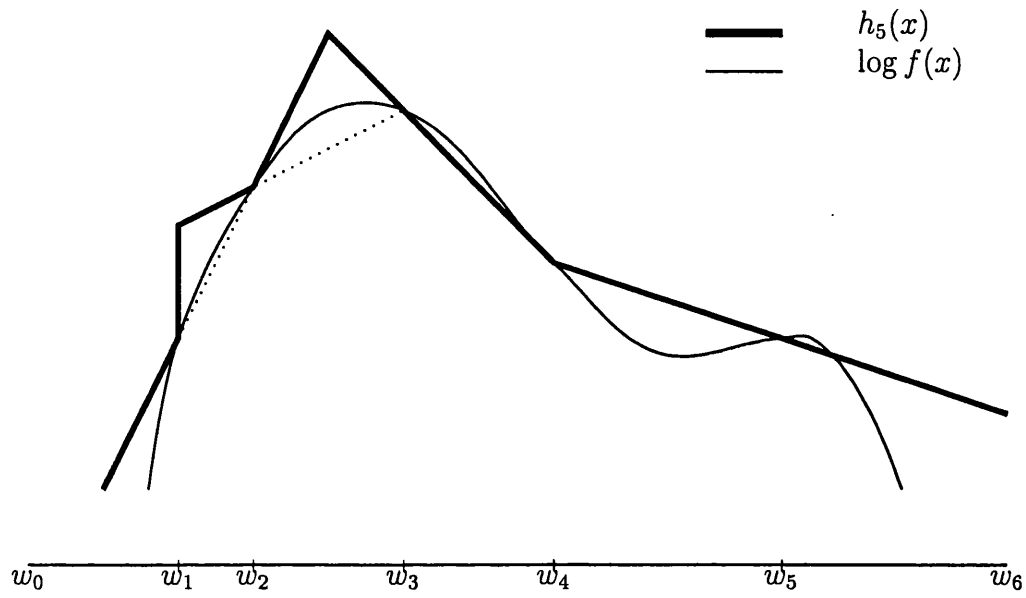
The adaptive rejection sampling (ARS) scheme, proposed by Gilks and Wild [75], and the adaptive rejection Metropolis sampling (ARMS) scheme, developed by Gilks, Best and Tan [72] are designed to improve the speed of Gibbs sampling. In Gibbs sampling, the full conditional density function for each component is derived and samples are drawn from the full conditional density functions one by one in each iteration. Frequently, it is necessary to use a rejection sampling scheme to sample from a full conditional density function. The acceptance rate of rejection sampling depends on the envelope function we choose. ARS is proposed only for sampling from log-concave target density functions, while ARMS works for general target densities. They have better acceptance rates for the candidate in

each iteration within Gibbs sampling than a standard rejection sampling scheme. As a result, Gibbs sampling using ARS and ARMS may be more efficient than Gibbs sampling using standard rejection sampling.

The ARS and ARMS schemes try to improve the efficiency of MCMC by combining extra information from the target density function and the envelope or proposal density function to create a new envelope or proposal density function each time a candidate sample is rejected. The updated envelope or proposal density function provides a higher acceptance rate for the next candidate sample. The envelope functions and proposal density functions are carefully designed so that they are close to the target density function and can be updated systematically after each rejection. Gilks and Wild [75] and Gilks, Best and Tan [72] choose the envelope and proposal density functions so that the log envelope and the log proposal density functions are piecewise linear and continuous, so a sample can be easily drawn from them.

The ARS scheme is based on a rejection sampling scheme. In the ARS scheme, an envelope function is generated automatically at the beginning. The envelope function is updated after each time a candidate is rejected. A new candidate is then generated with the updated envelope function. The procedure is repeated until one candidate is accepted. The ARMS scheme consists of two stages: the first stage is the ARS algorithm, and the second stage is a one-step Metropolis-Hastings algorithm. The ARS procedure is applied to generate and update the proposal density function. A candidate is generated from the current proposal density function at each iteration. If the candidate is rejected, then the proposal density function is updated. Then a new candidate is generated from the updated proposal density function. The procedure is repeated until one candidate is accepted. Since the proposal density function is not an envelope function, the accepted candidate is not simply the final sample we accept. The candidate accepted in the ARS stage is then judged by a one-step Metropolis-Hastings. The candidate is accepted as an ARMS sample if it is accepted by the Metropolis-Hastings step. Otherwise, the sample generated at the previous iteration is kept.

Figure 4.1: Adaptive rejection function $h_5(x)$ of $f(x)$ with $S_5 = \{w_0, \ldots, w_6\}$



The ARS scheme is design to sample from log-concave target density function, while the ARMS scheme works for general cases. The ARMS algorithm proposed by Gilks , Best and Tan [72] can automatically reduce to an ARS algorithm when the target density function is log-concave.

Define the adaptive rejection function h_k of $f(x)$ as follows [72]:

1. Let $S_k = \{w_0, w_1, \cdots, w_{k+1}\}$ denote the current set of abscissae in ascending order, where w_0 and w_{k+1} are the possibly infinite lower and upper limits of the sample space $\mathcal{X}$.

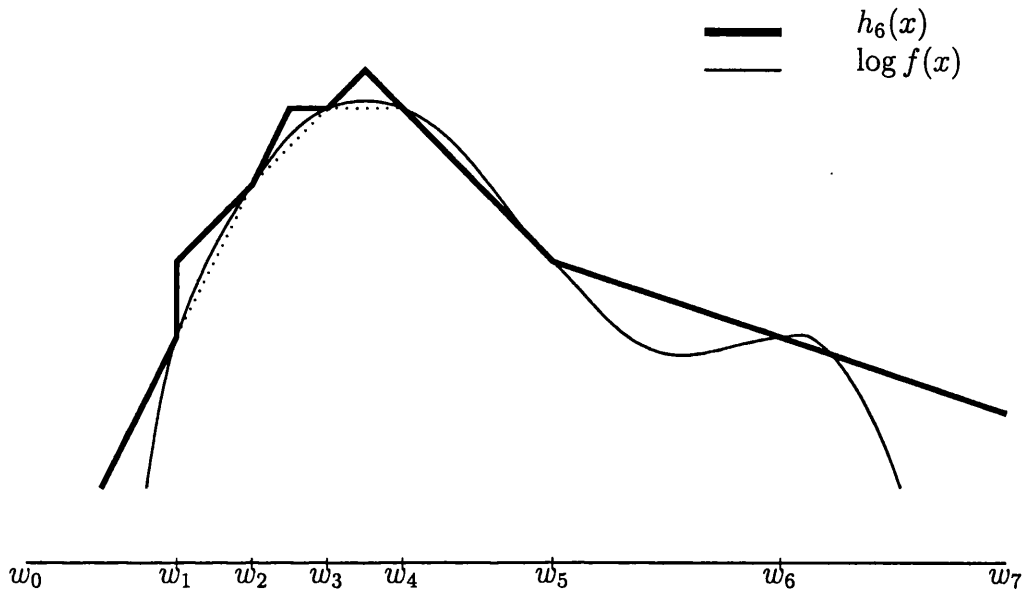
2. For $1 \leq i \leq j \leq k$ let $L_{i,j}(x; S_k)$ denote the straight line through points $[w_i, \ln f(w_i)]$ and $[w_j, \ln f(w_j)]$. For other (i, j), $L_{i,j}$ is not defined.

3. Define a piecewise linear function $h_k(x)$:

$$h_k(x) = \max[L_{i,i+1}(x; S_k), \min\{L_{i-1,i}(x; S_k), L_{i+1,i+2}(x; S_k)\}], \quad w_i \leq x < w_{i+1}, \tag{4.2}$$

where $\min(a, b) = \min(b, a) = \max(a, b) = \max(b, a) = a$ if b is undefined, and h_k depends on S_k.

Figure 4.2: Adaptive rejection function $h_6(x)$ of $f(x)$ in figure 4.1, where w_3 is the rejected value in step 3 of ARMS.



4. If the sample space $\mathcal{X}$ is not bounded on the left, the abscissae have to be chosen so that the gradient of $L_{1,2}(x; S_k)$ is positive. If $\mathcal{X}$ is not bounded on the right, the gradient of $L_{k-1,k}(x; S_k)$ is negative.

The adaptive rejection function $h_k(x)$ defined above is an envelope function when $f(x)$ is log-concave. When $f(x)$ is not log-concave, it is considered as a proposal density function. According to the definition, the function h_k is much closer to $\log f$ when k is larger, and if f is log-concave, $\exp h_k$ is an envelope function of f everywhere in $\mathcal{X}$. An example is given in figure 4.1. The graph shows the $h_5(x)$ of a non-log-concave function $f(x)$ and $\log f(x)$ when $S_5 = \{w_0, w_1, \ldots, w_6\}$. In this case, $\exp h_5(x)$ is a proposal density function instead of an envelope function of $f(x)$.

Consider an iteration of Gibbs sampling. Let $(X_1, X_2, \ldots, X_h)$ be the complete set of variables generated by the Gibbs sampler, and X_i be the current variable to be sampled from its full conditional density function $f(x_i)$ (simplified notation of $f(x_i|x^{[i]})$). Let X_{cur} denote the current value of x at a given iteration of the Gibbs sampler. The aim is to substitute for X_{cur} a new value X_M from f.

It is important that the starting abscissae is independent of X_{cur}. Let

$$g_k(x) = \frac{\exp h_k(x)}{\int \exp h_k(x)dx}.$$

The algorithm of ARMS is

Algorithm 4.3 *ARMS*

step 0, initialise k and S_k*;*

step 1, sample X from g_k*;*

step 2, sample U from $U(0,1)$*;*

step 3, if $U > f(X)/\exp h_k(X)$ *then* {

ARS rejection step:

set $S_{k+1} = S_k \cup \{X\}$*;*

relabel points in S_{k+1} *in ascending order;*

increment k and go back to step 1;}

else {

ARS acceptance step:

set $X_A = X$*;*}*;*

step 4, sample U from $U(0,1)$*;*

step 5, if $U > \min\left[1, \frac{f(X_A)\min\{f(X_{cur}), \exp h_k(X_{cur})\}}{f(X_{cur})\min\{f(X_A), \exp h_k(X_A)\}}\right]$ *then* {

Metropolis-Hastings rejection step:

set $X_M = X_{cur}$*;*}

else {

Metropolis-Hastings acceptance step:

set $X_M = X_A$*;*}*;*

step 6, return X_M*.*

Continue the example in figure 4.1. Now a candidate X is generated in step 1 and is then rejected in step 3, X is then added into S_5 to create S_6 (w_3 in figure 4.2 is the rejected X). The new adaptive rejection function is the $h_6(x)$ in figure 4.2. As one can see, $h_6(x)$ is closer to $\log f(x)$ than $h_5(x)$ to $\log f(x)$. ARMS is

in fact not a pure Gibbs sampling scheme but a more complicated Markov chain. ARMS is an application of the auxiliary variable method [14]. The proof that ARMS within Gibbs sampling yields a stationary Markov chain with the desired target distribution is given in [72].

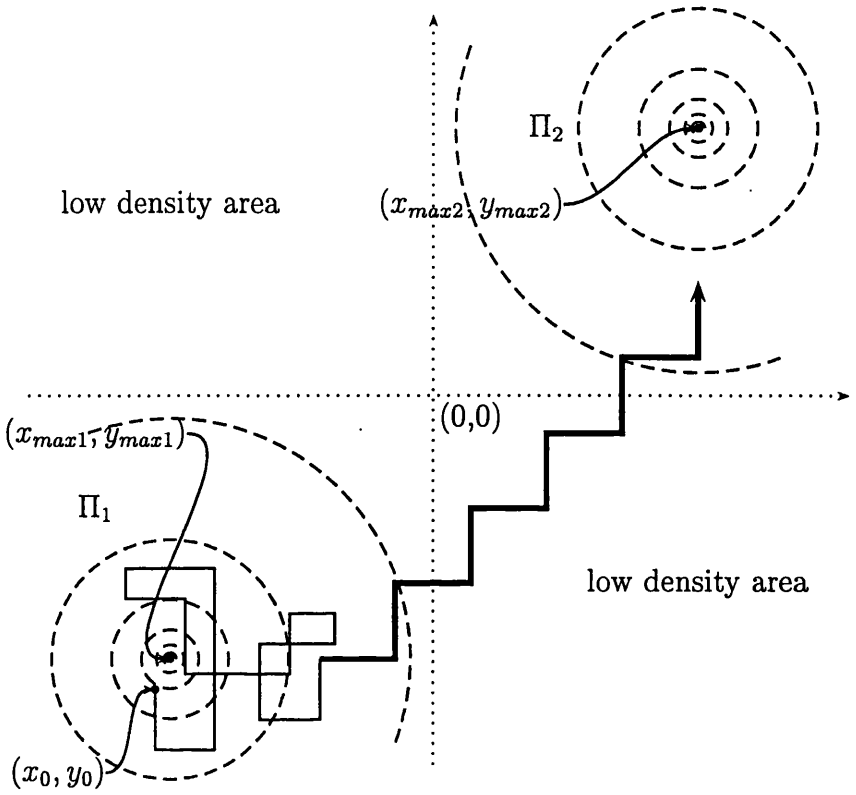
4.5 Multiple Sequences MCMC and Convergence Assessment

4.5.1 Multiple Sequences MCMC

In some cases, a single chain that has not been turned down by a convergence evaluation has not in fact converged to its target distribution. An example where this may happen is Gibbs sampling of a bivariate target distribution which has two well-separated and high density peaks along the diagonal of the plane. Suppose the two local maxima of the density function are at (x_{max1}, y_{max1}) for peak Π_1 in the 3rd quadrant and (x_{max2}, y_{max2}) for the peak Π_2 in the 1st quadrant, and the initial point (x_0, y_0) is located near (x_{max1}, y_{max1}). The 1st and the 3rd quadrants are areas with higher density and while the 2nd and the 4th quadrants are areas with lower density (see figure 4.3).

Before the sequence can travel to Π_2, it has to travel towards either the 2nd quadrant along the y-axis or the 4th quadrant along the x-axis. However, it is less likely to generate samples along the x-axis or y-axis away from a position near a local maximum. Figure 4.3 illustrates a possible sampling path from the initial point (x_0, y_0) near (x_{max1}, y_{max1}) towards peak Π_2. In order to get close to Π_2, the Gibbs sampler needs to generate a lot of samples in the lower density area given its previous sample in a comparatively higher density area. Thus, it is very difficult for the sequence to travel from the initial point at Π_1 to Π_2, and vice versa. On the other hand, it is possible for a Gibbs sampling sequence to travel only in peak Π_1. Consequently, the distribution estimated by these samples would often not be the true bimodal target distribution, but a uni-modal distribution.

Figure 4.3: The contour and a sampling path of Gibbs sampling for a bimodal bivariate distribution with well-separated peaks.



In two dimensional cases, we may be able to judge whether the estimation is right or not since we may already know the shape of the target density functions. However, it is very difficult to judge whether the estimation is right in high dimensional cases, because it is very difficult to know the shape of the target density function. In order to try to prevent this problem, Gelman and Rubin [67] suggested running several chains with overdispersed starting points and detect whether all chains converge to the sample target density function.

For multiple chain simulation, Gelman and Rubin [67] evaluated the convergence by comparing the within-sequence variance and the overall variance. When the ratio of these variances is far from one, the multiple sequences have not yet converged to the same distribution. If the sequences do not converge to the same distribution further strategies have to be considered [73]. The method using the ratio of the two variances is called the variance ratio method by Brooks and Roberts [11].

When some of the parameters that we want to learn about are highly correlated, a single sequence may wander around the sample space with high autocorrelation. Samples from such chains with finite iterations may approximate the target distribution well but the convergence assessment for the finite single chain fails. With multiple chains, one is able to detect whether the samples approximate the target distribution well.

4.5.2 Convergence Assessment: Variance Ratio Methods

The original variance ratio method was introduced by Gelman and Rubin [67], and later generalised by Brooks and Gelman [21]. For any scalar summary statistic ψ, which might be a function of several parameters, let ψ_{ij} denote the j^{th} value of ψ in the i^{th} chain. To implement these methods, a variance ratio is defined and $m > 2$ (typically four or more) independent sequences of MCMC are run for $2n$ iterations with n chosen so that the samples in the second half of the chains have a variance ratio less than 1.2 (a criterion suggested by Gelman [64]). The first n iterations are considered to be the *burn-in* period. Define

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\psi_{i.}-\psi_{..})^2, \quad \text{where} \quad \psi_{i.}=\frac{1}{n}\sum_{j=n+1}^{2n}\psi_{ij}, \psi_{..}=\frac{1}{m}\sum_{i=1}^{m}\psi_{i.}$$

$$W = \frac{1}{m}\sum_{i=1}^{m}s_i^2, \quad \text{where} \quad s_i^2=\frac{1}{n-1}\sum_{j=n+1}^{2n}(\psi_{ij}-\psi_{i.})^2,$$

where B/n is the variance between m sequences with means $\psi_{i.}$, and W is the mean of the m within-sequence variances s_i^2, which generally underestimates the variance of σ^2 because the individual sequences have not had time to explore all possible ψ. Then, define

$$\widehat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}B,$$

which overestimates σ^2, assuming the starting distribution of ψ is appropriately overdispersed. Gelman and Rubin [67] define

$$\sqrt{\widehat{R}} = \sqrt{\frac{\widehat{V}}{W}\frac{d}{d-2}},$$

where $\widehat{V} = \widehat{\sigma}^2 + B/(mn)$, $d = 2\widehat{V}^2/\widehat{\text{var}}(\widehat{V})$ and

$$\widehat{\text{var}}(\widehat{V}) = \left(\frac{n-1}{n}\right)^2 \frac{1}{m}\widehat{\text{var}}(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1}B^2$$

$$+2\frac{(m+1)(n-1)n}{m^2n^2}[\widehat{\text{cov}}(s_i^2, \psi_i.^2) - 2\psi_{..}\widehat{\text{cov}}(s_i^2, \psi_{i.})],$$

and where the estimated variance and covariances are obtained from the m sample values of $\psi_{i.}$ and s_i^2. When $n \to \infty$, the variance ratio $\widehat{R}$ should converge to 1 if the sequence converges.

Variations of the original variance ratio by Gelman and Rubin have also been suggested. Gelman [64] used the 'estimated potential scale reduction',

$$\sqrt{\widehat{R}_G} = \sqrt{\frac{\widehat{\sigma}^2}{W}}, \tag{4.3}$$

where $\widehat{R_G} = \widehat{\sigma}^2/W$ is the variance ratio. Brooks and Gelman [21] suggested using the potential scale reduction factor (PSRF)

$$\widehat{R}_C = \frac{(d+3)\widehat{V}}{(d+1)W}. \tag{4.4}$$

They also consider assessing more than one parameter simultaneously. Let ψ denote a vector of parameters. The estimate of the posterior variance-covariance matrix of ψ is

$$\widehat{V} = \frac{n-1}{n}W + \left(\frac{1+m}{m}\right)\frac{B}{n},$$

where

$$W = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{t=n+1}^{2n}(\psi_{it} - \psi_{i.})(\psi_{it} - \psi_{i.})^t$$

and

$$\frac{B}{n} = \frac{1}{m-1}\sum_{i=1}^{m}(\psi_{i.} - \psi_{..})(\psi_{i.} - \psi_{..})^t.$$

Let λ_1 be the largest eigenvalue of the symmetric and positive definite matrix $W^{-1}B/n$. Then define the the multivariate PSRF or MPSRF as

$$\widehat{R}^p = \frac{n-1}{n} + \frac{m+1}{m}\lambda_1.$$

The above methods rely on the assumption of normality for ψ. Brooks and Gelman [21] suggested two empirical methods to avoid the assumption. Firstly, an interval-based index $\widehat{R}_{interval}$ is introduced. From each individual chain, take the empirical $100(1-\alpha)\%$ interval, i.e. the $100\frac{\alpha}{2}\%$ and the $100(1-\frac{\alpha}{2})\%$ points of the n simulation draws as the within-sequence interval length estimates. Then calculate the empirical $100(1-\alpha)\%$ interval of total samples, to gain a total-sequence interval length estimate. The interval-based index is defined as

$$\widehat{R}_{interval} = \frac{\text{length of total-sequence interval}}{\text{mean length of the within-sequence intervals}}. \tag{4.5}$$

The other empirical method suggested by Brooks and Gelman [21] makes use of the empirical estimate of the central s^{th} ordered moments. The index is defined as

$$\widehat{R}_s = \frac{\frac{1}{mn-1}\sum_{j=1}^{m}\sum_{t=n+1}^{2n}|\psi_{jt}-\psi_{..}|^s}{\frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=n+1}^{2n}|\psi_{jt}-\psi_{j.}|^s},$$

for any s.

4.6 Sampling Plan

Due to the dimension and the complexity of the posterior distribution, our sampling scheme is based on the Gibbs sampler. In section 4.4, we introduced the ARMS algorithm, which is more efficient for generating a sample from a distribution with an arbitrary shape than generating a sample using the ordinary rejection sampler. There is no need to choose proper proposal distributions or envelope functions because the ARMS automatically builds them after assigning the initial abscissae set. Hybridisation of ARMS and Gibbs sampler can be easily applied to any model. The other methods (see section 4.7.1 for improving sampling efficiency) are much more difficult to set up and there is even a doubt whether they can really increase the efficiency for high dimensional cases.

The ARMS has in fact been implemented in the latest version of BUGS (MRC Biostatistics Unit, Cambridge, version 0.6) and WinBUGS package (MRC Biostatistics Unit, Cambridge ,version 1.3) [117] [116] as a strategy to improve the

sampling efficiency. 'BUGS' stands for **B***ayesian Inference* **U***sing* **G***ibbs* **S***ampling*, which is a computer software for the Bayesian analysis of complex statistical models using MCMC methods. The classical BUGS is a command-line language, while WinBUGS uses a graphical interface. One of the restrictions of the latest BUGS and WinBUGs package is that it is still not possible to place any structure on a covariance matrix given an inverse Wishart distribution in the packages. As a result, we cannot use these packages for our hierarchical model. Our MCMC sampler is implemented in MATLAB (The MathWorks, Inc. Version 5.3, 1999), which is a high-performance language for technical computing, especially matrix manipulation.

Multiple-chain MCMC sampling is also suggested for use here because it provides an easy way to assess convergence. It also provides a more reliable detection of whether the joint distribution of several parameters is multimodal, while a diagnostic method using single-chain MCMC may only converge to a unimodal distribution when there are several modes. Besides, when the correlation between parameters is high, it is possible for a single-chain MCMC to have a time series plot which looks like a multimodal one because the sequence may circle around a small area then move away. The multiple-chain MCMC approach can also help to judge whether this is a multimodal case or not. If it is only due to correlation, different chains are less likely to circle around at the same area.

All parameters are divided into several components. The full conditional density function of each component is required for generating samples. Generally each component is univariate, but not necessarily. For example, when several parameters' joint conditional distribution is a standard multivariate or matrix-variate distribution, they can be generated as a whole. The uniform and normal random generators which are required in our examples are built-in functions in MATLAB [122]. Random numbers from the inverse-gamma distribution can be obtained by transforming random numbers generated from a gamma distribution (see section 2.5.1). For the random gamma generator we use Best's rejection algorithm (see Devroye [49]). Random numbers from all the other distributions

are generated using the ARMS sampler.

Due to the complexity of the joint posterior models, some investigation of the posterior distribution needs to be done before running the main MCMC simulation, in order to locate the high density area and select the appropriate initial abscissae set for a model. When the dimension for the model is very low, one can plot the joint distribution or the conditional distributions in order to get an idea of the shape and the location of the high density area in the parameter space. When there are a lot of parameters, it is simply impossible to get any idea of the joint distribution by plotting the conditional distributions. In order to make efficient simulation, the initial abscissae should span the high density area of the distribution of the parameter that we want to generate from so that the initial rejection function has a shape that captures that general shape of the density function. If the initial rejection function is too flat, more rejections will happen in each iteration.

To select good initial abscissae sets for all parameters, we start a short run of MCMC simulation with wider spread abscissae sets. From the histograms of generated samples we find a rough location for the high-density area. Then we narrow down the range of our abscissae sets and start another run of MCMC. If the abscissae sets are still too widely spread, we narrow the range of the abscissae sets again, until the abscissae sets only spread over the high density area. The main simulation can then start.

4.7 Other Approaches

4.7.1 Improving efficiency

In this chapter we have focused on ARMS as a method for improving sampling efficiency. The sampling strategy we employ in this thesis for analysing our examples is to use multiple chains of ARMS within Gibbs sampling. The dimension of the problems in our examples, especially the hierarchical regression analysis is very high so that it is very difficult to construct an efficient MCMC sampler with

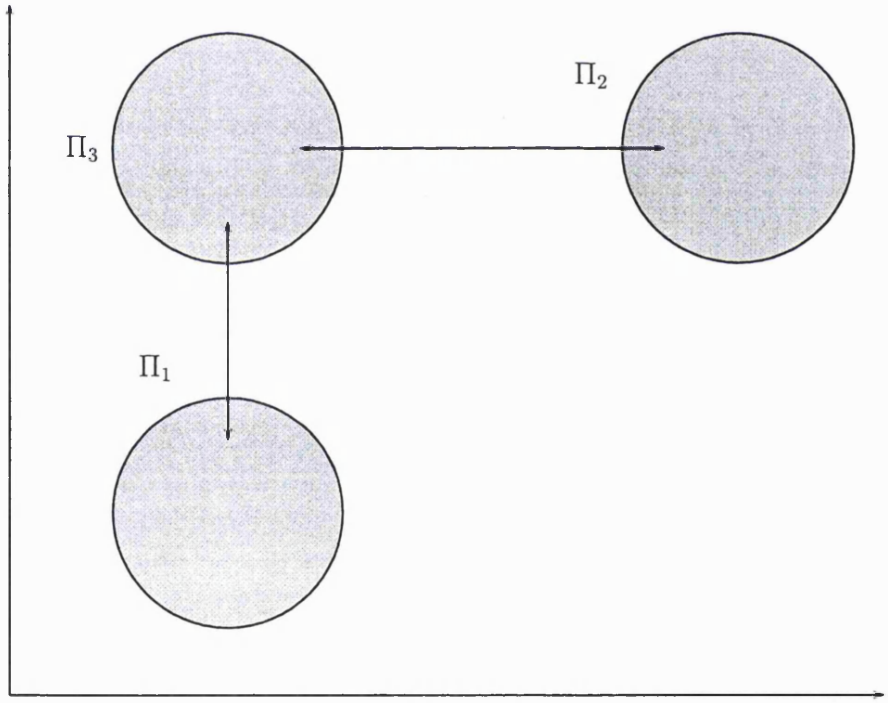
other strategies.

There are many strategies in the literature for improving the convergence speed. Choosing a good proposal density is just the first step. When the correlation between parameters of interest is high, the sampling procedure takes a long time to converge. Two ways of improving this situation are reparameterisation and the blocking of parameters. By proper reparameterisation, one may obtain new parameters with lower correlation. The best-known example is the reparameterisation of the regression coefficient of the linear regression model. More examples are given in Gilks *et al.* [73]. Sampling highly correlated parameters in a block may also remove the effect of correlation (see Gilks *et al.* [73]). However, there is a serious danger of expending a large amount of computing time in each iteration, since it is generally more difficult to sample a vector than to sample a scalar [110]. Random direction methods, such as the hit-and-run algorithm by Schmeiser and Chen [112] or the adaptive direction sampling by Gilks *et al.* [74] are also strategies for sampling highly correlated parameters. In Gibbs sampling, samples always move along the tracks parallel to the coordinate axes. This property causes slow convergence when Gibbs sampling is used for highly correlated parameters. Random direction methods allow each move to happen in any possible direction. For multimodal target density functions, the random direction methods may also work well.

Methods for improving sampling efficiency based on importance sampling [65] have been designed. Samples are drawn from a new target density function, which is a modification of the original target density function, then the importance sampling is applied to estimate the expectation of the random quantities we want to learn about. One useful modification for Gibbs sampling is to create another peak in the original target density function as a stepping stone so that the sample can jump to another peak more easily via the new peak[73]. See figure 4.4. Another frequently used modification is to flatten the original target density function f in order to allow a chain to travel from one peak to another peak more easily. The most common method is to take $f^* = f^{1/T}$ as the new (un-normalised) target

Figure 4.4: Adding a new peak

Adding a new peak Π_3 in the original target distribution with two separated peaks Π_1 and Π_2: it is much easier for a Gibbs sampling sequence to travel between Π_1 and Π_3 or travel between Π_2 and Π_3 rather than travel between Π_1 and Π_2 straight away.



density function in MCMC, where T is called temperature.

Geyer [70] proposed running m parallel MCMC chains with different target density functions $\{f_1^*, f_2^*, \cdots\}$, where $f_1^* = f$. For example, the f_i^*'s may be modified f's with different temperatures. Sampling starts from the chain of f_1^*. A sample is taken from the current chain at the current iteration, then an attempt is made to move to another chain using a Metropolis-Hastings step to generate the next sample. At the end, only samples from the chain of f_1^* are kept as the final samples. The method is called Metropolis-Coupled MCMC. A similar idea is simulated tempering or simulated annealing [71], which runs only one chain instead of m chains, but within the chain, the target density function switches, according to a Metropolis-Hastings step for the index of f_i^*. One problem in simulated tempering is that it requires the normalised constant of each target density function, which is rarely known analytically in Bayesian modelling. Therefore, the normalised constants have to be estimated.

Besag and Green [14] suggested that using auxiliary variables may increase the sampling efficiency in some cases. One example of using auxiliary variables is ARMS. This method adds extra variables into the model without affecting the target density function but the simulation is easier and more efficient with auxiliary variables than without them. More examples of using auxiliary variables are described in [73]. There are many more techniques for improving sampling efficiency, but it is impossible to describe every method in limited space.

4.7.2 Convergence Assessment

One important practical issue in the area of MCMC is how many iterations we need to get a good approximation for our posterior model. The MCMC estimation is based on the assumption that the samples generated after some burn-in period converge to the target distribution and can approximate the stationary distribution of the Markov chains well. If the convergence rate of an Markov sequence can be calculated analytically or approximately, one may know how many iterations are sufficient to reach the stationary distribution. However, convergence rates are

extremely difficult to calculate or estimate in practice, and it is even impossible to prove such rates exist [22]. Therefore, methods for convergence diagnostics that do not require knowing the convergence rate are very important in MCMC applications. Many techniques have been developed in order to assess whether MCMC chains have converged after a given number of iterations. However, these methods cannot guarantee the convergence of the Markov chains.

The variance-ratio approach introduced in section 4.5.2 is one example, which requires running several independent chains. Yu and Mykland's cusum method [130] is another method but is entirely different from the variance-ratio approach. The cusum method is a graphical based method that requires only a single Markov sequence. One judges a sequence by monitoring the cusum plot of the sequence. A "hairy" cusum plot indicates that the sequence converges well. Such judgement is subjective. Geweke [69] proposed another single sequence method which yields rather more objective judgement. It compares two sub-sequences in a MCMC sequence with a test statistic. If the testing is rejected, we conclude that the sequence has not converged yet.

Those methods introduced in the previous paragraph require only the output of the simulation to make convergence diagnostics. There are some methods that require more information from the Markov chains. For example, Liu, Liu and Rubin's L^2 [92] convergence diagnostic method further requires the transition kernel of the sampler. These methods are much more computationally expensive. Another property of Liu, Liu and Rubin's method is that it is only designed for the Gibbs sampler. Another example for the methods designed only for certain sampler is Mykland's *et al.* [98] approach, which needs the regenerative simulation. The methods introduced in the previous paragraph are some examples of the methods that apply to general MCMC samplers.

There have been so many methods in the literature that we do not intend to introduce them in details. Comprehensive reviews of many methods can be found in Brooks and Roberts [22], Cowles and Carlin [36] and Mengersen *et al.* [96]. MCMC sequences approved by these methods are not guaranteed that they ac-

tually converge to the right target distribution after the burn in period. Several methods can be used together in order to make the decision more correctly. Green and Murdoch [77] suggests that exact sampling should be the ultimate objective of Bayesian computation. However, using non-exact simulation is still a more realistic approach currently. As a result, convergence diagnostics is still an important topic in MCMC simulation.

Chapter 5

Modelling a High Dimensional Covariance Matrix

5.1 Introduction

Suppose the random vector $X = (X_1, X_2, \ldots, X_p)$ follows a multivariate normal distribution with mean zero and unknown covariance matrix Σ. A prior distribution is assigned to Σ under a Bayesian framework. When p is large and the number of observations is small, the prior distribution for the covariance matrix has a great effect on the posterior model. Since there is not much information in the data, correct prior information is desired. Very often, prior knowledge for Σ is either limited or difficult to formalise. In a conjugate analysis, we suppose $\Sigma \sim \mathcal{IW}(\delta; \Phi)$, where $\delta > 0$ and $\Phi > 0$ are given. However, we may not know what are the proper values for the hyperparameters δ and Φ, or these hyperparameters should not be fixed constants at all. As a result, we assign diffuse priors to the hyperparameters to indicate our prior ignorance. When p is small, one may consider every entry in the upper triangle of Φ as an individual hyperparameter and assign to each hyperparameter a prior distribution. However, when p is large, the number of hyperparameters is then so large that Bayesian modelling is very complicated. Instead, we consider the case when Φ has a structure with few hyperparameters. Since $(\delta - 2)E(\Sigma) = \Phi$, the structure of Φ should be consistent with our belief of

the structure of Σ. Structuring Φ instead of Σ implies the belief that the structure for Σ is not deterministic and the posterior covariance structure will be adjusted by the data.

Properties such as homoscedasticity, independence and exchangeability for variables can be assumed so that the number of parameters can be greatly reduced. These assumptions sometimes cohere with our belief in a model, while in some cases they are made to reduce the computational complexity. Homoscedasticity, independence and exchangeability represent different structural properties of Σ. When variables are homoscedastic, the variances of all the variables are the same; when variables are independent, Σ is a diagonal matrix; when variables are exchangeable, variances of all variables are the same and all the entries off the diagonal in Σ are the same. Under these properties, a model may be specified using the least possible number of parameters. We consider cases with slightly more complicated structures for the covariance matrix of the variables. We follow Brown's [23] suggestion of using a coherent structural covariance matrix.

Consider the model for the NIR applications in our examples. The NIR spectra of wheat samples are smooth random functions. The real process that generates NIR spectra of wheat samples is unknown, but empirically the measurements at different wavelengths are highly correlated, and the covariance function of the process should be continuous. Practically, setting $\Sigma = kI$ in a prior model leads to a posterior model which predicts reasonable well. Although ARMA-type correlation structures have been suggested, only an AR(1)-type structure has been used in practice [24]. We further apply an AR(2)-type correlation structure for which the correlation decays more slowly than an AR(1) process.

In this chapter, we first introduce the definition of a random function that we use to describe the NIR spectra. Then we build up the model step by step. Structural coherence will be briefly introduced. Some algebraic properties of the AR(1) and the AR(2) autocorrelation functions are presented. This model framework will be used for the NIR spectra throughout this thesis. By assigning strong hyper priors to the hyperparameters, we simulate some spectra from the models

with different covariance structures. We then compare the graphs of the spectra generated by the models, their sample covariance matrices and sample correlation matrices with those of the natural NIR spectra.

5.2 Random Function

Since the NIR absorption or reflection of samples can be detected at any wavelength within the NIR band, it would be appropriate to describe an NIR spectrum as a continuous-time random function. However, an instrument can only record NIR absorption or reflection on a discrete set of wavelengths. A discrete NIR spectrum recorded by an instrument may be thought of as a sub-sequence of a continuous-time random function.

For any arbitrary set $T \subset \mathbb{R}$, ξ is called a random function on T if $\xi = \{\xi(t)|\forall t \in T, \xi(t) \text{ is a random variable}\}$. Suppose S is a countable subspace of T, where $S = \{t_1, t_2, \ldots, t_n\}$, and n can be infinity, then $\xi_S = \{\xi(t_1), \xi(t_2), \ldots, \xi(t_n)\}$ is a random function on S, and is a subset of ξ. Define the distribution function for the random function as

$$F_{t_1,t_2,\ldots,t_n}(x_1, x_2, \ldots, x_n) = P(\xi(t_1) < x_1, \xi(t_2) < x_2, \ldots, \xi(t_n) < x_n), \qquad (5.1)$$

which is the joint distribution function of $\xi(t_1), \xi(t_2), \ldots, \xi(t_n)$. The distribution function of the random function satisfies two conditions (Yaglom [128]): firstly, the distribution function is the same under any permutation for the indices of t, and secondly,

$$F_{t_1,t_2,\ldots,t_m,t_{m+1},\ldots,t_n}(x_1, x_2, \ldots, x_m, \infty, \ldots, \infty) = F_{t_1,t_2,\ldots,t_m,}(x_1, x_2, \ldots, x_m)$$

for any $t_{m+1}, \ldots, t_n$ if $m < n$.

The random function ξ_S on S is *equally spaced* if

$$S = \{t_j|t_j = t_0 + jh, j \in \mathbb{Z}, n \geq j \geq 1\}.$$

In our examples, NIR spectra are recorded at equally spaced wavelengths. Therefore, we can consider these spectra as equally spaced random functions.

Let $X_{(q)} = (X_1, X_2, \ldots, X_q)$ be a random vector from a sample space $\mathcal{X} \subset \mathbf{R}^q$ with a distribution function $F(X_1, X_2, \ldots, X_q)$. We can re-define X_q as a random function on $T_{(q)}$, where $T_{(q)} = \{t_1, t_2, \ldots, t_q\}$. That is, there exists a one-to-one mapping between $T_{(q)}$ and the indices of $X_{(q)}$.

5.3 Normal Model

We consider a three-level hierarchical model. Let $X_{(q)} = (X_1, X_2, \ldots, X_q)$ be a Gaussian random function. For the first stage, suppose $X_{(q)}$ follows a multivariate normal distribution with covariance matrix Σ $(q \times q)$ and we assume the mean is zero since we would like to focus on the inference for Σ. We require Σ to be strictly positive-definite (denoted as $\Sigma > 0$) so that $X_{(q)}$ always exists on the q-dimensional sample space. A conventional choice for the proper prior distribution for Σ is an inverse-Wishart distribution with a shape parameter $\delta > 0$ (1×1) and scale matrix $\Phi > 0$ $(q \times q)$. This is a conjugate prior distribution for the normally distributed $X_{(q)}$. For the third level, we denote the prior density function for Φ as $\pi(\Phi)$. We denote the model for $X_{(q)}$ as

$$\begin{aligned} X_{(q)} &\sim \mathcal{N}(1, \Sigma), \\ \Sigma &\sim \mathcal{IW}(\delta; \Phi), \end{aligned} \tag{5.2}$$

with prior density function $\pi(\Phi)$ for Φ.

Suppose we observe independent 1 by q samples $x_{1(q)}, x_{2(q)}, \ldots, x_{n(q)}$ for $X_{(q)}$ and let $x = (x_{1(q)}^t, x_{2(q)}^t, \ldots, x_{n(q)}^t)^t$, which is an n by q matrix. Denote the ij^{th} entry of x as x_{ij}. The likelihood function for the model is $L(\Sigma) = p(x|\Sigma)$, where $p(X|\Sigma)$ is the density function of X given Σ, and the prior density function of Σ is $\pi(\Sigma|\Phi)$. The posterior density of Σ conditional on Φ is then

$$\begin{aligned} p(\Sigma|x, \Phi) &= \frac{L(\Sigma)\pi(\Sigma|\Phi)}{\int_\Sigma L(\Sigma, x)\pi(\Sigma|\Phi)d\Sigma} \\ &\propto L(\Sigma)\pi(\Sigma|\Phi), \end{aligned}$$

thus, it can be easily shown that $\Sigma|x, \Phi \sim \mathcal{IW}(\delta + n; x^t x + \Phi)$. This density function only exists on the q-dimensional space when $x^t x + \Phi$ is positive definite.

The joint posterior density of Σ and Φ is

$$\begin{aligned} p(\Sigma,\Phi|x) &= \frac{L(\Sigma)\pi(\Sigma|\Phi)\pi(\Phi)}{\int_{\Sigma,\Phi} L(\Sigma)\pi(\Sigma|\Phi)\pi(\Phi)d\Sigma d\Phi} \\ &\propto L(\Sigma)\pi(\Sigma|\Phi)\pi(\Phi) \\ &\propto |\Phi|^{\frac{\delta+q-1}{2}}|\Sigma|^{-\frac{\delta+2q+n}{2}}\exp[-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(x^tx+\Phi)]\pi(\Phi), \end{aligned} \tag{5.3}$$

and the marginal posterior densities of Σ and Φ are

$$\begin{aligned} p(\Sigma|x) &= \int_\Phi p(\Sigma,\Phi|x)d\Phi \\ &\propto \int_\Phi |\Phi|^{\frac{\delta+q-1}{2}}|\Sigma|^{-\frac{\delta+2q+n}{2}}\exp[-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(x^tx+\Phi)]\pi(\Phi)d\Phi, \end{aligned} \tag{5.4}$$

$$\begin{aligned} p(\Phi|x) &= \int_\Sigma p(\Sigma,\Phi|x)d\Sigma \\ &\propto \int_\Sigma |\Phi|^{\frac{\delta+q-1}{2}}|\Sigma|^{-\frac{\delta+2q+n}{2}}\exp[-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(x^tx+\Phi)]\pi(\Phi)d\Sigma \\ &\propto \frac{|\Phi|^{\frac{\delta+q-1}{2}}\pi(\Phi)}{|x^tx+\Phi|^{\frac{\delta+q+n-1}{2}}} \end{aligned} \tag{5.5}$$

respectively.

The posterior distributions of Σ and Φ and the predictive distribution of the future observation will usually be very complicated under the above model assumption given arbitrary $\pi(\Phi)$. When Φ has an improper prior density $\pi(\Phi) \propto |\Phi|^{-k/2}$, the marginal distribution of Σ is $\mathcal{IW}(n+k-2q; x^tx)$, the marginal posterior density of Φ is a matrix-F distribution $\mathcal{F}(\delta+2q-k, n-2q+k; x^tx)$, and the predictive distribution for a future value X_f is $\mathcal{T}(k+n-2q; 1, x^tx)$. These distribtions are well-defined only if $2q-n<k<\delta+2q$ and $x^tx>0$ is non-singular. When x^tx is singular, the marginal posterior density of Φ does not exist on the q-dimensional space but on a hyperplane in the q-dimensional space. Therefore, other forms of prior for Φ have to be used when x^tx is singular. A simpler case is when there is a simple structure in Φ (when Φ is a matrix function of a small number of parameters).

5.4 Coherence

We consider two principles of coherence in Brown [23]. Firstly, suppose we have q variables in our model. The number of variables q can be altered by either adding some variables or taking away some variables. The prior distribution in the model with fewer variables should be the marginalised prior distribution of the prior distribution of the model with more variables. The assumption of multivariate normal sampling distribution and inverse-Wishart prior distribution for the covariance matrix automatically satisfies this requirement (see also Lindley [90]). For example, suppose $Y_{(q)} \sim N(0, \Sigma_{(q)})$ and $\Sigma_{(q)} \sim \mathcal{IW}(\delta; \Phi_{(q)})$, and $Y_{(p)}$ is a refinement (sub-vector) of $Y_{(q)}$. Then, the covariance matrix $\Sigma_{(p)}$ of $Y_{(p)}$ would be the sub-matrix of $\Sigma_{(q)}$ which corresponds to $Y_{(q)}$, and the prior distribution of $\Sigma_{(p)}$ would be $\mathcal{IW}(\delta; \Phi_{(p)})$, where the indices of $\Phi_{(p)}$ in $\Phi_{(q)}$ are the same as the indices of $\Sigma_{(p)}$ in $\Sigma_{(q)}$.

Secondly, the prior distribution for the smaller random vector should be structurally generated by the same prior consideration that leads to the generation of the prior of the parental vector of the smaller random vector. This principle is called structural coherence. Suppose q is odd and $Y_{(q)} = (Y_1, Y_2, \ldots, Y_q)$ is an AR(1) process, which is an equally spaced random function. Then, the refinement $Y_{[q]} = (Y_1, Y_3, \ldots, Y_q)$ of $Y_{(q)}$ is also an AR(1) process, which is also equally spaced. Any equally spaced refinement of $Y_{(q)}$ is an AR(1) process. However, suppose $Y_{(q)}$ is a MA(1) process, then $Y_{[q]}$ will not be an MA(1) process. Any equally spaced refinement of $Y_{(q)}$ will not be an MA(1) process. Therefore, an AR(1) process is structurally coherent while an MA(1) process is not.

In our examples, the NIR spectrum of a sample consists of 100 absorptions measured at 100 equally spaced wavelengths. One may also obtain an NIR spectrum for the sample with another set of wavelengths. Since these spectra are in fact sub-sequences of the same continuous-time random function with a continuous covariance function, in assuming the prior model of the NIR spectra, we should consider a structurally coherent prior belief for the model.

5.5 Structural Covariance

Consider the model in section 5.3. Let $\Phi = \Lambda P \Lambda$, where $P > 0$, Λ is a diagonal matrix, every entry on the diagonal is greater than zero, and the diagonal of $\Lambda\Lambda$ is the same as the diagonal of Φ. Thus, the diagonal of P is a vector with all elements one. If $\Sigma \sim \mathcal{IW}(\delta; \Phi)$,

$$E(\Sigma) = \frac{\Lambda}{(\delta-2)^{\frac{1}{2}}} P \frac{\Lambda}{(\delta-2)^{\frac{1}{2}}},$$

so that P describes our belief about the correlation matrix, and the diagonal of $\Lambda/(\delta-2)^{0.5}$ represents our belief about the standard deviation of $X_{(q)}$. To structure Φ, let Φ be a matrix function of vector κ, whose number of elements is much smaller than $q(q+1)/2$. Assume $\Phi = \Phi(\kappa)$, $P = P(\kappa)$ and $\Lambda = \Lambda(\kappa)$, although P and Λ may not have common parameters. We focus on the assumption for P and assume a simple structure for Λ.

The simplest assumption for the expected covariance of $X_{(q)}$ is to assume $\Phi = \sigma^2 I_q$, where P is a q by q identity matrix, and Λ is a diagonal matrix with all elements on its diagonal equal to σ. It uses only one parameter σ. The implication of the assumption is that $X_1, X_2, \ldots X_q$ are expected to be mutually independent and homoscedastic. An intraclass form is also a simple structure, having

$$\Phi = \sigma^2(1-\gamma)I_q + \sigma^2\gamma J_q,$$

where J_q is a q by q matrix with all entries one, and $0 < \gamma < 1$ so that $\Lambda = \sigma I_q$ and P is a matrix with all diagonal elements equal to 1 and other entries γ. This structure corresponds to the assumption that the explanatory variables are expected to be exchangeable. Another class of structures for Q_{xx} is suggested in Brown [23]. It considers that $(X_1, X_2, \ldots X_q)$ are equally spaced sub-sampled from a weak stationary continuous parameter Gaussian process. One member of the sub-class in Brown's suggestion involves the correlation function

$$\rho(h) = \exp(-\alpha|h|^k),$$

where $0 < k \leq 2$. When $k = 1$, this implies a continuous AR(1) process. Another structure we will use is the autocorrelation function of an AR(2) process.

These autoregressive-type structures satisfy the structural coherence principle. For continuous-time ARMA(s,t)-type structure, we require $s > t$ in order that the process is stationary. More types of well-defined correlation functions are available in the literature of spatial statistics. Several frequently used correlation functions are introduced in Chilés and Delfiner(1999) [33]. For example, the triangle model, where $\rho(h) = 1 - h/a$ if $h \leq a$ and 0 if $h > a$; the Gaussian model, where $\rho(h) = \exp(-h^2/a^2)$ for $a > 0$; and the general Cauchy model, where $\rho(h) = (1 + h^2/a^2)^{-\alpha}$ for a, $\alpha > 0$.

5.6 AR(1) and AR(2) Correlation Functions

The correlation function of the p^{th} order autoregressive process has a general form:

$$\rho(h) = \sum_{\iota=1}^{p} b_\iota \mu_\iota^h.$$

The advantage of the AR-type correlation functions is that the analytical inverse matrices and determinant of their corresponding correlation matrices can be presented as simple functions. These help improve the accuracy and speed of numerical calculation. In this section, we focus on the correlation functions of AR(1) and AR(2) processes due to their simplicity. Consider the normal model (5.2) with $\Phi = \Lambda P \Lambda$ as in section 5.5 .

The correlation function of the first order autoregressive model is

$$\rho(h) = \tau^h, 1 > \tau \geq 0. \tag{5.6}$$

Therefore, the P matrix in section 5.5 with an AR(1) correlation structure is

$$P = \begin{bmatrix} 1 & \tau & \tau^2 & \dots & \tau^{q-1} \\ \tau & 1 & \tau & \dots & \tau^{q-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \tau^{q-1} & \tau^{q-2} & \dots & \tau & 1 \end{bmatrix},$$

when τ is zero, P is an identity matrix, which means the X_i's are independent.

The inverse of P is

$$P^{-1} = \frac{1}{1-\tau^2}\begin{bmatrix} 1 & -\tau & 0 & \dots & \dots & 0 \\ -\tau & 1+\tau^2 & -\tau & 0 & \dots & 0 \\ 0 & -\tau & 1+\tau^2 & -\tau & 0 & \dots \\ \vdots & \vdots & \dots & \dots & \vdots & \vdots \\ 0 & \dots & 0 & -\tau & 1+\tau^2 & -\tau \\ 0 & 0 & \dots & 0 & -\tau & 1 \end{bmatrix}.$$

The determinant of P is

$$|P| = (1-\tau^2)^{q-1}.$$

The correlation function of the 2nd order autoregressive process is

$$\rho(h) = b_1\mu_1^h + b_2\mu_2^h,$$

where

$$\begin{aligned} b_1 &= \frac{(1-\mu_2^2)\mu_1}{(\mu_1-\mu_2)(1+\mu_1\mu_2)}, \\ b_2 &= -\frac{(1-\mu_1^2)\mu_2}{(\mu_1-\mu_2)(1+\mu_1\mu_2)}. \end{aligned}$$

P with AR(2) autocorrelation structure is

$$\begin{bmatrix} 1 & b_1\mu_1+b_2\mu_2 & b_1\mu_1^2+b_2\mu_2^2 & \dots & b_1\mu_1^{q-1}+b_2\mu_2^{q-1} \\ b_1\mu_1+b_2\mu_2 & 1 & b_1\mu_1+b_2\mu_2 & \dots & b_1\mu_1^{q-2}+b_2\mu_2^{q-2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ b_1\mu_1^{q-1}+b_2\mu_2^{q-1} & b_1\mu_1^{q-2}+b_2\mu_2^{q-2} & \dots & \dots & 1 \end{bmatrix}. \quad (5.7)$$

For the convenience of manipulation, we reparameterize μ_1 and μ_2 as

$$\begin{aligned} \phi_1 &= \mu_1+\mu_2, \\ \phi_2 &= -\mu_1\mu_2. \end{aligned}$$

Under the asymptotic stationarity condition for an AR(2) process (see Box *et al.* [17]), we must have

$$-1 < \phi_2 < 1, \quad \phi_2+\phi_1 < 1, \quad \phi_2-\phi_1 < 1, \quad (5.8)$$

and then (5.7) is non-singular for arbitrary q. In order for μ_1 and μ_2 to be real, we need $\phi_1^2 + 4\phi_2 \geq 0$ (see Box *et al.* [17]). For complex solution for μ_1 and μ_2, we require $\phi_1^2 + 4\phi_2 < 0$.

The two parameters μ_1 and μ_2 are

$$\mu_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2}, \quad \mu_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2}.$$

We may rewrite μ_1 and μ_2 in the complex form

$$\mu_1 = D\exp(i2\pi f_0), \quad \mu_2 = D\exp(-i2\pi f_0), \tag{5.9}$$

where $D = \sqrt{-\phi_2}$. In the case with complex μ_1 and μ_2, the AR(2) autocorrection function is a damped sine wave with frequency f_0 and dampling factor D. (Box *et al.* [17]). When D is smaller, the autocorrelation function decays to zero faster. When D is greater, the autocorrelation function decays to zero slower.

When $p \geq 5$, $P^{-1} = U^t U$, with U an upper triangular matrix

$$U = \frac{1}{\sqrt{m}} \begin{bmatrix} 1 & -\phi_1 & -\phi_2 & 0 & \dots & \dots & 0 \\ 0 & 1 & -\phi_1 & -\phi_2 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & 0 & 1 & -\phi_1 & -\phi_2 \\ 0 & 0 & \dots & \dots & 0 & k & j \\ 0 & 0 & \dots & \dots & \dots & 0 & l \end{bmatrix},$$

where $k^2 = 1 - \phi_2^2$, $kj = -\phi_1\phi_2 - \phi_1$, $l^2 = 1 - \phi_2^2 - j^2$, and $m = 1 - \phi_1\rho_1 - \phi_2\rho_2$, where $\rho_1 = b_1\mu_1 + b_2\mu_2$ and $\rho_2 = b_1\mu_1^2 + b_2\mu_2^2$.

The inverse of P can be expressed as

$$\frac{1}{m}\begin{bmatrix} 1 & -\phi_1 & -\phi_2 & 0 & \dots & \dots & 0 \\ -\phi_1 & 1+\phi_1^2 & \phi_1\phi_2-\phi_1 & -\phi_2 & 0 & \dots & 0 \\ -\phi_2 & \phi_1\phi_2-\phi_1 & 1+\phi_1^2+\phi_2^2 & \phi_1\phi_2-\phi_1 & -\phi_2 & 0 & \dots \\ 0 & -\phi_2 & \phi_1\phi_2-\phi_1 & 1+\phi_1^2+\phi_2^2 & \dots & \dots & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \dots & 0 & -\phi_2 & \phi_1\phi_2-\phi_1 & 1+\phi_1^2+\phi_2^2 & \phi_1\phi_2-\phi_1 & -\phi_2 \\ 0 & \dots & 0 & -\phi_2 & \phi_1\phi_2-\phi_1 & 1+\phi_1^2 & -\phi_1 \\ 0 & \dots & \dots & 0 & -\phi_2 & -\phi_1 & 1 \end{bmatrix}$$

and

$$|P| = \frac{m^q}{(1+\phi_2)^2[(1-\phi_2)^2-\phi_1^2]}.$$

In practice, NIR spectra are considered to be differentiable. First and second derivatives of the NIR spectra are frequently used. Since the first derivative of $\rho(0)$ for a continuous AR(1) does not exist, it can be shown that a continuous AR(1) process is not stochastically differentiable. For a continuous AR(2) process, the first derivative of the spectra exists, but is nowhere continuous, i.e. the 2^{nd} derivative of a continuous AR(2) process does not exist (see Priestley [104]). However, we do not attempt to pursue a much more appropriate theoretical model for the NIR spectra in this thesis.

5.7 Example

In this section, we display spectra generated from several prior models for our second example in chapter 3. We consider models for original spectra and the 2^{nd} derivative spectra (figure 3.4 and 3.5). We regard the mean of the spectra as known and denote a 1 by q mean-corrected spectrum as $X = (X_1, X_2, \dots X_q)$, which we model as an equally spaced Gaussian random function on wavelengths $T = (t_1, t_2, \dots, t_q)$. Therefore, $X \sim \mathcal{N}(1, \Sigma)$, the q by q covariance matrix $\Sigma \sim \mathcal{IW}(\delta; \Phi)$ and $\Phi_{q\times q} = \Lambda P \Lambda$, where Λ is a q by q diagonal matrix and the diagonal of the q by q matrix P is the same as the diagonal of Φ. The five models in table

Model	Data type	Structure of P
M.a	original spectra	identity matrix
M.b	original spectra	AR(1) correlation matrix
M.c	original spectra	AR(2) correlation matrix
M.d	2nd derivative spectra	identity matrix
M.e	2nd derivative spectra	AR(2) correlation matrix

Table 5.1: Five combinations of data and structural correlation matrix

5.1 are based on whether data are original spectra or the 2^{nd} derivative spectra and on different structures for P. The AR(1) structure has not been considered for the 2^{nd} derivative spectra because the sample correlation matrix of the 2^{nd} derivative spectra in our example (see figure 5.4 (a.3)) is very different from any AR(1) correlation matrix with an autocorrelation function (5.6).

Let $(\lambda_1, \lambda_2, \ldots, \lambda_q)$ be the diagonal of Λ, then $\lambda_\iota^2/(\delta - 2)$ is the prior mean of the variance of X_ι in X. Figure 5.1(a) shows the sample standard deviation of X_ι at wavelength t_ι, and figure 5.1(b) indicates that

$$\log\left[(\text{sample standard deviation of } X_\iota)/\sqrt{|\text{sample mean of } X_\iota|}\right]$$

is approximately a linear function of the index of variables ι. Therefore, we assume

$$\lambda_\iota = \sqrt{|\mu_\iota|}\exp(a_0 + b_0\iota), \tag{5.10}$$

where μ_ι is estimated by the sample mean of X_ι. In order to reduce the correlation between parameters, (5.10) is reparameterised as

$$\lambda_\iota = \sqrt{|\mu_\iota|}\exp[a + b(\iota - \bar{\iota})], \tag{5.11}$$

where $\bar{\iota} = q^{-1}\sum\iota$. The parameter a controls the average level of standard deviation and b controls the average slope of the standard deviation as a function of wavelengths. When b is fixed, the variance increases when a goes up.

Figure 5.2 shows that the curve of the sample standard deviation of the 2^{nd} derivative spectrum has similar pattern as the curve of $\sqrt{|\text{sample mean spectra}|}$. Although these two curves are not exactly the same, their peaks and valleys of the

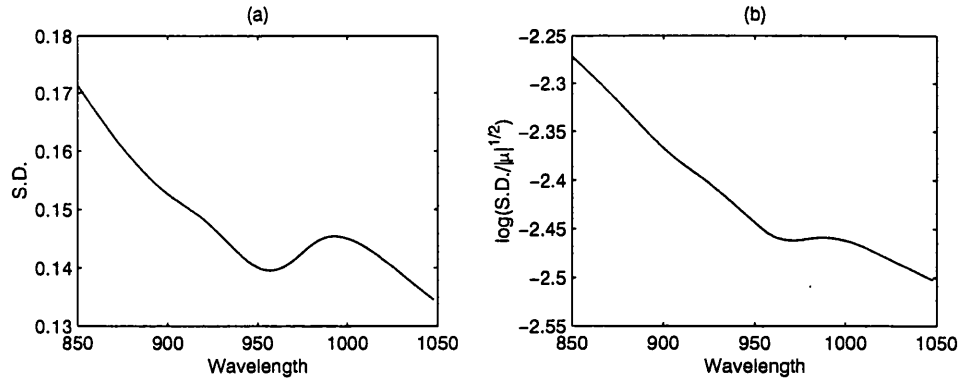


Figure 5.1: Sample standard deviation (S.D.) the NIR spectra in the first example in chapter 3.4.2

curves are almost at the same positions. Therefore, we assume

$$\lambda_\iota = \sqrt{|\mu_\iota|} \exp a, \tag{5.12}$$

for the 2nd derivative spectra. That is, we assume that b is a constant 0 in equation (5.10).

We sample some spectra from the five models in table 5.1 in order to see how similar they are to the real NIR spectra. Due to the hierarchical structure of the models, we can generate them in a simple way. Firstly, a set of hyperparameters is drawn from the hyper prior distributions of the hyperparameters to form a scalar matrix $\widetilde{\Phi}$ for the prior distribution of Σ. Given $\delta = 3$, a covariance matrix $\widetilde{\Sigma}$ is generated from $\mathcal{IW}(3; \widetilde{\Phi})$. A spectrum is then drawn from $\mathcal{N}(1, \widetilde{\Sigma})$. The hyper priors we use will be specified later. According to our preliminary investigation, the models are not sensitive to the shape of hyper priors, but very sensitivity to the domains of the hyperparameters. Therefore, we use uniform distributions as hyper priors for all the hyperparameters. We use MATLAB as a computing tool, which provides random generators for the uniform[0,1] and the standard matrix normal distribution. Random matrices from an inverse-Wishart distribution can be generated easily when δ is an integer and is greater than 2 using the following

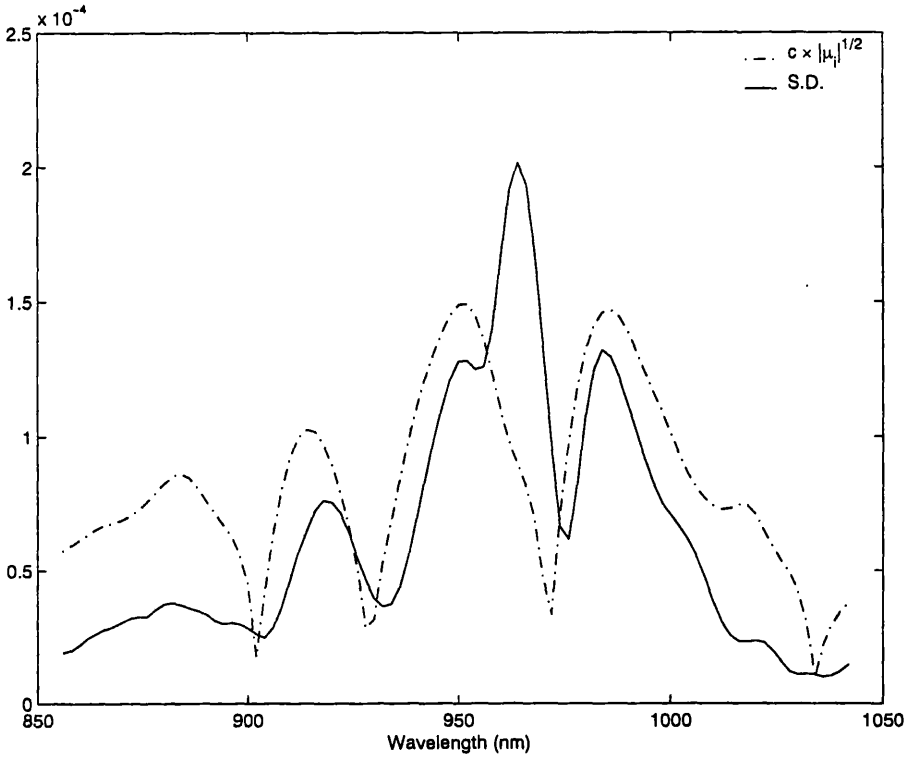


Figure 5.2: Sample standard deviation and a multiple of |sample mean|$^{0.5}$ of the 2nd derivative spectra in the first example in chapter 3.5.2

distribution properties (see section 2.5)

$$
\begin{aligned}
&X \sim \mathcal{N}(I_\nu, I_p) && (\nu \times p),\\
&V^{-1} = X^t X \sim \mathcal{W}(\nu; I_p) && (p \times p),\\
&V \sim \mathcal{IW}(\delta; I_p), && \nu = \delta + p - 1,\\
&\Sigma = AVA^t \sim \mathcal{IW}(\delta; \Phi) && (p \times p),
\end{aligned}
$$

where $\Phi = AA^t$ and A $(p \times p)$ can be a Cholesky factor of Φ.

Since eqn (5.11) and eqn (5.12) are always non-negative, a can possibly be any real number. According to figure 5.1 (b), $\log(\lambda_\iota/\sqrt{\mu_\iota})$ has an obvious downward trend as ι increases. Therefore, b may possibly be any negative real value. For AR(1) structure, τ only need to be a value in $[0, 1)$. For AR(2) structure, ϕ_1 and ϕ_2 have to satisfy the stationarily condition (5.8). For τ, ϕ_1 and ϕ_2, we use uniform priors. We could assign a very diffuse normal prior for a and perhaps a very diffuse gamma prior for $-b$. If we assign the above parameter spaces to our hyper

Model	Hyperparameter	Parameter space and Constraint
M.a	a,b	$a \in [-2.2, -2.6]$ $b \in [-0.00215, -0.00225]$
M.b	a,b,τ	$a \in [-2.2, -2.6]$ $b \in [-0.00215, -0.00225]$ $\tau \in [0.99965, 0.99975]$
M.c	a,b,ϕ_1	$a \in [-2.2, -2.6]$ $b \in [-0.00215, -0.00225]$ $\phi_1 \in [1.983, 1.984]$ $\phi_2 = -\phi_1 + 0.999985$
M.d	a	$a \in [-7, -5]$
M.e	a,b,ϕ_1	$a \in [-7, -5]$ $\phi_1 \in [1.96, 1.98]$ $\phi_2 = -\phi_1^2[4\cos(2\pi/50)^2]^{-1}$

Table 5.2: Prior settings for the example in section 5.7

priors, the spectra generated from these models are widely spread. Therefore, we restrict the ranges of these parameters to smaller parameter spaces so that the generated spectra are close (subjectively) to the real NIR spectra, and use flat priors for them over these ranges. Table 5.2 shows the prior settings for the models we draw spectral samples from. For M.c, a constraint between ϕ_1 and ϕ_2 has been used in order to reduce the number of parameters. In order to get a spectrum with high and slow-decaying autocorrelation, ϕ_1 and ϕ_2 have to be very close to the stationarily condition boundary $\phi_2 = -\phi_1 + 1$. We found that the generated spectra are acceptably close to the natural spectra (figure 5.3) when ϕ_1 is around 1.984 and ϕ_2 is around -0.984015, hence we choose the constraint $\phi_2 = -\phi_1 + 0.999985$. For M.e, we again set up a constraint to eliminate a parameter. According to the sample correlation matrix shown in figure 5.4 (a.3), an AR(2) autocorrelation function with the pattern of a damped sine wave should be able to describe the process, and there should be two periods between wavelength 850nm and 1048nm. Therefore, we assign the frequency f_0 in the complex form of μ_1 and μ_2 to be 50, i.e. we use 100nm (50 variables) as a period. Since the relationship between ϕ_1, ϕ_2 and f_0 is $\phi_2 = -\phi_1^2[4\cos(2\pi/f_0)^2]^{-1}$ (Box *et al.* [17]), we then have a constraint for ϕ_1 and ϕ_2 by fixing f_0. Since the damping factor is $\sqrt{-\phi_2}$, with different values

for ϕ_1 the autocorrelation function of the AR(2) process decays at different rates.

Figure 5.3 (a.1) shows 50 centred NIR spectra from wheat samples. Their smoothness indicates the high correlation between variables. Figure 5.3 (a.3) presents the 3-dimensional visualisation of the sample correlation matrix of the spectra, which looks like a saddle. Figure 5.3 (a.2) and (a.3) show that the covariance and correlation matrix are smooth functions of wavelength as well. The spectra generated by model M.a are shown in figure 5.3 (b.1). The identity structure of P causes each spectrum to be a white noise process. Figure 5.3 (b.2) and(b.3) visualise the sample covariance and correlation matrix for the model.

In comparison with M.a, the shape of sample covariance and correlation matrix of spectra generated by model M.b shown in figure 5.3 (c.2) and (c.3) are closer to the covariance and correlation matrix of the real NIR spectra, although they are not smooth enough. The samples are shown in figure 5.3 (c.1), and they fluctuate a great deal. The spectra generated by the prior model M.c with the AR(2) correlation structure for P are much smoother (see figure 5.3 (d.1)). The smooth correlation matrix of M.c (figure 5.3 (d.3) does not have the shape of a saddle, but its short term autocorrelation is much closer to that of the real NIR spectra than M.a and M.b.

Figure 5.4 (a.1) shows the centred 2^{nd} derivative NIR spectra of the wheat samples. Their sample covariance matrix and sample correlation matrix are shown in figure 5.4 (a.2) and figure 5.4 (a.3). Since the structure of P in model M.d has the same structure as P (identity matrix) in model M.a, we can expect that a spectrum generated from M.d will be a white noise process as well. Graphs for M.d are not shown. Curves generated from M.e are shown in figure 5.4(b.1). The graph shows that the spectra have larger variation in several common sections, which are connected to their neighbour sections with areas with much smaller spectral variation. This special pattern of the variation of the spectra is in fact associated with the function of standard deviation (5.12). The sample covariance matrix and the sample correlation of M.e are shown in figure 5.4(b.2) and (b.3), respectively.

5.8 Remark

In this chapter, we consider AR(1) and AR(2) correlation functions as the matrix functions for Φ in the NIR example. We suppose natural NIR spectra are smooth. The spectra are NIR absorptions at 100 equally spaced wavelengths and the distance between two successive wavelengths is 2nm. Under an appropriate prior belief for the hyperparameters, the prior models with AR(1) and AR(2) assumptions generate smooth spectra. For the AR(1) structure, the spectra are smoother when τ is close to 1. For the AR(2) structure, smooth spectra are produced when the ϕ_1 and ϕ_2 are very close to a boundary ($\phi_2 = -\phi_1 + 1$) of the parameter space for a stationary AR(2) process. Justification for using AR(1) and AR(2) correlation function is based on the macro pattern of the correlation matrix and computational simplicity.

In section 5.7, we assume hyperparameters of the models are from proper uniform distributions. According to our experience, the model is more sensitive to the chosen ranges of the parameters rather than the shape of the prior distributions of hyperparameters over these ranges. This is because when the number of variables is vary large, the marginal prior density function of the hyperparameters is strongly denominated by the prior distribution of Σ and the effect caused by the hyper prior density functions is ignorable unless hyper prior density functions we choose are highly concentrated.

Suppose we observe n samples of X, say x, which is an n by q matrix, each row representing an independent sample. If the rank of $x^t x$ is less than q, then $x^t x$ is not full rank. Suppose $\Phi(\kappa)$ is a q by q positive definite matrix,

$$\text{Rank}(x^t x + \Phi(\kappa)) \leq \text{Rank}(x^t x) + \text{Rank}(\Phi(\kappa)).$$

The posterior density function for κ is

$$\frac{|\Phi(\kappa)|^{\frac{\delta+q}{2}} \pi(\kappa)}{|x^t x + \Phi(\kappa)|^{\frac{\delta+q+n-1}{2}}}.$$

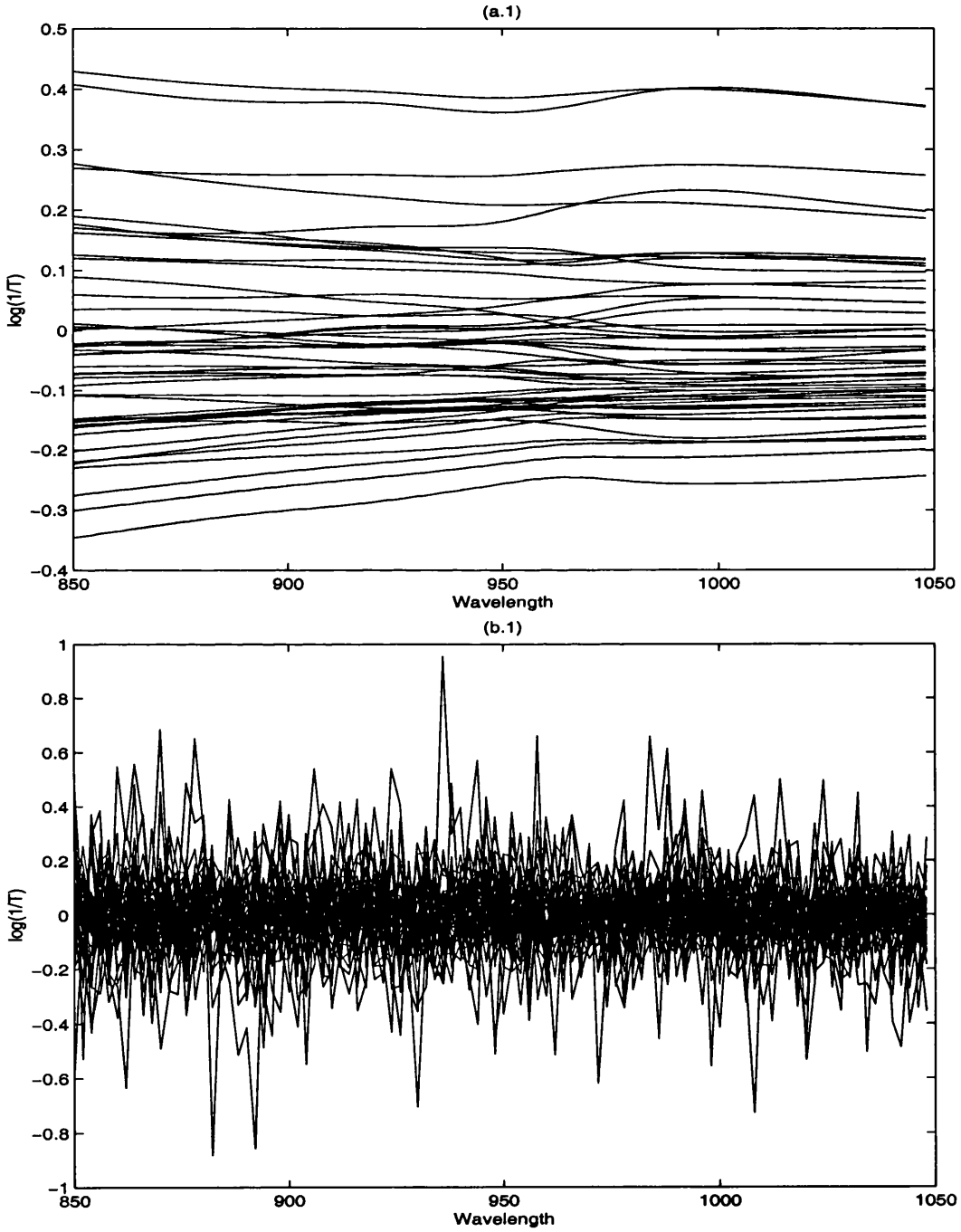
In our models, κ are different combinations of a, b, τ, ϕ_1 and ϕ_2 for different models. For model M.b, Mc, and M.e, our prior settings for the values of τ and ϕ_1

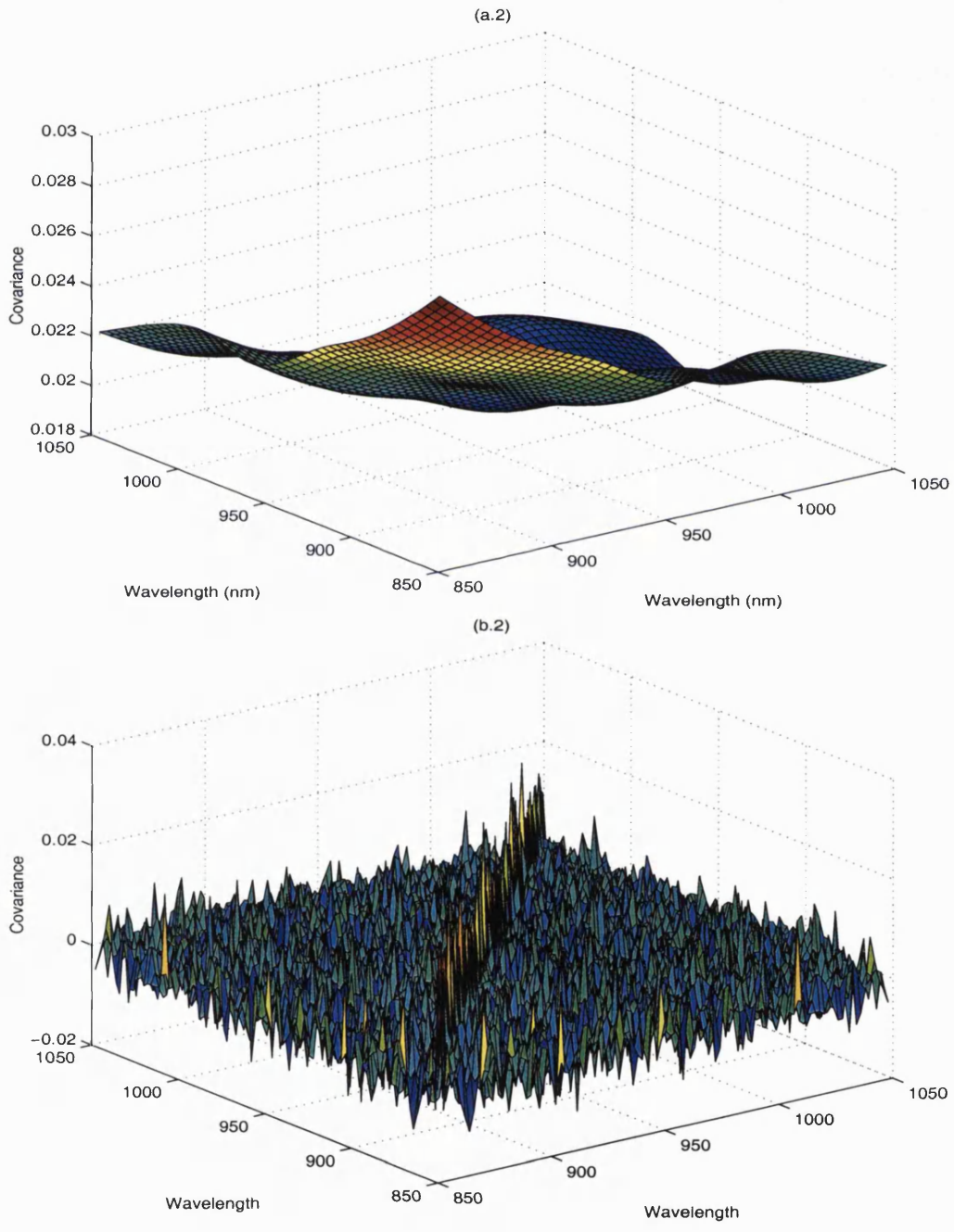
that generate the smooth spectra are very close to the boundaries of the parameter spaces (which for τ is 1 and for ϕ_1 and ϕ_2 is $\phi_2 = -\phi_1 + 1$). When κ reaches these boundaries, the rank of $\Phi(\kappa)$ is dramatically reduced to one so that $x^t x + \Phi(\kappa)$ is no longer non-singular. The determinants of $x^t x + \Phi(\kappa)$ and $\Phi(\kappa)$ converge to zero as κ goes toward these boundaries. By L'Hospital's rule,

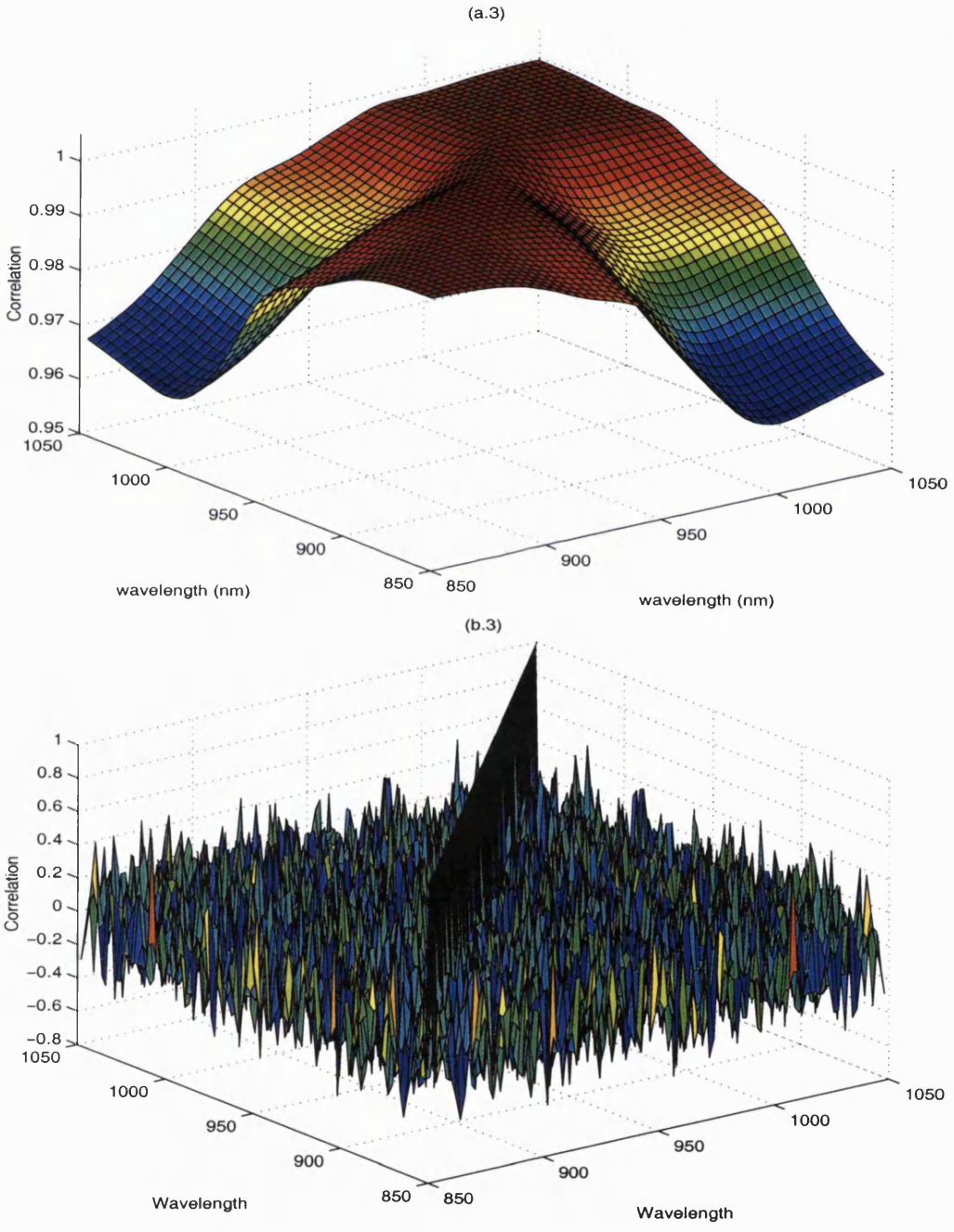
$$\lim_{\kappa \to w} \frac{|\Phi(\kappa)|^{\frac{\delta+q}{2}} \pi(\kappa)}{|x^t x + \Phi(\kappa)|^{\frac{\delta+q+n-1}{2}}} = \infty,$$

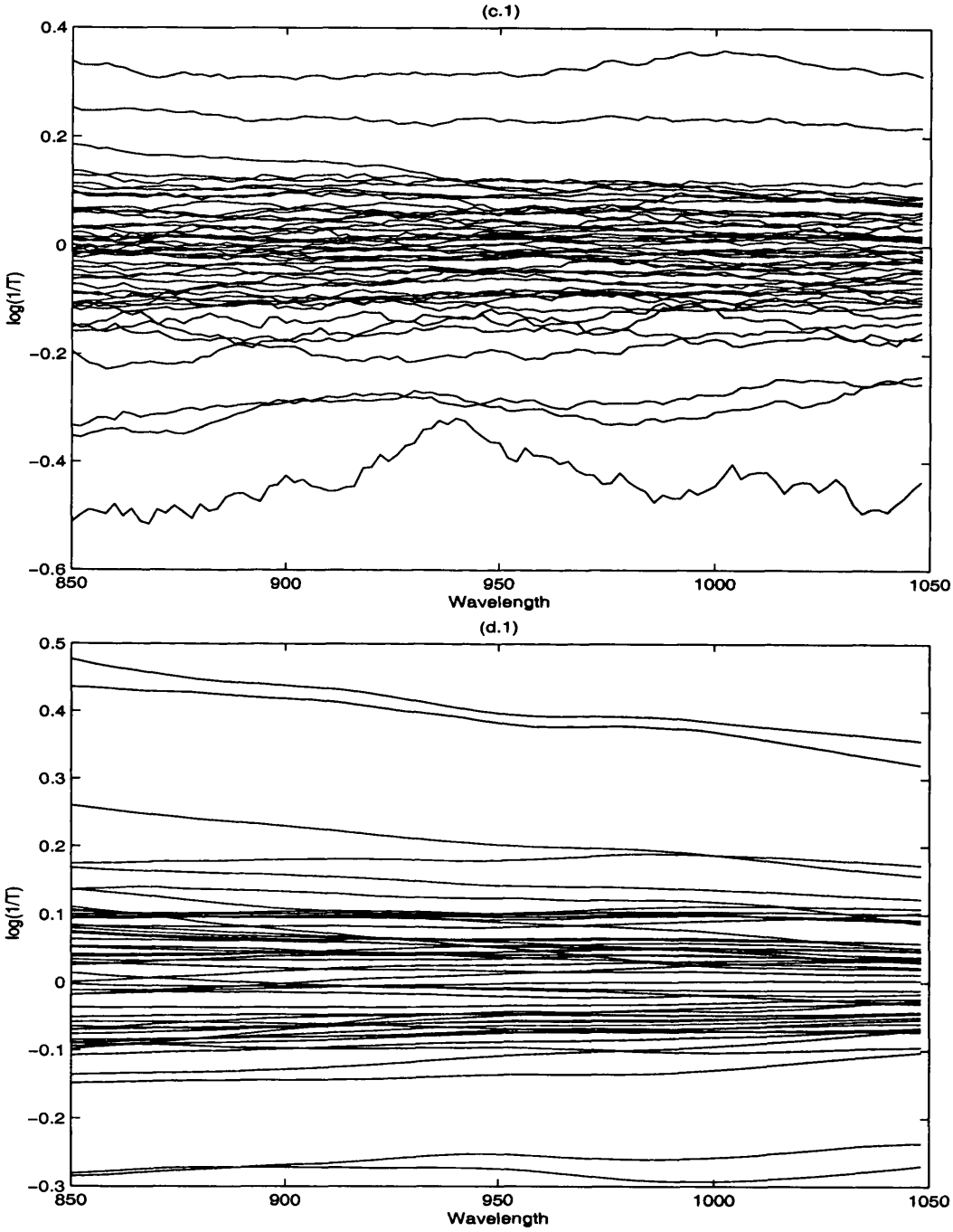
where w represents the boundary points, because the order of $|\Phi(\kappa)|^{\frac{\delta+q}{2}} \pi(\kappa)$ is less than the order of $|x^t x + \Phi(\kappa)|^{\frac{\delta+q+n-1}{2}}$. Therefore, the maximum of the posterior density of κ always happens at the boundary. The information from the data is not enough to form a local maximum value near the boundary. Hence, we have to limit our parameter spaces away from these boundaries or all our MCMC simulations will be absorbed into them. The posterior inference may be sensitive to the chosen space.

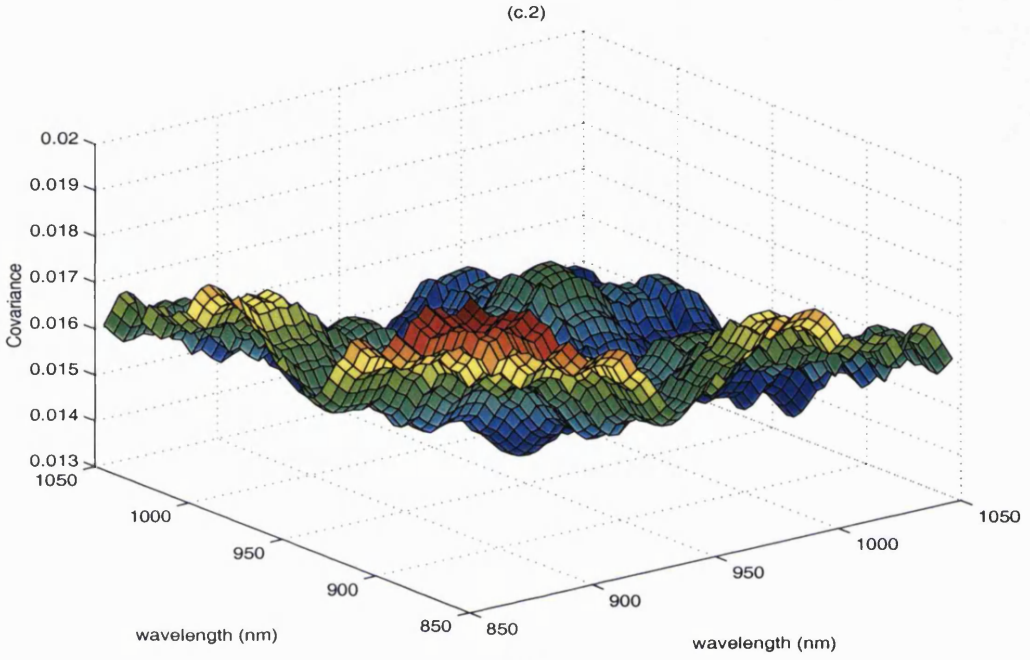
Figure 5.3: Spectra, their covariance matrices and correlation matrices.

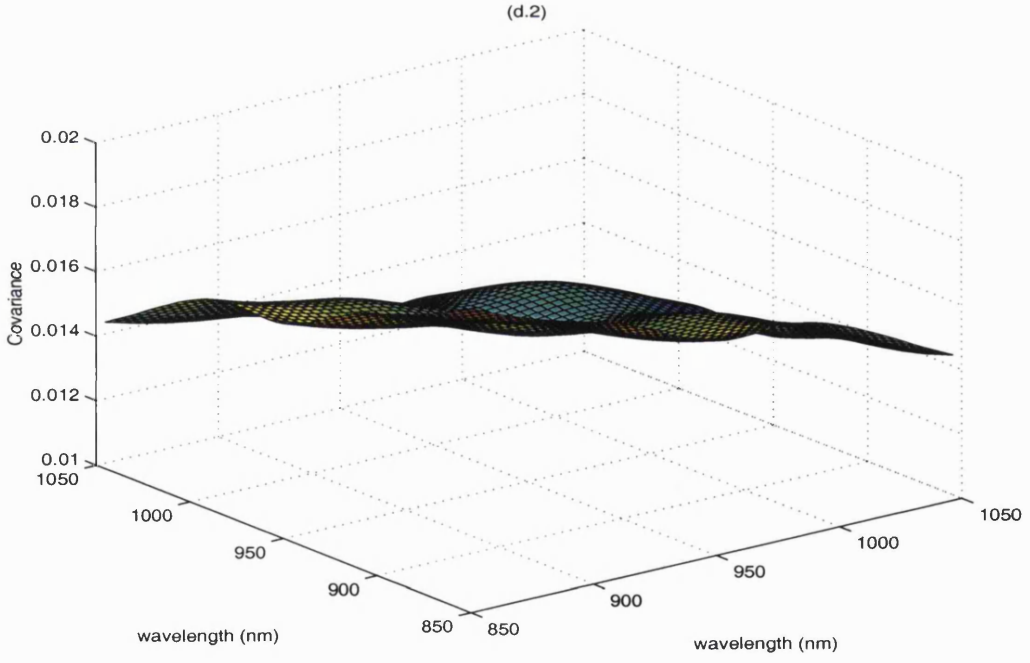


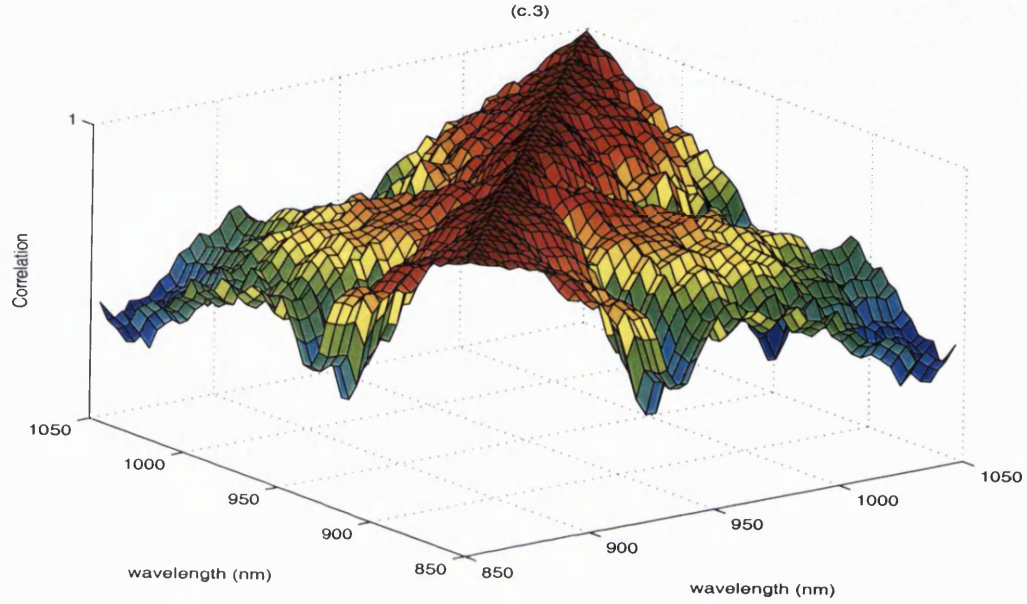

(a.2)
0.03
0.028
0.026
0.024
0.022
0.02
0.018
Covariance
1050
1000
950
900
850
Wavelength (nm)
850
900
950
1000
1050
Wavelength (nm)
(b.2)
0.04
0.02
0
−0.02
Covariance
1050
1000
950
900
850
Wavelength
850
900
950
1000
1050
Wavelength


(a.3)
Correlation
1
0.99
0.98
0.97
0.96
0.95
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)
(b.3)
Correlation
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
1050
1000
950
900
850
Wavelength
850
900
950
1000
1050
Wavelength


(c.1)
0.4
0.2
0
-0.2
-0.4
-0.6
log(1/T)
850
900
950
1000
1050
Wavelength
(d.1)
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
log(1/T)
850
900
950
1000
1050
Wavelength


(c.2)
Covariance
0.02
0.019
0.018
0.017
0.016
0.015
0.014
0.013
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)


(d.2)
Covariance
0.02
0.018
0.016
0.014
0.012
0.01
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)


(c.3)
Correlation
1
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)

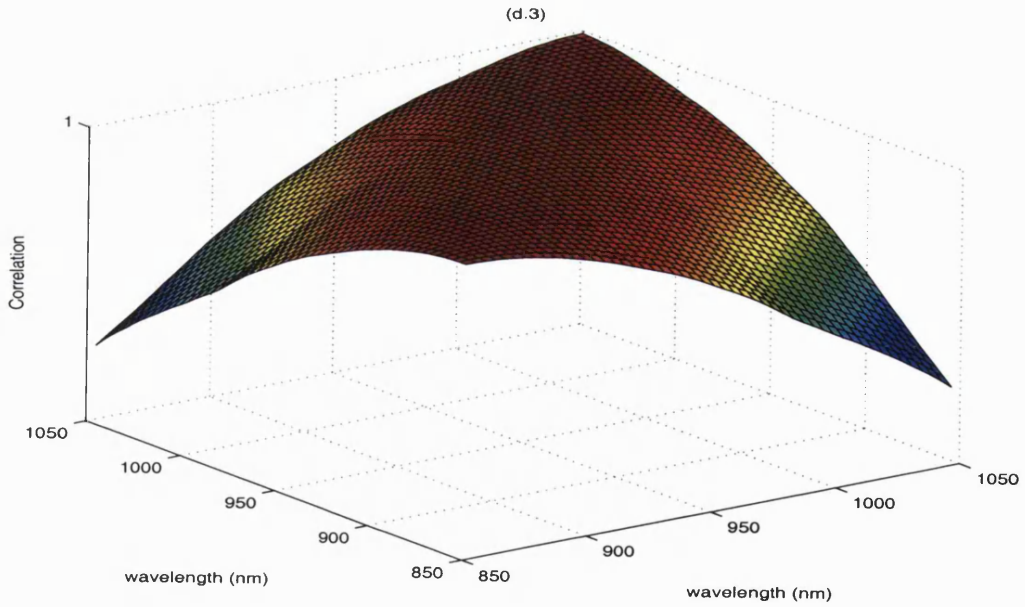
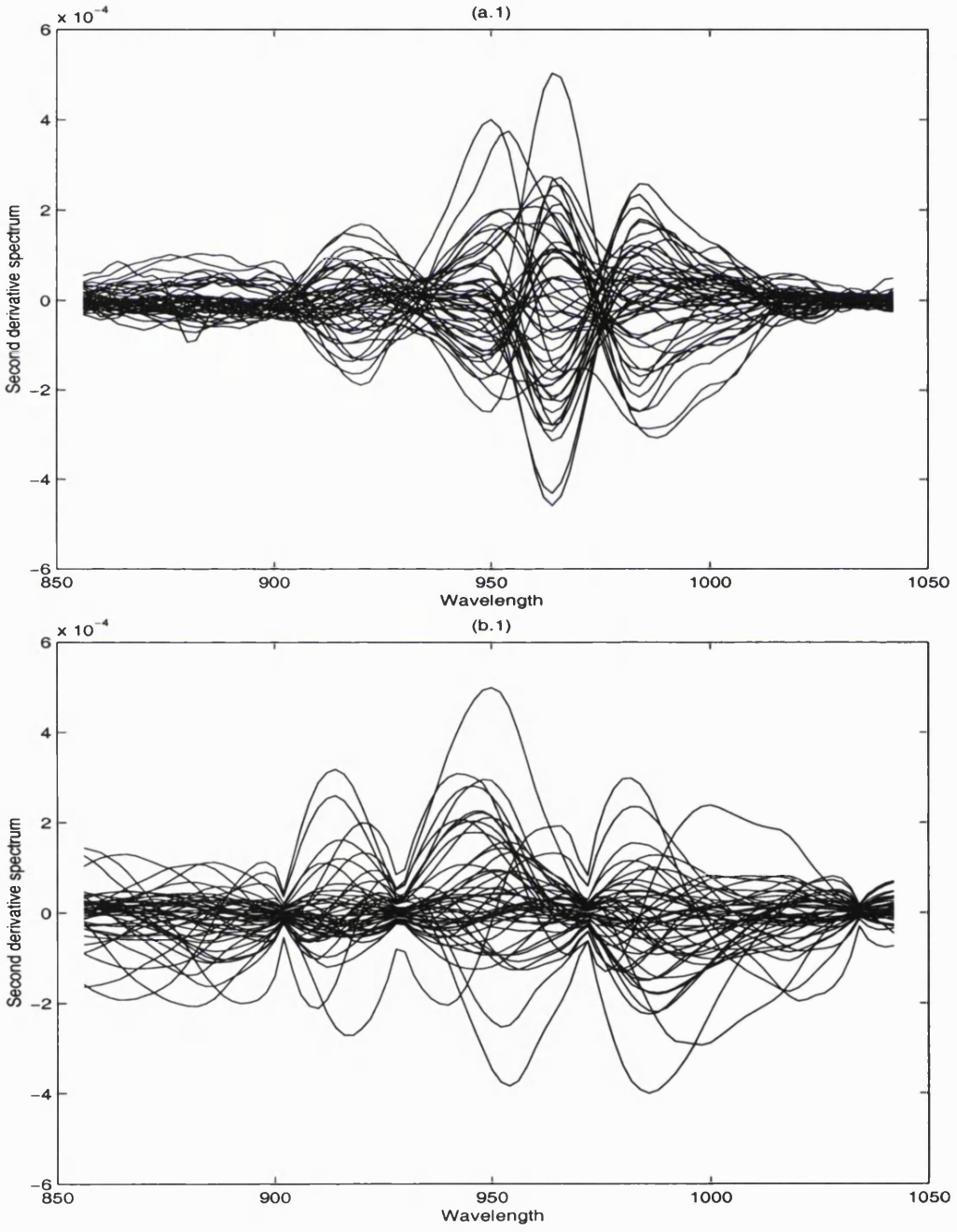
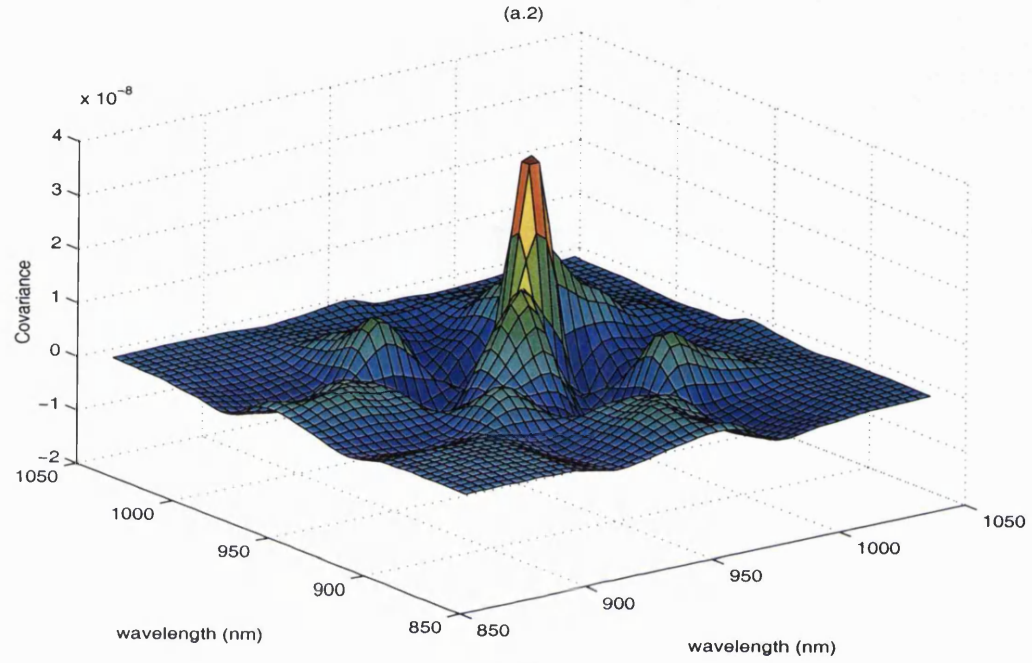
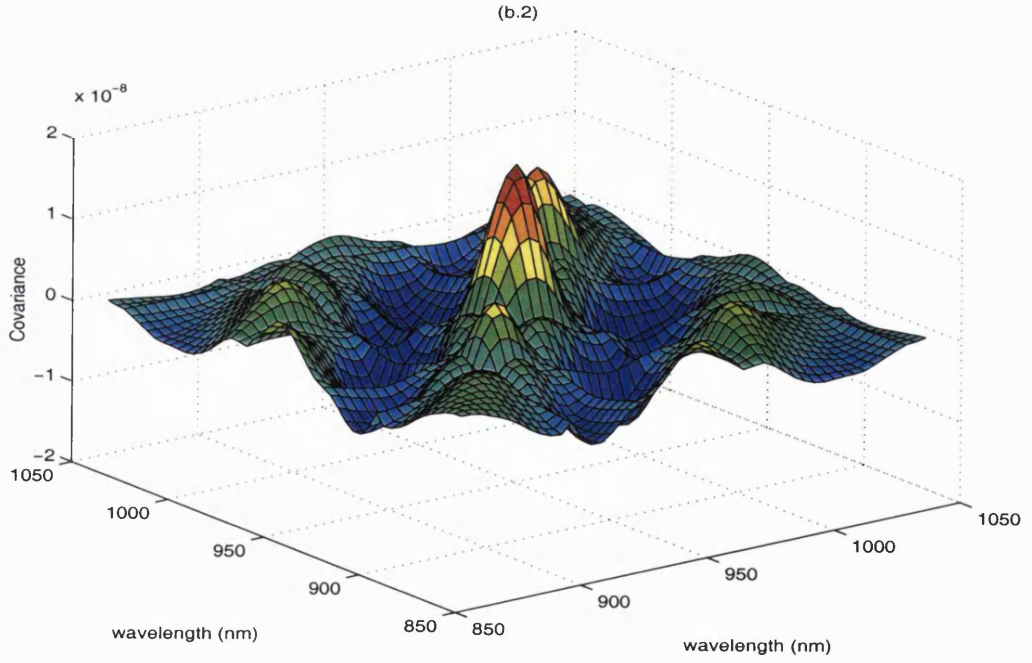
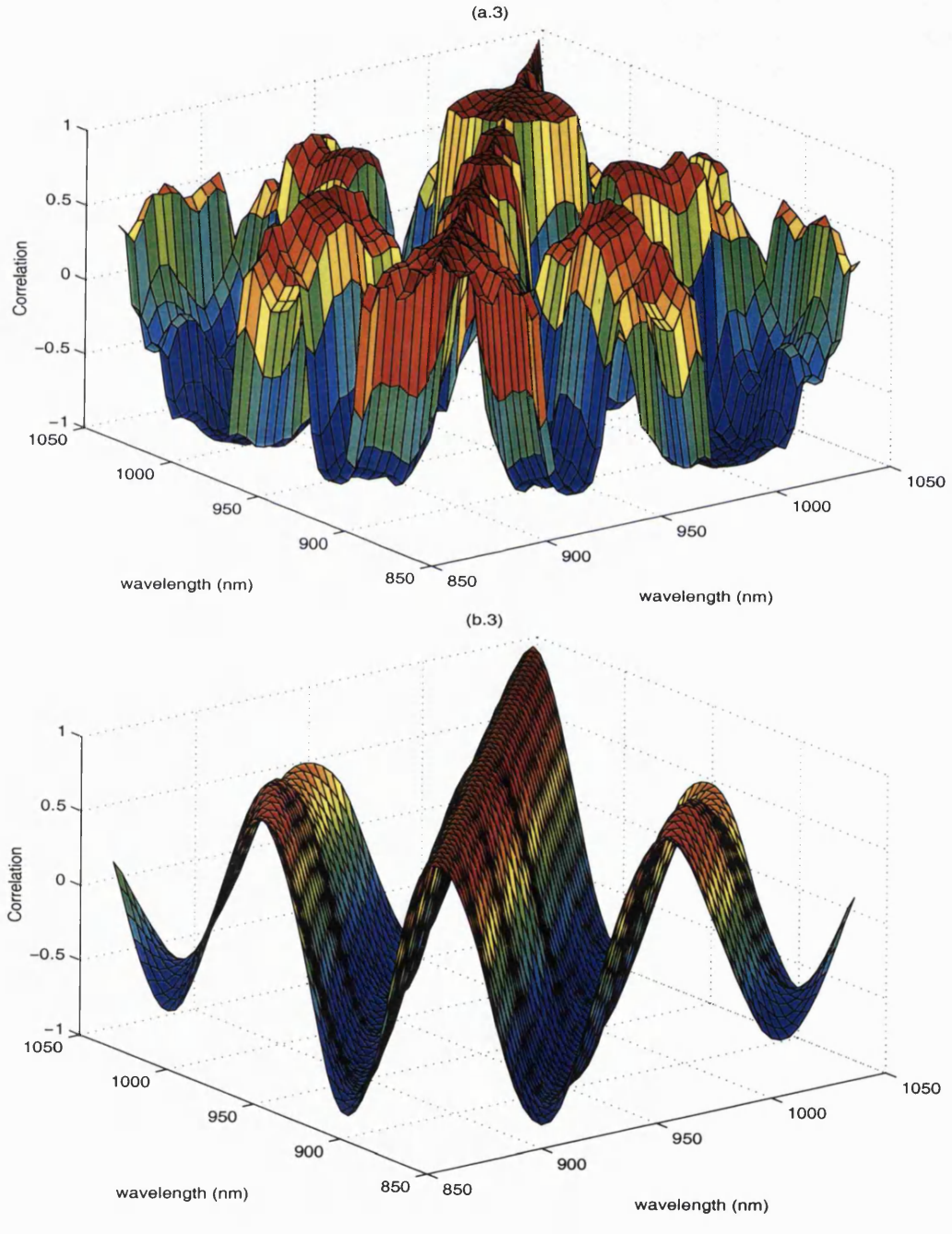

(d.3)
Correlation
1
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)

Figure 5.4: The 2nd derivative spectra, their covariance matrices and correlation matrices.




(a.2)
x 10^-8
4
3
2
1
0
−1
−2
Covariance
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)


(b.2)
x 10^-8
2
1
0
−1
−2
Covariance
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)


(a.3)
Correlation
1
0.5
0
−0.5
−1
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)
(b.3)
Correlation
1
0.5
0
−0.5
−1
1050
1000
950
900
850
wavelength (nm)
850
900
950
1000
1050
wavelength (nm)

Chapter 6

Bayesian Regression with Many Variables

6.1 Introduction

When considering multiple regression modelling using a Bayesian approach, the relative numbers of samples and variables does not present quite the same problem as it does in the classical approach. In classical regression, the number of training samples has to be greater than the number of explanatory variables in order that the sample covariance matrix of the explanatory variables is invertible and consequently the maximum likelihood estimates and the least squares estimates of the regression coefficients exist. Therefore, techniques for variable-selection or regularised regression methods have to be considered for the classical approach in order to reduce the number of variables. In a Bayesian framework, this constraint does not exist. When the number of observations is much larger than the number of variables, improper non-informative prior distributions for the regression coefficients in the controlled regression analysis or for the covariance matrix of regressors in random regression are often chosen. Then the posterior means of the regression coefficients are effectively the same as the classical solution. However, when the number of variables is large, the use of improper priors frequently leads to degenerate posterior distributions for the parameters of the regression model.

The problem can be avoided by using proper prior density functions instead of improper prior density functions. It is known that when the number of observations is small, the prior density functions are very informative and hence very influential for the posterior results. Therefore, care has to be taken when selecting prior density functions.

The approach to choosing prior density functions in this thesis follows the idea in Brown [23] considering the problem of Bayesian regression with many variables (see chapter 5). He uses a natural conjugate prior for the normal regression model and supposes there is a simple pattern in the covariance matrix of regression coefficients or in the expected covariance matrix of the regressors. The structure assumption for the covariance matrix or expected covariance matrix aims to limit the number of parameters and keep the computational simplicity. Structural coherence is another principle Brown [23] suggested when considering refinements of regressors (see section 5.5). Implementation of similar idea can be found in Brown *et al.* [28] [25], where the focus is on a methodology for variable selection. The correlation structure used there is an identity matrix, which keeps the number of parameters to a minimum and keeps the posterior density function relatively simple. In this thesis, we also consider covariance structures that provide better prior information whilst still not making the posterior density function too complicated.

We consider random regression models where the number of regressors can in principle be increased to infinity. We use an NIR calibration problem (the first example in chapter 5) as an example. We regress the protein content of wheat samples on the NIR spectra of wheat samples in our example. NIR absorption of wheat samples has been measured at 100 wavelengths from 850nm to 1048nm with 2nm increments at each step, denoted as $X_{\lambda_{850}}, X_{\lambda_{852}}, \ldots, X_{\lambda_{1048}}$, where X_{λ_ω} (1×1) represents the absorption measured at wavelength ω nm. These measurements are used as regressors to predict the protein content. Absorption measurements can be taken at any number of wavelengths within the NIR band. Hence, the NIR spectra can be considered as continuous-time (continuous-wavelength) random function, and $\{X_{\lambda_{850}}, X_{\lambda_{852}}, \ldots, X_{\lambda_{1048}}\}$ is then a discrete subset of the random function. A

special property of the NIR spectra is that although we can increase the number of wavelengths where we measure the absorption with a fixed range, the information about protein content contained in an NIR spectrum within this fixed range is limited. Even if the number of wavelengths we choose is infinite within this range, the protein content will not be predicted perfectly by these regressors. One possible way to gather more information about protein content would be to extend the coverage of the wavelengths at which the spectra are recorded.

Suppose Y is the response variable and $X = (X_1, X_2, \ldots, X_p)$ are the explanatory variables or regressors and p can be arbitrarily large. The equation of the multiple regression model is

$$Y = X\beta + E.$$

In random regression, Y and X are assumed jointly normally distributed centred at zero with an unknown covariance matrix. A conventional assumption for the prior distribution of the covariance matrix is an inverse-Wishart distribution. The inverse-Wishart distribution is a conjugate prior distribution for the covariance matrix of the normal sampling distribution. It has been shown by Dawid [43] that this natural conjugate prior assumption for the parameters of a random regression implies that a response can be predicted perfectly using an infinite number of regressors given the hyperparameters. However, in our example and most practical cases it is believed that the response variables should not be able to be predicted perfectly even when there is an infinite number of regressors. A non-conjugate random regression model was conceptually suggested by Mäkeläinen and Brown [93] in order to avoid perfect prediction using an infinite number of regressors. The sampling distribution for the non-conjugate model is an extension of the sampling distribution for the conjugate model. It assumes that Y is the sum of two latent variables η and α. The first, η, and the regressors are jointly multivariate normally distributed while α is independent of all the regressors. When the prior distribution of the covariance matrix of η and the regressors is an inverse-Wishart distribution, the predictive error of η using the regressors tends to zero as p goes to infinity whatever the scale matrix or shape parameters of the inverse-Wishart distribution

are. However, the variation of Y will never be explained completely by X since α is independent of X. Fang and Dawid [54] investigated the asymptotic properties of the non-conjugate model when the number of explanatory variables goes to infinity given the prior expected covariance of variables. Brown *et al.* [25] applied the non-conjugate model with a simulation-based variable selection method to a practical case.

In this chapter, the non-conjugate regression model is applied to our example. We further assign hyper priors to the hyperparameters. We also apply the concept of using a structural covariance matrix for the joint distribution of Y and X. Several coherent structural assumptions for the covariance of X have been introduced in chapter 5. The assumption for the structure of the covariance between Y and X is kept simple in order to focus on the performance of the models arising from the structural assumptions for the covariance matrix of the regressors. We use ARMS within Gibbs sampling as our sampling scheme (see chapter 4). A comparison of models with different covariance structures will be presented.

6.2 Sampling Model

Consider the case of a random regression model with one response variable and p explanatory variables. The number of i.i.d. samples is n, and each sample is denoted by $(Y_{[i]}, X_{[i]})$ for $i = 1 \ldots n$, where $Y_{[i]}$ is the response variable and $X_{[i]} = (X_{[i]1}, X_{[i]2}, \ldots X_{[i]p})$ are the regressors. Here $Y_{[i]}$ is 1 by 1 and $X_{[i]}$ is 1 by p. Let

$$Y_{n\times 1} = \begin{bmatrix} Y_{[1]} \\ Y_{[2]} \\ \vdots \\ Y_{[n]} \end{bmatrix}, \quad X_{n\times p} = \begin{bmatrix} X_{[1]} \\ X_{[2]} \\ \vdots \\ X_{[n]} \end{bmatrix}.$$

Suppose $Y_{[i]}$ is the sum of two unobservable independent variables $\eta_{[i]}$ and $\alpha_{[i]}$. The joint distribution of $\eta_{[i]}$ and $X_{[i]}$ is a multivariate normal distribution with mean zero and covariance matrix Σ. The latent variable $\alpha_{[i]}$ is normally distributed with mean zero. Let $\eta = (\eta_{[1]}, \eta_{[2]}, \ldots, \eta_{[n]})^t$ and $\alpha = (\alpha_{[1]}, \alpha_{[2]}, \ldots, \alpha_{[n]})^t$, where both of

these are n by 1 vectors. The model is expressed as

$$\begin{aligned} &Y = \eta + \alpha, \\ &(\eta, X) \sim \mathcal{N}(I_n, \Sigma), \\ &\alpha \sim \mathcal{N}(I_n, \Phi), \\ &\alpha \perp\!\!\!\perp (\eta, X) | (\Sigma, \Phi), \end{aligned}$$

where Σ is a $(p+1)$ by $(p+1)$ positive definite symmetric matrix, Φ is a positive scalar, and I_n is an n th order identity matrix. According to the above assumptions, the joint distribution of Y and X is matrix-normal

$$(Y, X) \sim \mathcal{N}\left(I_n, \Sigma + \begin{bmatrix} \Phi & 0 \\ 0 & 0 \end{bmatrix}\right). \tag{6.1}$$

The covariance matrix Σ can be partitioned as

$$\begin{bmatrix} \Sigma_{\eta\eta} & \Sigma_{\eta x} \\ \Sigma_{x\eta} & \Sigma_{xx} \end{bmatrix}$$

where $\Sigma_{\eta\eta}$ (1×1) is the variance of $\eta_{[i]}$, Σ_{xx} $(p \times p)$ is the covariance matrix of $X_{[i]}$, $\Sigma_{\eta x}$ $(1 \times p)$ is the covariance vector of $\eta_{[i]}$ and $X_{[i]}$, and $\Sigma_{\eta x} = \Sigma_{x\eta}^t$. The prior variance of Y is $\Sigma_{\eta\eta} + \Phi$.

Define

$$\begin{aligned} &\beta = \Sigma_{xx}^{-1}\Sigma_{x\eta} && (p \text{ by } 1), \\ &\Gamma = \Sigma_{\eta\eta.x} = \Sigma_{\eta\eta} - \Sigma_{\eta x}\Sigma_{xx}^{-1}\Sigma_{x\eta} && (1 \text{ by } 1). \end{aligned}$$

Since Σ is positive definite, Γ is positive. It can be easily shown that $(\Sigma_{xx}, \beta, \Gamma)$ is an one-to-one function of Σ. It can be shown that the joint distribution of Y and X in equation (6.1) is equivalent to the joint distribution of $Y|X$ and X, for which the sampling model is

$$\begin{cases} Y|X \sim X\beta + \mathcal{N}(I_n, \sigma^2), \\ X \sim \mathcal{N}(I_n, \Sigma_{xx}), \end{cases} \tag{6.2}$$

where $\sigma^2 = \Gamma + \Phi$ is a scalar. The first equation in (6.2) is in the standard form of a regression model with regression coefficients β and standard deviation σ. Given

$\Phi = 0$, the non-conjugate model reduces to the conjugate case and $\sigma^2 = \Gamma$. The methodology we use for modelling the non-conjugate case can be applied to the conjugate case. For hierarchical modelling the conjugate case has similar numerical problems to the non-conjugate case.

The second equation in (6.2), which provides the information of the variation of X, is often neglected in Bayesian regression analysis. By doing this, we are implicitly assuming that X is fixed. Although the prior distributions of β and Γ are derived from the joint distribution of X and Y, their posterior distributions do not contain the information for the variation of X because the information for the variation of X is only carried by the likelihood function of X, which is automatically dropped in a standard regression analysis conditional on X. In the spirit of the random regression analysis with a good informative prior, a full hierarchical model should include the likelihood function of X so that the posterior distributions of the hyperparameters related to X can be updated with the maximum information about X.

6.3 Prior Assumptions

Let

$$\begin{aligned} \Sigma|\delta, Q &\sim \mathcal{IW}(\delta; Q), \\ \Phi|\nu, K &\sim \mathcal{IW}(\nu; K), \end{aligned} \tag{6.3}$$

where $Q > 0$ ($p+1$ by $p+1$), $K > 0$ (1 by 1), $\delta > 0$ (1 by 1) and $\nu > 0$ (1 by 1), and $\Sigma \perp\!\!\!\perp \Phi|(\delta, Q, \nu, K)$. When δ and ν are larger than two, the expectations of Σ and Φ exist and are equal to $Q/(\delta - 2)$ and $K/(\nu - 2)$ respectively.

From the distribution of Σ in (6.3), it can be deduced that the distributions of Σ_{pp}, β and Γ given δ, Q, ν and K are

$$\begin{aligned} \Sigma_{xx} &\sim \mathcal{IW}(\delta; Q_{xx}), \\ \beta|\Gamma &\sim Q_{xx}^{-1}Q_{x\eta} + \mathcal{N}(Q_{xx}^{-1}, \Gamma), \\ \Gamma &\sim \mathcal{IW}(\delta + p; Q_{\eta\eta.x}), \end{aligned} \tag{6.4}$$

where

$$Q = \begin{bmatrix} Q_{\eta\eta} & Q_{\eta x} \\ Q_{x\eta} & Q_{xx} \end{bmatrix},$$

and

$$Q_{\eta\eta.x} = Q_{\eta\eta} - Q_{\eta x} Q_{xx}^{-1} Q_{x\eta}.$$

Since Q is positive definite, $Q_{\eta\eta.x}$ is greater than zero. According to the natural property of an inverse-Wishart distribution, Σ_{xx} is independent of (β, Γ) conditional on Q. Since, Φ and Σ are independent, Φ and (β, Γ) are also independent. For the convenience of computing, we define $\theta = \Phi\,\Gamma^{-1}$ (scalar), which is the ratio of Φ to Γ and leads to $\sigma^2 = (1+\theta)\,\Gamma$. The distribution of θ is a matrix-F distribution $\mathcal{F}(\delta + p, \nu; KQ_{\eta\eta.x}^{-1})$ and $\theta|\Gamma, K \sim \mathcal{IW}(\nu, K\Gamma^{-1})$.

Suppose the expected covariance matrix of (Y, X) has a deterministic parametric structure. That is, the scale matrix Q of the prior distribution of Σ has a deterministic structure. We consider Q as a matrix function of some hyperparameters $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_m)$, where the number of hyperparameters m is much less than $p(p+1)/2$. The scale matrix Q is decomposed as in chapter 5. Let $Q = \Lambda R \Lambda$, where Λ is a diagonal matrix and the diagonal elements of R are one's. Given Q, the elements of the diagonal of $\Lambda\Lambda/(\delta - 2)$ are just the expectation of the variance of $(\eta_{[i]}, X_{[i]})$. We partition R as $R_{\eta\eta}$, R_{xx}, $R_{\eta x}$, and $R_{x\eta}$, and partition Λ into $\Lambda_{\eta\eta}$ Λ_{xx} $\Lambda_{\eta x}$ and $\Lambda_{x\eta}$ so that $Q_{\eta\eta} = \Lambda_{\eta\eta}\Lambda_{\eta\eta}$, $Q_{xx} = \Lambda_{xx}R_{xx}\Lambda_{xx}$, $Q_{x\eta} = \Lambda_{xx}R_{x\eta}\Lambda_{\eta\eta}$. Since Λ is a diagonal matrix, $\Lambda_{x\eta}$ and $\Lambda_{\eta x}$ are zero vectors, and $\Lambda_{\eta\eta}$ is a scalar and Λ_{xx} is diagonal. Several structures for R_{xx} have been introduced in chapter 5. Denote the prior density functions of κ and K (in model 6.3) as $\pi(\kappa)$ and $\pi(K)$ respectively. For simplicity and lack of prior knowledge, we suppose all the hyperparameters (elememts of κ and K) are independent. The hierarchical structure of the model can be summarised by the directed graph in figure 6.1. It is important to use proper prior density functions for K and κ. Using improper prior distributions may yield improper marginal posterior distributions for parameters. Since the model is so sophisticated, it is difficult to examine whether a posterior density function is proper or not. By using proper prior distributions for parameters, the

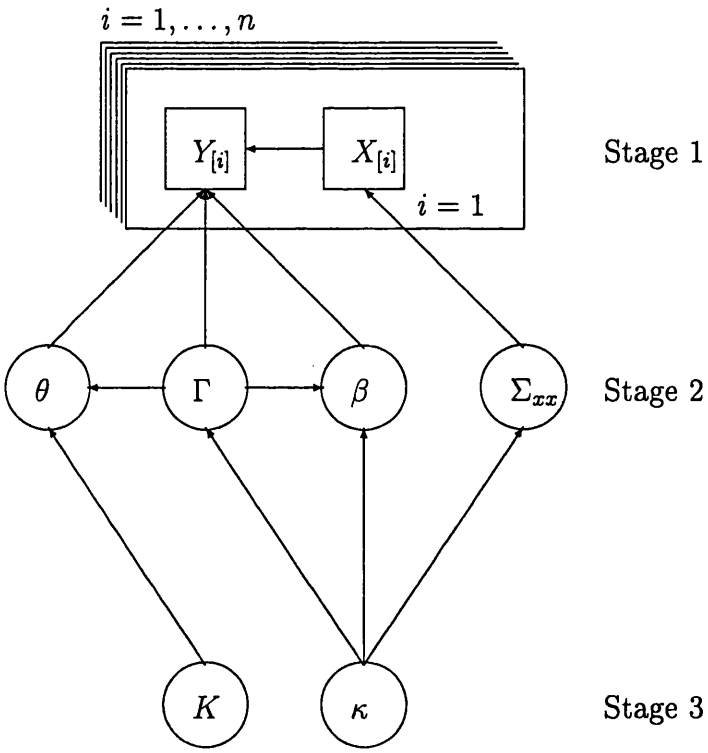


Figure 6.1: Graphical presentation of the non-conjugate random regression model

danger can be avoided.

6.4 Estimation

Suppose the training data matrix (Y_t, X_t) is an n by $p+1$ matrix, and has the same model as (Y, X). Given the prior assumptions in the last section, the joint posterior density function of all the parameters is

$$
\begin{aligned}
& P(\Sigma_{xx}, \beta, \theta, \Gamma, K, \kappa | Y_t, X_t) \\
\propto\ & \underbrace{P(Y_t|X_t, \beta, \theta, \Gamma)P(X_t|\Sigma_{xx})}_{\text{likelihood function}} \underbrace{\pi(\Sigma_{xx}|\kappa)\pi(\beta|\Gamma, \kappa)\pi(\theta|\Gamma, K)\pi(\Gamma|\kappa)}_{2^{\text{nd}}\text{ stage}} \underbrace{\pi(K)\pi(\kappa)}_{3^{\text{rd}}\text{ stage}} \\
=\ & \theta^{-\frac{\nu+2}{2}}(1+\theta)^{-\frac{n}{2}}|\Gamma|^{-\frac{\delta+\nu+2p+n+2}{2}}|Q_{xx}|^{\frac{\delta+p}{2}}|Q_{\eta\eta.x}|^{\frac{\delta+p}{2}}|K|^{\frac{\nu}{2}}|\Sigma_{xx}|^{-\frac{\delta+n+2p}{2}} \\
& \exp[-\frac{1}{2}\mathrm{tr}\Gamma^{-1}(\beta - D_{xx}^{-1}D_{x\eta})^t D_{xx}(\beta - D_{xx}^{-1}D_{x\eta})] \\
& \exp[-\frac{1}{2}\mathrm{tr}\Gamma^{-1}(D_{\eta\eta.x} + \theta^{-1}K)] \exp[-\frac{1}{2}\mathrm{tr}\Sigma_{xx}^{-1}(Q_{xx} + {X_t}^t X_t)] \\
& \pi(K)\pi(\kappa),
\end{aligned}
$$

where

$$
\begin{aligned}
D_{\eta\eta} &= Q_{\eta\eta} + (1+\theta)^{-1}Y_t^t Y_t \quad (1 \text{ by } 1), \\
D_{xx} &= Q_{xx} + (1+\theta)^{-1}{X_t}^t X_t \quad (p \text{ by } p), \\
D_{x\eta} &= Q_{x\eta} + (1+\theta)^{-1}{X_t}^t Y_t \quad (p \text{ by } 1), \\
D_{\eta x} &= D_{x\eta}^t \quad (1 \text{ by } p).
\end{aligned}
$$

and $D_{\eta\eta.x} = D_{\eta\eta} - D_{\eta x}D_{xx}D_{x\eta}$. Let

$$
D \equiv \begin{bmatrix} D_{\eta\eta} & D_{\eta x} \\ D_{x\eta} & D_{xx} \end{bmatrix} = Q + \frac{1}{1+\theta}\begin{bmatrix} \eta & X_t \end{bmatrix}^t \begin{bmatrix} \eta & X_t \end{bmatrix}.
$$

Since Q is positive definite, D is also positive definite, therefore, $D_{\eta\eta.x} > 0$ and $D_{\eta\eta.x} + \theta^{-1}K > 0$.

Due to the large number of parameters and the complexity of the model, MCMC is the most appropriate method for inference. ARMS within Gibbs sampling is a general efficient method for stochastic simulation when the full conditional density functions cannot be sampled directly. In Gibbs sampling, samples

are generated using the full conditional density functions of parameters, and any full conditional density function only needs to be known up to a multiple of its exact density function. Due to the complexity of the joint posterior density function, multiple MCMC chains are run in order to check whether MCMC converges to unique stationary distribution. A more detailed description of MCMC and our sampling plan have been introduced in chapter 4.

ARMS within Gibbs sampling requires the full conditional density functions for the parameters. The full conditional distributions of the parameters β, Γ, and Σ_{xx} belong to the same distribution families as their prior distributions

$$\begin{aligned}
&\beta|X_t, Y_t, \Gamma, \theta, \kappa \sim H + \mathcal{N}(D_{xx}^{-1}, \Gamma), \qquad (6.5)\\
&\Gamma|X_t, Y_t, \beta, \theta, \kappa, K \sim \mathcal{IW}(\delta + \nu + n + 2p; (\beta - H)^t D_{xx}(\beta - H) + D_{\eta\eta.x} + \frac{K}{\theta}),\\
&\Sigma_{xx}|X_t, \kappa \sim \mathcal{IW}(\delta + n; Q_{xx} + X_t^t X_t),
\end{aligned}$$

where $H = D_{xx}^{-1} D_{x\eta}$. The full conditional density function of θ is

$$f(\theta|X_t, Y_t, \beta, \Gamma, K) \propto \theta^{-\frac{\nu+2}{2}}(1+\theta)^{-\frac{n}{2}} \exp -\frac{1}{2}\mathrm{tr}\Gamma^{-1}\left[\frac{(Y_t - X_t\beta)^t(Y_t - X_t\beta)}{1+\theta} + \frac{K}{\theta}\right]$$

for $\theta > 0$ and 0 elsewhere. The full conditional density function of K is

$$\pi(K)|K|^{\frac{\nu}{2}} \exp(\Gamma^{-1}\theta^{-1}K).$$

Some of the full conditional density functions are only known up to multiples. These are sufficient for MCMC. Due to the size of Σ_{xx} when p is large and the possibility that the scale matrix $Q_{xx} + {X_t}^t X_t$ is ill-conditioned given all other parameters, it is beneficial to avoid sampling from the distribution of Σ_{xx}. Since the full conditional distribution of Σ_{xx} is an inverse-Wishart, a straightforward approach is to marginalise over it. The full conditional density functions of β, Γ, and θ are not affected by this (see figure 6.1), but the full conditional density functions of the hyperparameters relating to Σ_{xx} are changed. As a result, the full conditional density function for κ given other parameters is proportional to

$$\begin{aligned}
&\pi(\kappa)|Q_{xx}(\kappa)|^{\frac{\delta+p}{2}}|Q_{xx}(\kappa) + {X_t}^t X_t|^{-\frac{\delta+n+p-1}{2}}|Q_{\eta\eta.x}(\kappa)|^{\frac{\delta+p}{2}}\\
&\times \exp\{-\frac{1}{2}\mathrm{tr}\Gamma_p^{-1}[(\beta - \hat{\beta}(\kappa))^t Q_{xx}(\kappa)(\beta - \hat{\beta}(\kappa)) + Q_{xx.\eta}(\kappa)]\}, \qquad (6.6)
\end{aligned}$$

where $\hat{\beta} = {Q_{xx}}^{-1}(\kappa)Q_{x\eta}(\kappa)$. Denote the full conditional density function of κ_i (a component of κ) as $f(\kappa_i|\kappa_{(-i)})$, which is proportional to equation (6.6), and $\kappa_{(-i)}$ represents all parameters of the entire model except κ_i itself. In our sampling plan, we generate the components of κ one by one instead of generating κ as a whole. The elements of κ will be specified in section 6.6.

The MCMC estimates of β, Γ, θ, K and κ are denoted as $\hat{\beta}$, $\hat{\Gamma}$, $\hat{\theta}$, $\hat{K}$ and $\hat{\kappa}$, are approximations of their posterior means using the sample means of the data generated by MCMC. The posterior variance or covariance matrices of these parameters are also obtained using the sample variance or covariance matrix. Suppose we observe a future case X_f, the future response Y_f can be summarised by its predictive density function $p(Y_f|X_f)$. The expectation of Y_f under the predictive distribution $E(Y_f|X_f)$ is equal to $\mathrm{E}[E(Y_f|X_t, X_f, \beta, \Gamma, \theta)]$. Due to the hierarchical structure of the model,

$$E[E(Y_f|Y_t, X_t, X_f, \beta, \Gamma, \theta)] = E(X_f\beta|Y_t, X_t) = X_f E(\beta|Y_t, X_t),$$

and

$$E(\beta|Y_t, X_t) = E[E(\beta|Y_t, X_t, \beta_{(-1)})],$$

where $\beta_{(-1)}$ represents all parameters of the model except β. Suppose $\beta_{(-1)}^{(1)}, \ldots, \beta_{(-1)}^{(m)}$ are m MCMC samples of $\beta_{(-1)}$ drawn from the target density function of $\beta_{(-1)}$, the Rao-Blackwellised estimated density of $\beta|X_t$ is

$$\hat{p}(\beta|Y_t, X_t) = \frac{1}{m}\sum_{l=1}^{m} p(\beta|Y_t, X_t, \beta_{(-1)}^{(l)}),$$

which provides a good estimate for the density of $\beta|X_t$ (see Gelfand and Smith [63]) and do not requires the MCMC samples of β to make the estimation. Using the Rao-Blackwellised estimate for the density of $\beta|X_t$, we can obtain an estimate for $E(\beta|X_t)$

$$\hat{\beta}_{RB} = \frac{1}{m}\sum_{l=1}^{m} E(\beta|Y_t, X_t, \beta_{(-1)}^{(l)}),$$

where $E(\beta|X_t, \beta_{(-1)}^{(l)})$ is the expectation of the conditional distribution of β (see equation 6.6) given the l^{th} sample of $\beta_{(-1)}$. Hence, the expectation of $Y_f|X_f$ can be estimated by $X_f\hat{\beta}_{RB}$.

6.5 Some Numerical Problems

Numerical problems need to be taken into account when operations on large and ill-conditioned matrices are carried out in computing. These problems are efficiency, precision and overflow. Three operations that may cause most of the problems in our analysis are matrix determinant evaluation, matrix inversion and taking the square root of a square matrix. The matrices we deal with are Q_{xx}, D_{xx} and $Q_{xx}(\kappa) + X_t{}^t X_t$.

When $X_t{}^t X_t$ is singular and $Q_{xx}(\kappa)$ is nearly singular,

$$|Q_{xx}(\kappa) + X_t{}^t X_t|^{-\frac{\delta+n+p-1}{2}}$$

may be so large that even its logarithm overflows. By re-arranging

$$|Q_{xx}(\kappa)|^{\frac{\delta+p}{2}} |Q_{xx}(\kappa) + X_t{}^t X_t|^{-\frac{\delta+n+p-1}{2}}$$

in equation (6.6) as

$$|Q_{xx}(\kappa)|^{-\frac{n-1}{2}} |I + Q_{xx}(\kappa)^{-1} X_t{}^t X_t|^{-\frac{\delta+n+p-1}{2}},$$

the problem can be ameliorated. The analytical determinant of $Q_{xx}(\kappa)$ is available in simple form for the structures which have been introduced in chapter 5. However, there is no simple form for $|I + Q_{xx}(\kappa)^{-1} X_t{}^t X_t|$. When a matrix is ill-conditioned, the ratio of the largest eigenvalue to the smallest eignevalue is large, the computing error for both inversion and taking determinants tends to be large. We rescaling an ill-conditioned matrix so that the ratio of the largest to the smallest eigenvalue becomes smaller because it helps to reduce the computing error (Atkinson [3]).

After the re-arrangement of equation (6.6), the unnormalised $f(\kappa_i|\kappa_{(-i)})$ at its maximum in the parameter space can still be larger than the maximum real number the computer can handle. A further idea to overcome this problem is to rescale $f(\kappa_i|\kappa_{(-i)})$ to $g(\kappa_i|\kappa_{(-i)})$ so that the maximum of $g(\kappa_i|\kappa_{(-i)})$ is not so large. Theoretically, one can choose $g(\kappa_i|\kappa_{(-i)})$ as $f(\kappa_i|\kappa_{(-i)})/m$, where m is the maximum of $f(\kappa_i|\kappa_{(-i)})$ in the parameter space. As a result, the maximum of

$g(\kappa|\kappa_{(-i)})$ is close to 1. When using of ARMS within Gibbs sampling, one can choose

$$m = \max_{\kappa_i \in S} f(\kappa_i|\kappa_{(-i)}),$$

where S is the initial set of abscissae for κ_i in each iteration of Gibbs sampling (see chapter 4). With a good choice of S, m should be fairly close to the maximum of $f(\kappa_i|\kappa_{(-i)})$ in its parameter space.

The frequently used methods for inverting a symmetric matrix are the Cholesky decomposition, Gauss-Jordan elimination, LU decomposition, and singular value decomposition. Our MCMC algorithm is implemented using Matlab, a mathematical environment specialised for linear algebra and matrix manipulation. These methods can easily be implemented in Matlab. In calculating $C^{-1}b$ where C is a square matrix, Matlab provides a *backslash* operator '\', which applies the algorithms for solving linear systems so that $C\backslash b$ is a much more efficient way of calculating $C^{-1}b$ than inverting C first then multiplying C^{-1} by b. Since the analytical inverse of $Q_{xx}(\kappa)$ is also available, we can apply the binomial inverse theorem (see Brown [23]) when inverting D_{xx} for simulating β. The actual matrix we need to invert is then reduced to a n by n matrix, whose size is much smaller and whose condition is much better than D_{xx}. Efficiency and precision can then be improved.

In order to simulate β, we also need to calculate the square root of D_{xx}, since the full conditional distribution of β is equal to

$$D_{xx}{}^{-1}D_{x\eta} + A\mathcal{N}(I_p, 1)\Gamma^{\frac{1}{2}},$$

where $D_{xx} = AA^t$. Since D_{xx} is symmetric and positive definite, an obvious choice of A is the factor in the Cholesky decomposition for D_{xx}. One can also consider the principal components decomposition that $D_{xx} = V\Lambda V^t$, where the Λ is a diagonal matrix and the elements on its diagonal are the eigenvalues of D_{xx}, and columns of V are the corresponding eigenvectors so that $A = V\Lambda^{\frac{1}{2}}$. In our computation, we use the first one.

6.6 Example

We analyse the first example in chapter 3 using the hierarchical non-conjugate random regression model. The data contains 50 samples of wheat. For each wheat sample we have a measurement of the percentage of protein content as the response, and 100 measurements of an NIR transmission spectrum at 100 fixed equally spaced wavelengths as the regressors. Similar examples have been analysed successfully using partial least squares, principal components regression and other methods [94]. We consider the 5 prior models M.a, M.b, M.c, M.d, and M.e for X which have been discussed in chapter 5 (see table 5.1). The first three models are for the original spectra while the last two models are for the second derivative NIR spectra. The difference in the settings of the five models is only in Λ_{xx} and R_{xx}, while $\Lambda_{\eta\eta}$ and $R_{\eta x}$ are the same for these five models. The focus in this chapter is in the effect of the variation of the structure of R_{xx}, so we suppose $R_{\eta x} = (\rho, \rho, \ldots, \rho)$ for all models. The structure of Λ_{xx} is suggested by equation (5.10) in chapter 5

$$\Lambda_{xx}(i,i) = \sqrt{|\mu_i|}\exp(a + bi), \quad a, b \in \mathrm{R},$$

where a and b are random hyperparameters in the models in models for the original spectra. In the models for the second derivative spectra, a is a random hyperparameter while $b = 0$. In order to reduce the correlation of parameters, equation (5.10) is reparameterised as

$$\Lambda_{xx}(i,i) = \sqrt{|\mu_i|}\exp[a + b(i - \bar{i})],$$

where $\bar{i} = p^{-1}\sum_{i=1}^{p} i$.

With the AR(1) structure for expected correlation, there is one parameter τ in R_{xx}, where $R_{xx}(i,j) = \tau^{|i-j|}$, $0 \leq \tau < 1$. For an AR(2) correlation function, there are originally two hyperparameters ϕ_1 and ϕ_2 in R_{xx} (see chapter 5). The parameter space of (ϕ_1, ϕ_2) has to satisfy the stationary condition of an AR(2) process that

$$-1 < \phi_2 < 1, \quad \phi_2 + \phi_1 < 1, \quad \phi_2 - \phi_1 < 1.$$

To make computing easier, we further eliminate one hyperparameter by fixing the relationship between ϕ_1 and ϕ_2, and take ϕ_1 as the only parameter in R_{xx}. As in chapter 5, we suppose $\phi_2 = -\phi_1 + 0.99985$ for M.c and $\phi_2 = -\phi_1^2/[4\cos(2\pi/50)^2]$ for M.e.

According to our preliminary investigations, the posterior models are not sensitive to the shape of the hyper prior distribution, although they may be sensitive to the chosen ranges for some hyperparameters. Originally we have the following assumptions for the hyperparameters. We suppose the prior for a is $N(0, 100)$ for all five models. For b in M.a, M.b, and M.c, their prior density functions are proportional to the density of $N(0, 100)$ restricted to the parameter space $(-\infty, 0]$. The τ in M.b has a uniform(0,1) prior distribution. The ϕ_1 in M.c and M.e are also uniformly distributed in the stationary area under the constraints with ϕ_2. In all our models, ρ to satisfy a constraint such that R is non-singular or equivalently $|R| > 0$. Since

$$|R| = |R_{xx}||1 - R_{\eta x}R_{xx}R_{x\eta}|,$$

and $|R_{xx}| > 0$ under our assumptions for τ and ϕ_1, an essential condition is that $|1 - R_{\eta x}R_{xx}R_{x\eta}| > 0$. That is, $lb < \rho < rb$, where $rb = -lb = \rho_{\text{limit}}$ and

$$\rho_{\text{limit}} = \frac{1}{\sqrt{\text{sum of all elements in } R_{XX}^{-1}}}. \tag{6.7}$$

For the original NIR spectra, we found that in the sample correlation between Y_t and X_t are all negative, hence we assume $rb = 0$ and $lb = -\rho_{\text{limit}}$. For the 2^{nd} derivative spectra, the sample correlation between Y_t and X_t can also be positive, therefore, we retain the right bound of the parameter space of ρ as (6.7). The prior distribution of $\Lambda_{\eta\eta}$ is a diffuse $\mathcal{W}(1; 100)$. For the prior distribution for K, we use the diffuse $\mathcal{W}(1; 100)$ as well. In practice, we actually further constrained the parameter spaces of τ and ϕ_1 in our inference for numerical reasons. We shall discuss this later.

6.7 Scoring Rule and Sensitivity Analysis

In order to evaluate the model, cross-validation is applied. Since the number of samples is small and the number of variables large, every sample can be very influential to the model. A leave-one-out validation makes the best use of all available observations. However, it is time-consuming for our complicated model. The leave-one-block-out validation method has been considered as an alternative to the leave-one-out method in many practical cases taking into account the cost of computing the score. In a leave-one-block-out approach, the data are divided into several blocks, denoted as $(Y_{(1)}, X_{(1)}), (Y_{(2)}, X_{(2)}), \ldots, (Y_{(k)}, X_{(k)})$, with m_i samples in the i-th block and the number of blocks is much less than the number of samples. Denote $(Y_{(-i)}, X_{(-i)})$ as the data set contains all the data in (Y, X) except $(Y_{(i)}, X_{(i)})$. The response variable of the i-th block is estimated by $X_{(i)}\hat{\beta}_{(-i)}$, where $\hat{\beta}_{(-i)}$ is the MCMC estimation of β using data $(Y_{(-i)}, X_{(-i)})$. Define $r_{(i)} = Y_{(i)} - X_{(i)}\hat{\beta}_{(-i)}$, which is the vector of predictive residuals of the i-th block, and $r_{(i)} = (r_{(i),1}, r_{(i),2}, \ldots, r_{(i),m_i})$. The mean square error of prediction (MSEP) (see Martens and Næs [94]) of the model is

$$MSEP = \frac{1}{k}\sum_i \sum_j \frac{r_{(i),j}^2}{m_i}.$$

In our example, samples are randomly divided into 5 blocks.

The concern in this chapter is the performance of the model on the NIR spectral data with different structural assumptions for the covariance matrix. Their posterior performance can be compared using MSEP as a scoring rule. Due to the cost of computing, only the effect of the structural assumption for Q_{xx} will be investigated. We do not vary the hyper priors for the hyperparameters. Instead, we fixed τ and ϕ_1 to see how different values of them affect the inference. According to our preliminary investigation, the posterior distributions of the parameters are not sensitive to the shape assumption for the density function of the hyperparameters in Q_{xx}.

Table 6.1: Parameters in the five models

Note: $\Omega = \{\phi_1 | -1 < \phi_2 < 1, \phi_2 + \phi_1 < 1, \phi_2 - \phi_1 < 1\}$

Model	Parameters
M.a	$\beta \in \mathrm{R}^p$, $\Gamma > 0$, $\theta > 0$, $K > 0$, $\Lambda_{\eta\eta} > 0$, $a \in \mathrm{R}$, $b < 0$, $-\rho_{\text{limit}} < \rho < 0$
M.b	$\beta \in \mathrm{R}^p$, $\Gamma > 0$, $\theta > 0$, $K > 0$, $\Lambda_{\eta\eta} > 0$, $a \in \mathrm{R}$, $b < 0$, $-\rho_{\text{limit}} < \rho < 0$, $0 \leq \tau < 1$
M.c	$\beta \in \mathrm{R}^p$, $\Gamma > 0$, $\theta > 0$, $K > 0$, $\Lambda_{\eta\eta} > 0$, $a \in \mathrm{R}$, $b < 0$, $-\rho_{\text{limit}} < \rho < 0$, $\phi_1 \in \{\phi_1 \mid \phi_1 \in \Omega, \phi_2 = -\phi_1 + 0.99985\}$
M.d	$\beta \in \mathrm{R}^p$, $\Gamma > 0$, $\theta > 0$, $K > 0$, $\Lambda_{\eta\eta} > 0$, $a \in \mathrm{R}$, $-\rho_{\text{limit}} < \rho < \rho_{\text{limit}}$
M.e	$\beta \in \mathrm{R}^p$, $\Gamma > 0$, $\theta > 0$, $K > 0$, $\Lambda_{\eta\eta} > 0$, $a \in \mathrm{R}$, $-\rho_{\text{limit}} < \rho < \rho_{\text{limit}}$, $\phi_1 \in \{\phi_1 \mid \phi_1 \in \Omega, \phi_2 = -\phi_1^2/[4\cos(2\pi/50)^2]\}$

Sampling model and the first stage priors
$Y_t \mid X_t, \beta, \Gamma, \theta \sim X\beta + \mathcal{N}(I_n, \Gamma + \theta\Gamma)$,
$X_t \mid \Sigma_{xx} \sim \mathcal{N}(I_n; \Sigma_{xx})$,
$\Sigma_{xx} \sim \mathcal{IW}(\delta; Q_{xx})$,
$\beta \mid \Gamma, a, b, \rho, \Lambda_{\eta\eta}, \xi \sim Q_{xx}^{-1} Q_{x\eta} + \mathcal{N}(Q_{xx}^{-1}, \Gamma)$,
$\Gamma \mid a, b, \rho, \Lambda_{\eta\eta}, \xi \sim \mathcal{IW}(\delta + p; Q_{\eta\eta} - Q_{\eta x} Q_{xx}^{-1} Q_{x\eta})$,
$\theta \mid \Gamma, K \sim \mathcal{IW}(\nu, K\Gamma^{-1})$.
In M.a and M.d, ξ is not used.
In M.b, $\xi \equiv \tau$.
In M.c and M.e, $\xi \equiv \phi_1$.

6.8 Results and Diagnostics

6.8.1 Settings for Parameters

The parameters in the five models are summarised in table 6.1. The sampling model and the first stage priors are also briefly stated again in this table. Assume $\delta = 3$ and $\nu = 2$. The biggest possible parameter spaces for the hyperparameters a, b, ρ, τ, and ϕ_1 in five models have been introduced in section 6.4 (also shown in table 6.1). The regression coefficient β is a real vector, while Γ, $Q_{\eta\eta}$, θ, and K are all non-negative parameters. Ideally we should fit the models given the parameter spaces in table 6.1. However, we have to restrict some of them due to some numerical problems. For M.a, posterior density of θ is a very long tailed density function. The MCMC sequences for θ converge very slowly. It may be due to the high correlation of θ with other parameters, or maybe the posterior density of θ is multimodal. In order to keep the situation simple in this case, we introduce an upper bound 3000 for θ so that there is no problem of sampling from a multimodal distribution, and MCMC sequences converge in a reasonable time. For M.b, the MCMC sequences for τ converge to 1 very quickly producing singular Q_{xx} which causes overflow problems. Therefore, we assign an upper bound for τ which is near 1 but less than 1 so that Q_{xx} will not cause numerical problems. Since the density of τ increases dramatically as τ increases, the MCMC samples for τ are very likely be around the new upper bound, and the MCMC estimate for τ is very close to the upper bounds. Therefore, we choose the upper bound for τ as 0.9998 so that the estimate is not too far from our belief, which has been discussed in section 5.7. For M.c, the MCMC sequences converge slowly under the setting in table 6.1. We find that in order to improve the slow convergence, the easiest way is to use stronger priors. Instead of changing the shape of the prior density, we restrict the parameter space of ϕ_1 in a small area $[1.9835, 1.9840]$, which is also around the range we use in section 5.7 but smaller. These strong priors for τ and ϕ_1 force the posterior τ and ϕ_1 almost to be fixed values. There is not any computing problem for M.d and M.e with the settings in table 6.1. Therefore, we

do not use any stronger belief for their parameters.

6.8.2 MCMC Results

Our Bayesian models have been fitted using 4 independent MCMC chains with 2000 iterations in each chain. The first 1000 iterations of each chain are taken as the burn-in period. The second half of the MCMC chains have been tested by the variance ratio methods described in section 4.5.2. The table of variance ratios and the plots for the second-half MCMC sequences of models M.a-M.e are shown in table B.1 and figures B.1-B.5 in appendix B. Table 6.2 shows the means and standard deviations of the posterior distributions of parameters. Histograms of the MCMC samples in figures (B.6)-(B.10) in appendix B indicate the shape of the marginal distributions of parameters. In our simulation, the regression coefficients β had actually been generated, but the samples of β were only used in each iteration then discarded due to storage limitations. Since the full conditional distribution of β is a multi-normal distribution, the Rao-Blackwellised estimates for the density function of β can be obtained easily without using the MCMC samples of β (see section 6.4). According to the MCMC sequence plots in figures B.1-B.5, a, b, τ, ϕ_1, ρ, and K generally converge very well, while θ, Γ, and $\Lambda_{\eta\eta}$ do not. The sequence plots of Γ and $\Lambda_{\eta\eta}$ have very similar behaviour. Figure 6.2 shows the pairwise scatter plots of simulated samples of K, θ, Γ, and $\Lambda_{\eta\eta}$ for model M.b, which provides further evidence that θ, Γ, and $\Lambda_{\eta\eta}$ strongly depend on each other. Such posterior dependency of parameters causes the problems of slow convergence in MCMC in the five models. As we can see, MCMC for the parameters related to the covariance matrix of X only are very stable, while the unstable parameters are related to Y.

The predictive performance of five Bayesian models is evaluated by the MSEP and summarised in table 6.3. Performance of PCR is also included in the table as a reference. The PCR model for evaluating each block of validation data contains a subset of principal components (PC's) selected from a set of candidate PC's (the first 10 PC's) using the stepwise function (with default setting) in S-

Table 6.2: MCMC sample means and s.d. (in parentheses) of parameters

M.a				
block	a	b		ρ
1	-9.4737(0.0262)	-0.0011(0.0005)		-0.0076(0.0056)
2	-9.4323(0.0254)	-0.0015(0.0006)		-0.0078(0.0059)
3	-9.4483(0.0260)	-0.0015(0.0006)		-0.0078(0.0058)
4	-9.5084(0.0257)	-0.0007(0.0005)		-0.0078(0.0059)
5	-9.4291(0.0260)	-0.0016(0.0006)		-0.0077(0.0058)
block	$\Lambda_{\eta\eta}$	K	$\theta/100$	Γ
1	0.1163(0.0397)	0.3057(0.2750)	9.7441(6.1248)	0.0001(0.0021)
2	0.1128(0.0365)	0.3155(0.2936)	10.8452(6.7503)	0.0001(0.0001)
3	0.1193(0.0343)	0.3573(0.3326)	10.9235(6.8845)	0.0001(0.0001)
4	0.1225(0.0330)	0.2841(0.2690)	8.3899(5.6273)	0.0002(0.0001)
5	0.1082(0.0329)	0.3695(0.3332)	13.3549(7.0861)	0.0001(0.0001)

M.b				
block	a	b	τ	ρ
1	-5.4955(0.0260)	-0.0009(0.0005)	0.9998(0.0000)	-0.0821(0.0615)
2	-5.4463(0.0265)	-0.0010(0.0005)	0.9998(0.0000)	-0.0852(0.0617)
3	-5.4585(0.0261)	-0.0012(0.0005)	0.9998(0.0000)	-0.0861(0.0630)
4	-5.5234(0.0259)	-0.0005(0.0003)	0.9998(0.0000)	-0.0817(0.0591)
5	-5.4651(0.0261)	-0.0014(0.0005)	0.9998(0.0000)	-0.0885(0.0644)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	0.6064(0.1644)	0.2310(0.2260)	28.1638(22.6438)	0.0038(0.0021)
2	0.6123(0.1709)	0.2483(0.2545)	30.2323(28.6466)	0.0039(0.0023)
3	0.6061(0.1630)	0.3037(0.3156)	45.0840(82.8540)	0.0038(0.0020)
4	0.5944(0.1363)	0.2442(0.2444)	30.3446(29.6986)	0.0036(0.0016)
5	0.5279(0.1665)	0.3344(0.3322)	56.6378(54.5472)	0.0030(0.0020)

M.c				
block	a	b	ϕ_1	ρ
1	-2.0240(0.0241)	-0.0012(0.0003)	1.9838(0.0001)	-0.1774(0.1064)
2	-1.9770(0.0240)	-0.0013(0.0003)	1.9838(0.0001)	-0.2173(0.1295)
3	-1.9872(0.0241)	-0.0015(0.0003)	1.9838(0.0001)	-0.2172(0.1189)
4	-2.0297(0.0243)	-0.0008(0.0003)	1.9838(0.0001)	-0.1676(0.1052)
5	-1.9942(0.0240)	-0.0013(0.0003)	1.9838(0.0001)	-0.2836(0.1310)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	1.6418(0.3436)	0.2672(0.3468)	4.9483(6.6422)	0.0264(0.0103)
2	1.4715(0.3567)	0.3992(0.3433)	9.6249(10.5512)	0.0212(0.0103)
3	1.7163(0.4030)	0.4056(0.5149)	8.2539(14.9839)	0.0287(0.0128)
4	1.6212(0.3329)	0.3135(0.3685)	5.4936(6.0620)	0.0259(0.0106)
5	1.4621(0.3336)	0.4358(0.4822)	10.9251(11.9974)	0.0200(0.0093)

Table 6.2 (*Continued*)

M.d				
block	a			ρ
1	-6.7427(0.0272)			-0.0107(0.0139)
2	-6.7187(0.0264)			-0.0058(0.0145)
3	-6.7341(0.0272)			-0.0080(0.0135)
4	-6.8027(0.0270)			-0.0140(0.0139)
5	–6.6980(0.0268)			-0.0154(0.0147)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	0.5369(0.1479)	0.2384(0.2266)	38.5081(44.0036)	0.0031(0.0017)
2	0.4790(0.1320)	0.2636(0.2558)	51.5324(53.2392)	0.0025(0.0014)
3	0.5631(0.1535)	0.2974(0.2813)	39.7652(31.1963)	0.0035(0.0019)
4	0.5110(0.1226)	0.2420(0.2418)	39.6331(44.9819)	0.0028(0.0013)
5	0.5049(0.1524)	0.2882(0.3322)	56.3277(58.6502)	0.0028(0.0016)

M.e				
block	a		ϕ_1	ρ
1	-6.0956(0.0353)		1.2826(0.0172)	-0.0613(0.0615)
2	-6.0752(0.0349)		1.2836(0.0172)	-0.0179(0.0618)
3	-6.1019(0.0355)		1.2818(0.0170)	-0.0287(0.0705)
4	-6.1934(0.0367)		1.2446(0.0184)	-0.0317(0.0620)
5	-6.0800(0.0349)		1.2872(0.0171)	-0.0688(0.0622)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	0.9597(0.2402)	0.2890(0.2848)	15.3272(14.1281)	0.0089(0.0044)
2	0.8783(0.2093)	0.2831(0.2930)	16.3790(17.2386)	0.0079(0.0038)
3	0.8783(0.2344)	0.3785(0.3745)	22.6370(19.5167)	0.0078(0.0042)
4	0.8480(0.2279)	0.3114(0.3226)	21.4872(22.0519)	0.0074(0.0037)
5	0.8931(0.2405)	0.3417(0.3381)	21.2276(18.5951)	0.0076(0.0041)

Model Type	Sum of Squared Errors of Prediction					MSEP
	block 1	block 2	block 3	block 4	block 5	
PCR (original)	1.1034	1.6951	0.6786	1.289	0.9438	0.1142
M.a	1.3766	1.4277	0.5106	1.1241	0.9174	0.1071
M.b	1.2662	1.2397	0.5212	0.8595	1.0670	0.0991
M.c	1.6669	2.1867	0.7422	1.3148	2.7161	0.1725
PCR (2^{nd} deriv.)	0.8618	0.9045	0.5088	1.0973	0.7506	0.0841
M.d	1.0034	1.3252	0.4488	0.9423	1.1000	0.0984
M.e	1.1397	1.6944	0.7485	1.3095	1.2514	0.1254

Table 6.3: MSEP of a PCR model and three Bayesian models for the original spectra and a PCR and two Bayesian models for the second derivative spectra.

Plus (MathSoft, Seattle, Version 5.1) with the corresponding training data set. The table shows that M.a is slightly better than PCR and M.b is slightly better than M.a, but the difference is not much, while M.c is surprisingly, worse than PCR, M.a and M.b. For the second derivative data, PCR is a better model than M.d, and M.d is better than M.e. The performance of PCR model varies when the set of candidate PC's changes.

A summary of the posterior distributions of the parameters for model M.a-M.e is given in table 6.2. Suppose ϵ is a parameter in a model, we denote its posterior mean in model π by $\hat{\epsilon}_\pi$. The parameters a, ρ, $\Lambda_{\eta\eta}$, K, θ, Γ exist in every model, while b only exists in M.a, M.b and M.c (models for the original spectra). The hyperparameters τ and ϕ_1 are associated only with the correlation structure of X. Considering the models for the original data, i.e. models M.a-M.c, the posterior distributions of a, ρ, $\Lambda_{\eta\eta}$, θ, and Γ are very different and have obvious trends for different covariance structures such that

$$\begin{aligned} \hat{a}_{M.a} &< \hat{a}_{M.b} < \hat{a}_{M.c}, \\ \hat{\rho}_{M.a} &> \hat{\rho}_{M.b} > \hat{\rho}_{M.c}, \\ \hat{\theta}_{M.a} &> \hat{\theta}_{M.b} > \hat{\theta}_{M.c}, \\ \hat{\Gamma}_{M.a} &< \hat{\Gamma}_{M.b} < \hat{\Gamma}_{M.c} \end{aligned} \tag{6.8}$$

and

$$\widehat{\Lambda}_{\eta\eta\, M.a} < \widehat{\Lambda}_{\eta\eta\, M.b} < \widehat{\Lambda}_{\eta\eta\, M.c}.$$

Figure 6.2: Scatter plots of MCMC samples of K, θ, $\Lambda_{\eta\eta}$ and Γ

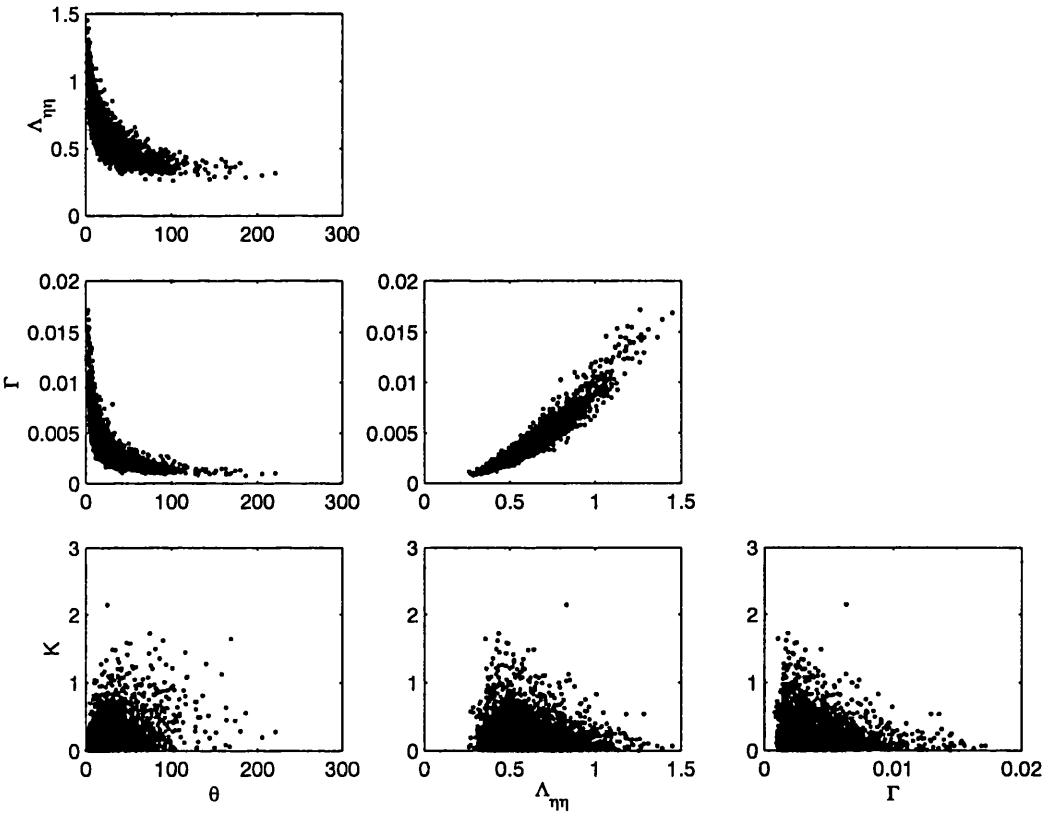
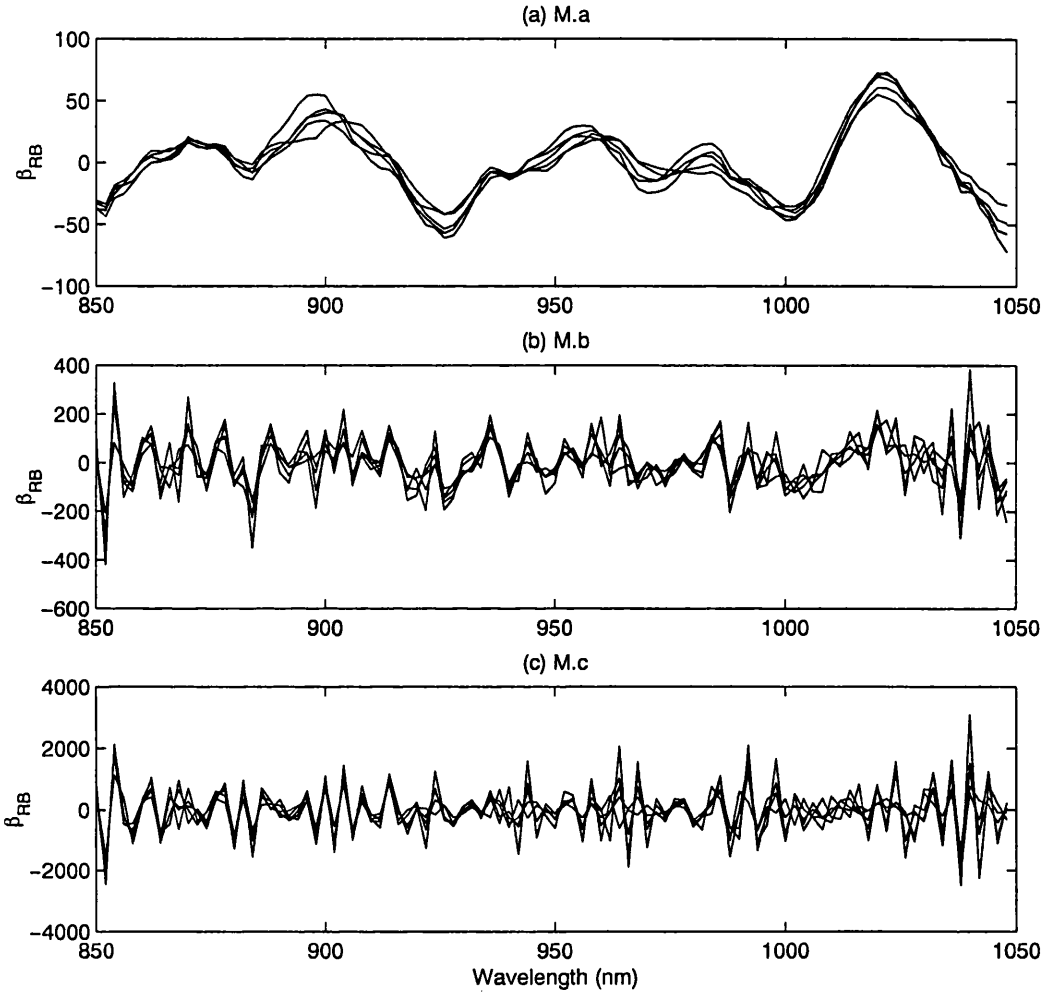


Figure 6.3: Rao-Blackwellised estimates of regression coefficients

For each model, there are five sets of Rao-Blackwellised estimated β because there are 5 blocks in our cross-validation.



The situation for the second derivative spectral data is similar,

$$\begin{aligned} \hat{a}_{M.d} &< \hat{a}_{M.e}, \\ \hat{\rho}_{M.d} &> \hat{\rho}_{M.e}, \\ \hat{\theta}_{M.d} &> \hat{\theta}_{M.e}, \\ \hat{\Gamma}_{M.d} &< \hat{\Gamma}_{M.e} \end{aligned} \tag{6.9}$$

and

$$\widehat{\Lambda}_{\eta\eta\, M.d} < \widehat{\Lambda}_{\eta\eta\, M.e}.$$

Figure 6.3 shows the posterior Rao-Blackwellised expected regression coeffi-

Figure 6.3 (*continued*)

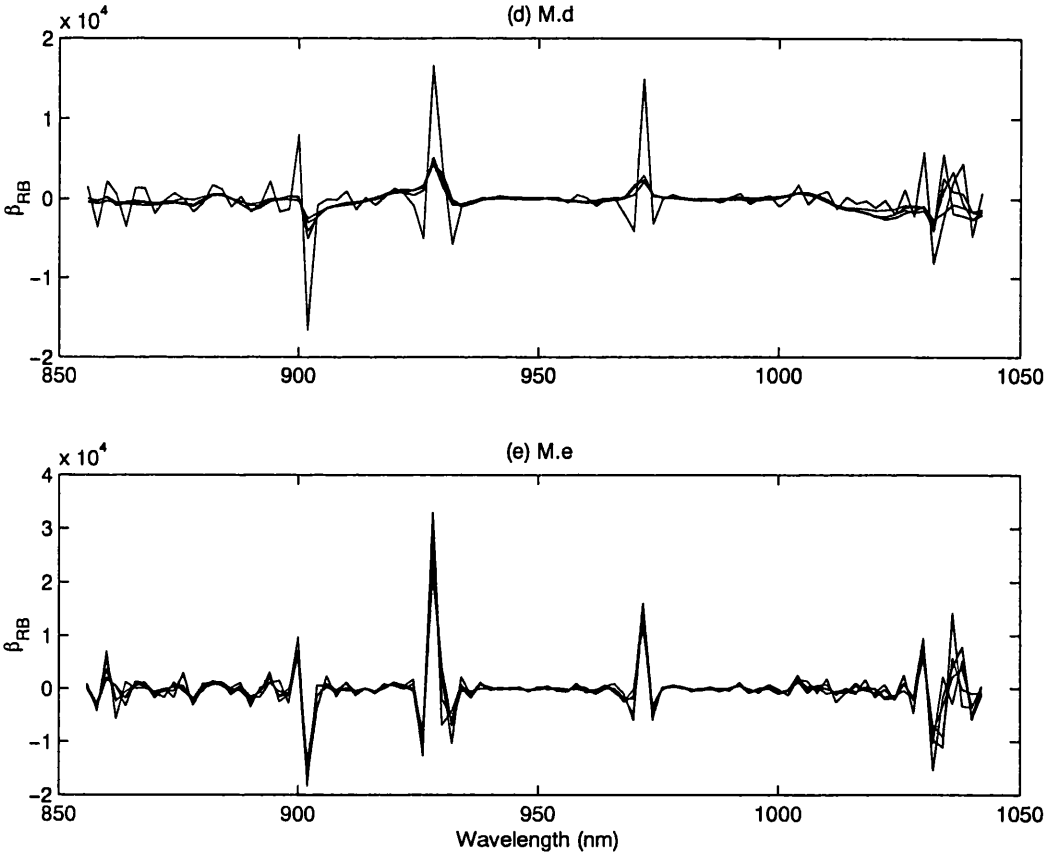
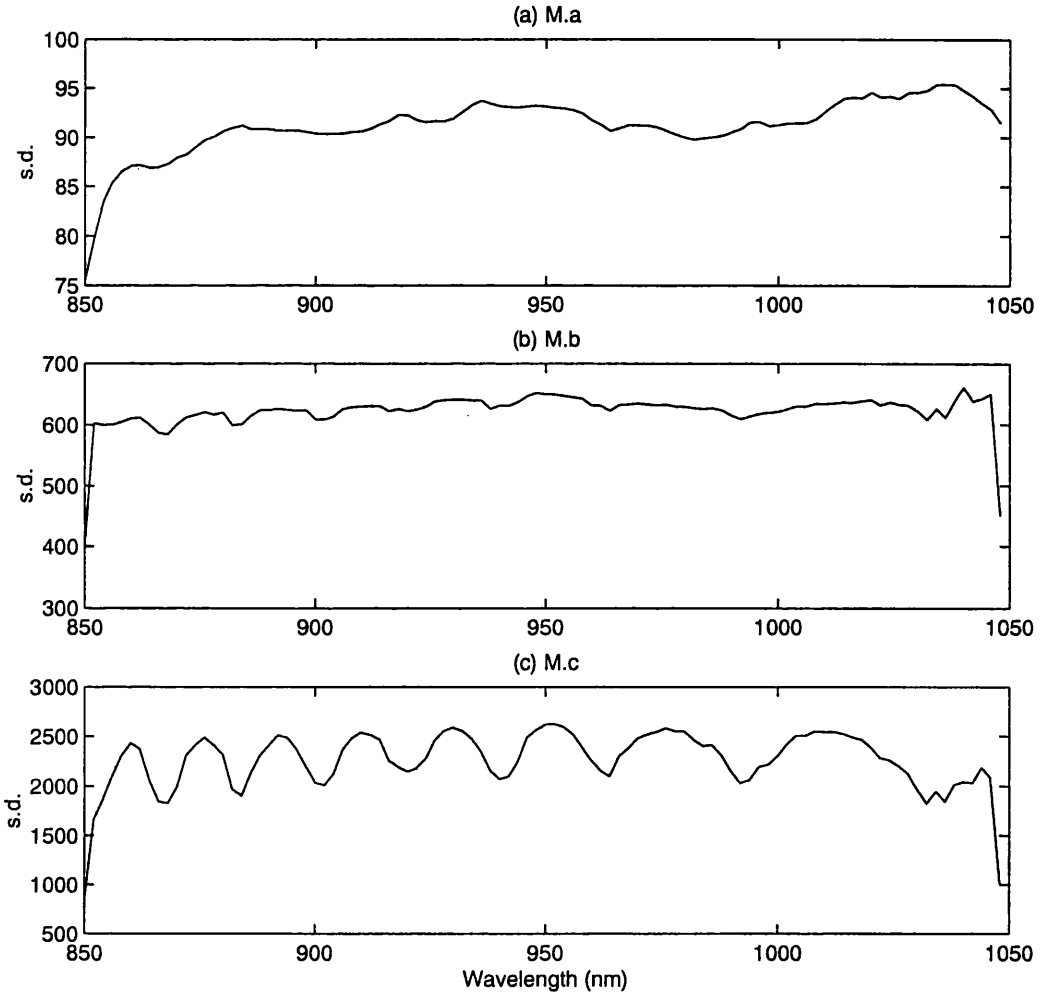


Figure 6.4: Estimated standard deviations of the posterior β



cients of the five models yielded by the 5 blocks of data. According to the graphs, the regression coefficient curves (RCC's) of the five models have different patterns. For the original spectra, the RCC's of M.a are the smoothest ones, and those of M.c are the least smooth. However, M.d and M.e for the second derivative spectra have very similar RCC patterns. The regression coefficients for measurements around wavelengths 900nm, 930nm, 970nm and from 1030nm to 1048nm are spiky for both models.

The covariance matrix of β conditional on data and all other parameters is $D_{xx}^{-1}\Gamma$. Since

$$\mathcal{C}(\beta|Y_t, X_t) = \mathcal{C}[E(\beta|Y_t, X_t)] + E[\mathcal{C}(\beta|Y_t, X_t)],$$

where $\mathcal{C}(\,)$ means the covariance matrix of the random vector in the parentheses, the covariance matrix of β can be estimated again using the Rao-Blackwellised estimate of the density function of β. Figure (6.4) shows the estimated standard deviations of the regression coefficients for M.a, M.b, and M.c on one block of training data. The standard deviations of regression coefficients are generally large. For M.a, M.b and M.c, M.a has the smallest variances for the regression coefficeints, while M.c has the largest among the three.

6.8.3 Model Assessment

Figure 6.5 shows simple model diagnostics. Figure 6.5 (a1), (b1) and (c1) show the observed-fitted value plots on the first block of training data (40 samples) by M.a, M.b and M.c respectively. These 40 fitted values are predicted by a model that is trained by these 40 samples. If a model is seriously overfitted, the points on the plot should line up on the diagonal perfectly of the plot. Figure 6.5 (a2), (b2) and (c2) are the observed-fitted value plots of the validation data by M.a, M.b and M.c respectively. In these graphs, the fitted value of a sample is predicted by a model that is trained by the corresponding block of training data the sample does not belong to. If a regression model makes sense, the 50 data points should scatter around the diagonal. These observed-fitted value plots do not show a significant problem.

Figure 6.5: Model Diagnostics for M.a, M.b and M.c

(a1), (b1) and (c1) show the observed-fitted value plots on one block of training data (40 samples) for M.a, M.b and M.c respectively. These fitted values are predicted by models which are trained by these samples being fitted. (a2), (b2) and (c2) are the observed-fitted value plots of the validation data (50 samples, 10 in each block) by M.a, M.b and M.c respectively.

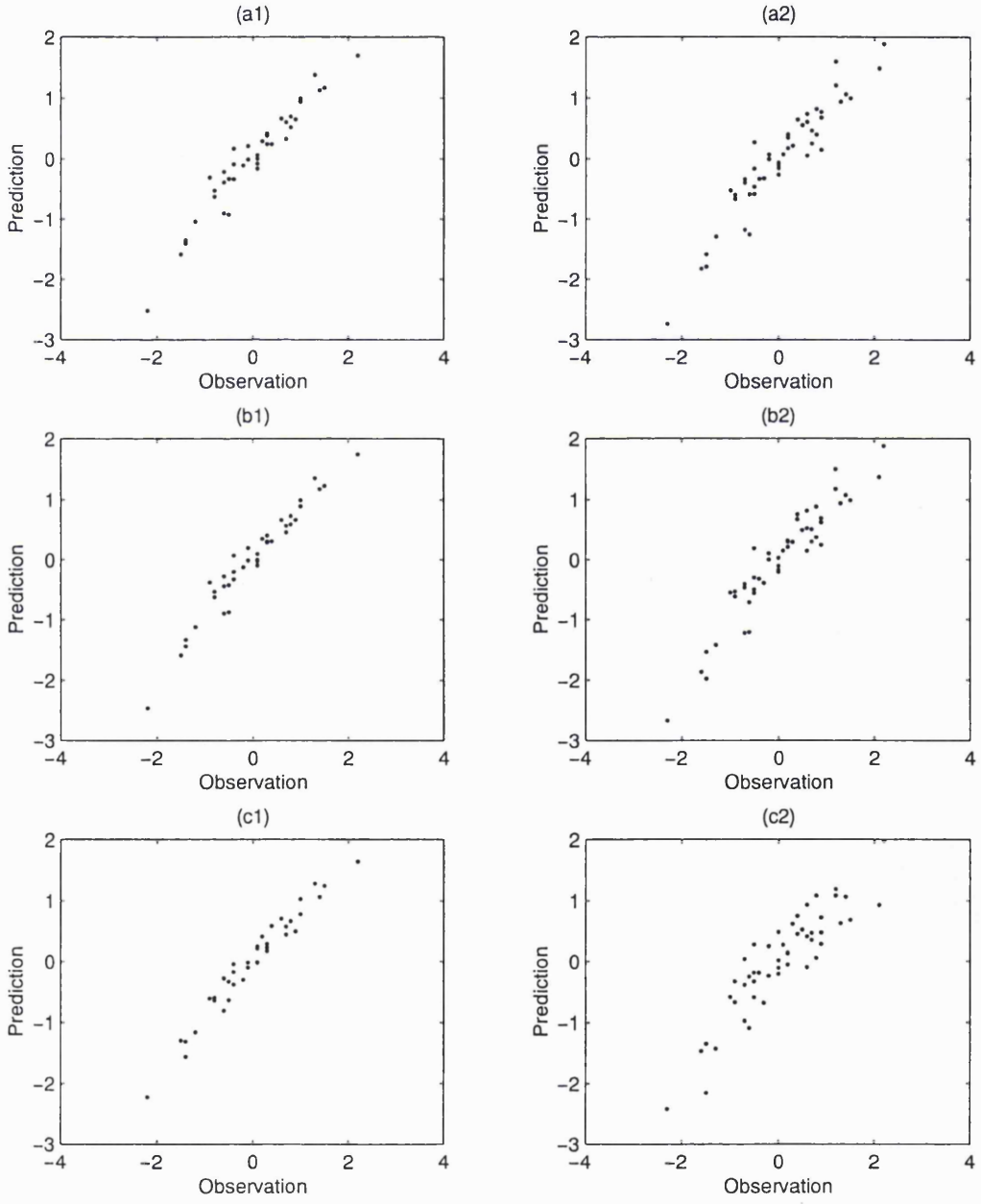


Table 6.4: M.b with fixed τ

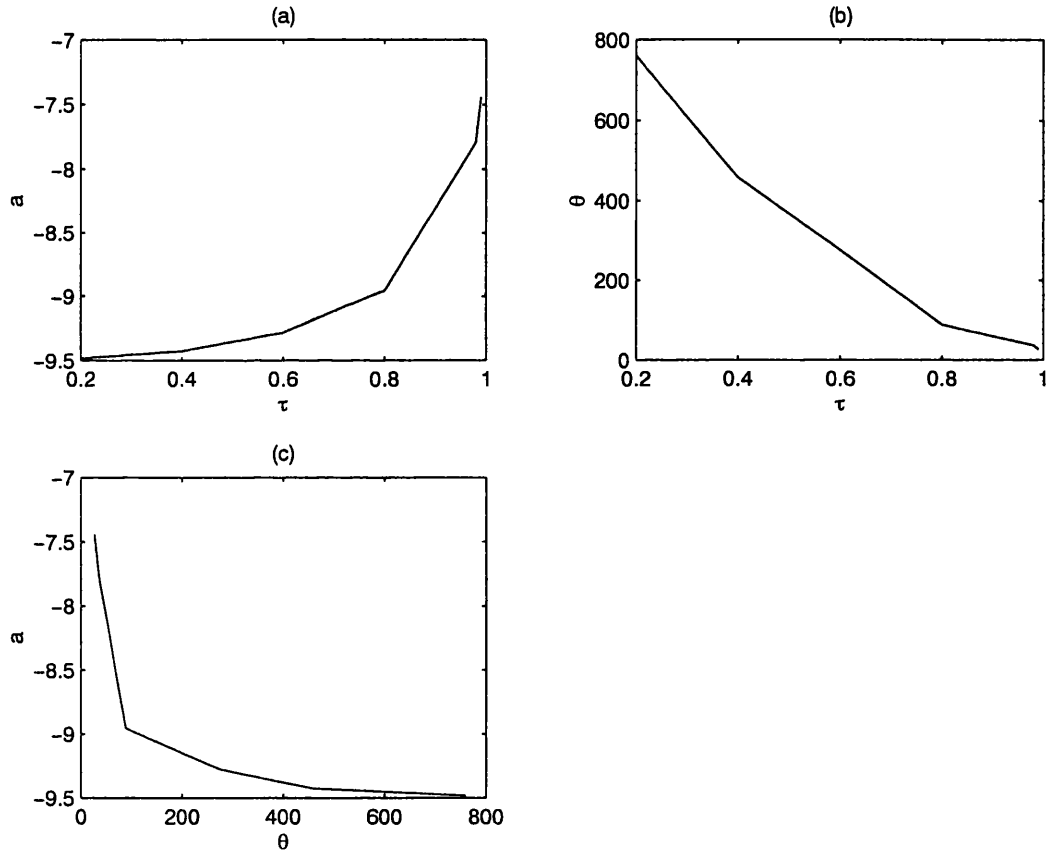
τ	0	0.2	0.4	0.6	0.8	0.98	0.99
MSEP	0.1071	0.1093	0.1082	0.1129	0.1089	0.1044	0.0985

Table 6.5: M.e with fixed ϕ_1

ϕ_1	0	0.5	1	1.5
MSEP	0.0984	0.0991	0.1146	0.1326

In table 6.2, we found that the standard deviation of τ in M.b is very small that the posterior mean is almost a fixed number at 0.9998, which is the upper bound for τ we chose to prevent the MCMC sequences of τ converge to 1. Now, we would like to check whether the posterior model is sensitive to the upper bound we choose or not. We fix τ at several values (however, no greater than 0.9998) and fit the models using MCMC. The MSEP's of these models are given table 6.5. According to the table, MSEP does not have an obvious trend. Again, we fix ϕ_1 for M.e at several values and calculate the MSEP's. The results are given in table 6.5. According to the table, MSEP increases as ϕ_1 increases. According to this analysis, we find that the sensitivity to prior assumptions is not the same for different models. M.a is quite stable given different values of τ, while M.e is sensitive to the value of ϕ_1, with the best prediction being given by $\phi_1 = 0$ (which is simply M.d). Figure 6.6 (a), (b), and (c) show the τ-a, τ-θ and θ-a relationship in M.a given τ on the first block of training data. They show that the distributions of a and θ given τ are strongly dependent on τ. Moreover, θ and a also have a strong correlation.

Figure 6.6: τ-a, τ-θ and θ-a relationships given τ



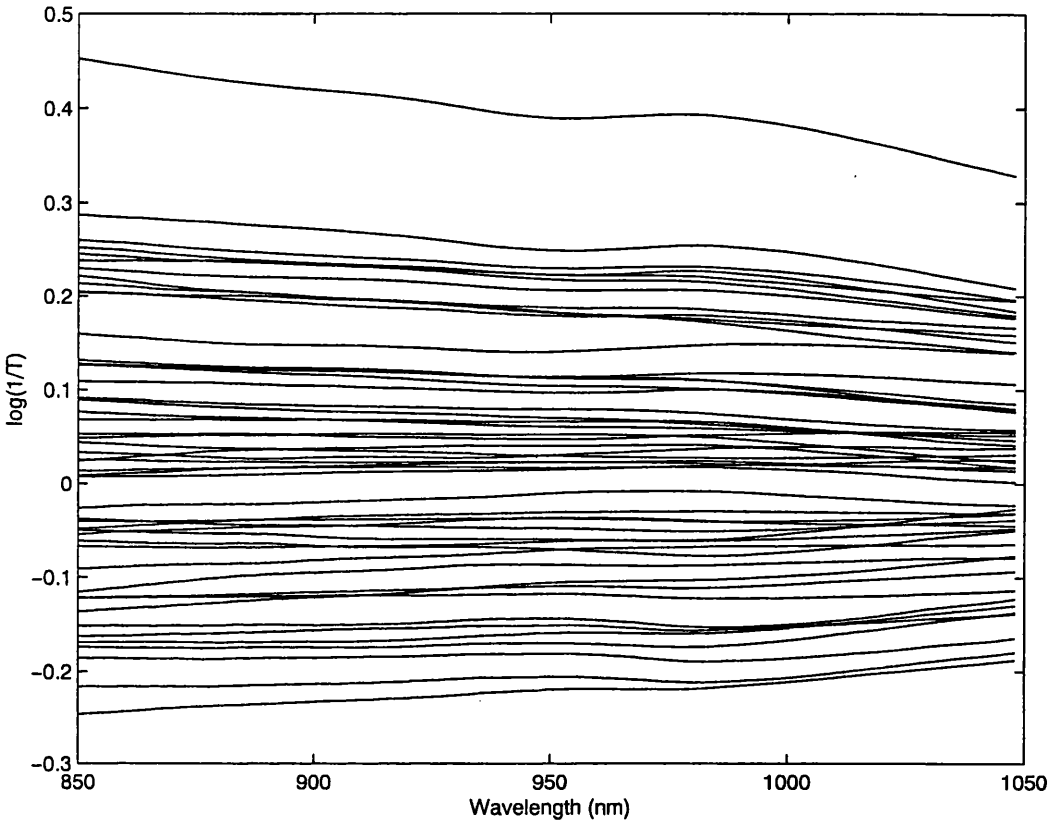


Figure 6.7: 50 centred artificial spectra

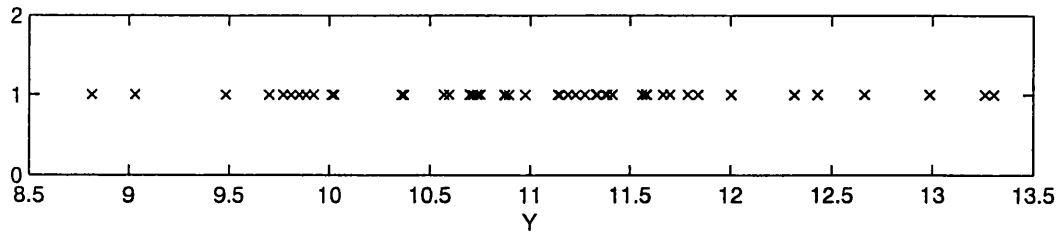


Figure 6.8: Dot plot of the 50 artificial data for Y

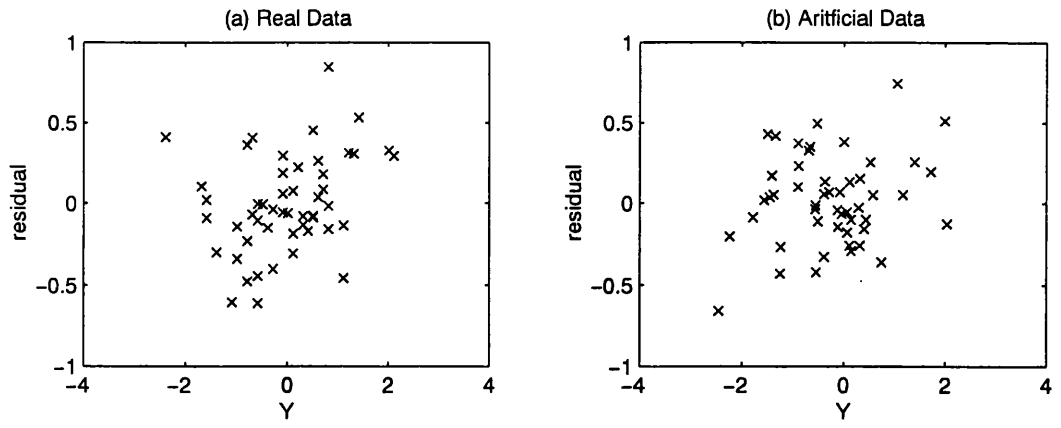


Figure 6.9: Sample Y against residual for the real data in and the artificial data using the given β.

6.9 Modelling with Artificial Data

6.9.1 Artificial Data

According to section 6.8, the hierarchical non-conjugate models do not generally provide better predictions for the real NIR dataset. M.c, which we supposed a priori to be a better model, produced the worst result. We know that none of our models are the correct model for the real data. The way we justisfied M.c to be a better model was rather subjective. In this section, we generate a set of artificial samples in which the spectra are from a process with an AR(2) correlation structure. We use the artificial data to fit our three Bayesian models, and we may see whether M.c with AR(2) correlation structure makes better predictions and whether it estimates the parameters well.

Fifty samples for X (100 explanatory variables) are generated from $\mu + \mathcal{N}(1, \Sigma_{xx})$, where μ is the sample mean of the spectra of the real example, and 50 samples for Y (one response variable) are generated from $X\beta + \mathcal{N}(1, \sigma^2)$, where $\sigma^2 = \Gamma + \Phi$, given $\beta = \beta_{(0)}$, $\Gamma = \Gamma_{(0)}$, $\Phi = \Phi_{(0)}$ and $\Sigma_{xx} = \Sigma_{xx(0)}$, where $\beta_{(0)}$ is in fact the PCR estimate for the regression coefficient vector for the real example we used before, and $(\Gamma_{(0)}, \Sigma_{xx(0)})$ is a random sample pair generated from the prior distributions in equation (6.4), using M.c [AR(2)] model structure. The other

random sample $\Phi_{(0)}$ is generated from its inverse-Wishart prior distribution in equation (6.3).

For hyperparameters, we suppose $a = -2.2$, $b = -0.002$, $\phi_1 = 1.9837$, $\phi_2 = -\phi_1 + 1 - 0.000015$, $\rho = -0.2$, $\Lambda_{\eta\eta} = 1.5$, and $K = .05$. The two degrees of freedom δ and ν are 3 and 2 respectively. The spectra (centered) we generated are shown in figure 6.7 and the values for Y are shown in the dot plot in figure 6.8. Figure 6.9 (a) and (b) shows the residual plots for the real data and the artificial data given the true β respectively. The artificial spectra are visually similar to the natural spectra except there are more cross-overs at 960nm for the natural spectra. The dot plots of Y for the real data and the artificial data are quite similar as well. The pattern of the residuals for the artificial data is not very different from the residual pattern of the real data.

6.9.2 Results

PCR, M.a, M.b, and M.c models were fitted to the artificial data. We applied the procedures used before. For MCMC, 4 independent chains with 2000 iterations were run for M.a and M.b, while 8 independent sequences were run for M.c with the same number of iterations because of slow convergence. Convergence of MCMC has been checked by the variance ratio method (not shown). M.a and M.b do not have strong evidence of lack of convergence, while M.c converges slowly. Appendix C contains the histograms of MCMC samples for all the parameters except β for our three Bayesian models. The histograms of the MCMC samples for the parameters for M.c in figure C.3 show that the posterior marginal distributions of some parameters in M.c have more than one mode. This is the reason for the slow convergence of the MCMC for M.c.

The sample posterior means and the sample standard deviations of the MCMC samples for the parameters are shown in table 6.6. Table 6.7 shows the mean squared errors of predictions for predicting using the original β, PCR, M.a, M.b, and M.c models. According to table 6.7, M.a has the least MSEP among three Bayesian models, while M.c is the worst in predicting Y. Figure 6.10 shows

the original β that generated the samples, and the Rao-Blackwellised estimates for the three Bayesian models for β. The estimate of β in M.a is again the smoothest, while the estimate of β in M.c is the least smooth one. The estimate of β for M.a is the closest to the β that generated the samples. The posterior means of a, ρ, b and $\Lambda_{\eta\eta}$ for M.c are very close to the true values, while K is not. For M.a and M.b, the posterior means for the common hyperparameters in the three models are obviously different from their true values except b is close to the true value in all three models. The posterior distribtuion of ϕ_1 does not necessary have its maximum at the limit of the parameter space. The posterior mean is close to the true value. This suggests that the spectra with the AR(2) structure can provide more effective information in fitting the model with AR(2) correlation structure.

6.10 Discussion

We have incorporated very strong structural belief in our regression models for the real example. For the original spectra, we considered M.c as a much more realistic model, while M.a was simply convenient for computations. For the second derivative spectra, M.e should be a more realistic assumption than M.d. However, the predictive performances of these models on the real example shows that what we preferred in advance does not predict the validation data well. We then fitted M.a, M.b and M.c with a set of artificial data generated from M.c given β and the hyperparameters. In the artificial case, M.c does not have good predictive performance, either. Since we do not know the real model for the real data, we cannot evaluate whether the posterior models of the parameters estimate the true values well. In the artificial data case, we know the true correlation structure for parameters and the true values for the hyperparameters. With the correct correlation structure, most of the parameters can be estimated well except the most important one for regression analysis: β.

By monitoring the detail of the MCMC sampling process, we found that the hyperparameters related to Q_{xx} such as a, b, τ and ϕ_1 are strongly dominated by the information in ${X_t}^t X_t$ alone. That is, the effect of the likelihood of $Y_t|X_t$ is

Table 6.6: MCMC sample means and s.d. (in parentheses) of parameters for the artificial samples

M.a				
block	a	b		ρ
1	-9.3525(0.0312)	-0.0024(0.0008)		-0.0078(0.0058)
2	-9.2938(0.0308)	-0.0019(0.0007)		-0.0078(0.0059)
3	-9.3604(0.0316)	-0.0022(0.0008)		-0.0079(0.0059)
4	-9.3306(0.0303)	-0.0026(0.0008)		-0.0078(0.0059)
5	-9.3549(0.0304)	-0.0018(0.0007)		-0.0078(0.0060)
block	$\Lambda_{\eta\eta}$	K	$\theta/10^5$	$\Gamma \times 10000$
1	0.0826(0.0274)	0.1483(0.1432)	0.0157(0.0080)	0.6653(0.3981)
2	0.0908(0.0232)	0.1230(0.1186)	0.0177(0.0073)	0.8528(0.4735)
3	0.0802(0.0210)	0.1974(0.1812)	0.0192(0.0068)	0.6653(0.3981)
4	0.0941(0.0275)	0.1674(0.1590)	0.0179(0.0074)	0.9341(0.6467)
5	0.0984(0.0334)	0.1698(0.1638)	0.0138(0.0080)	1.0500(0.7690)

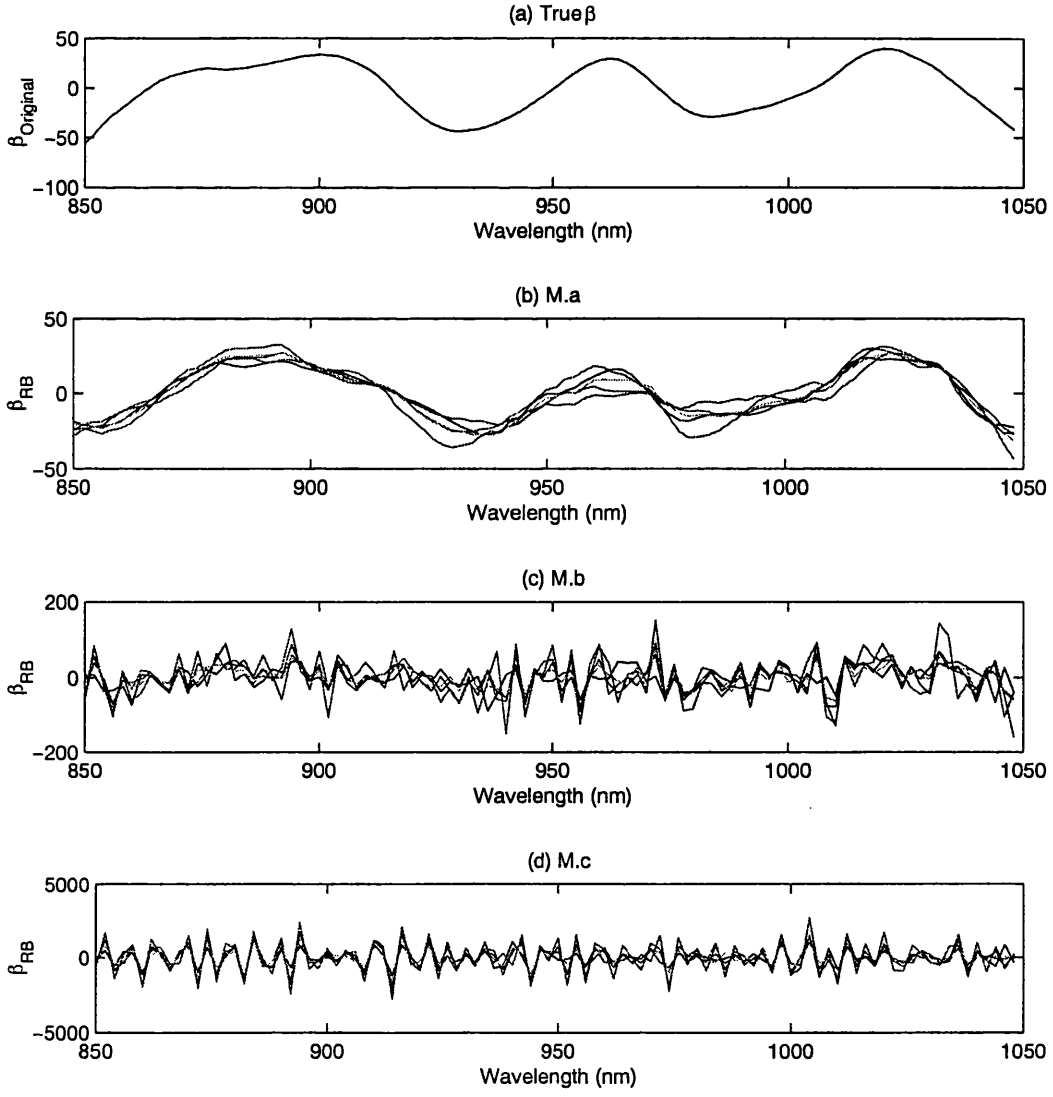
M.b				
block	a	b	τ	ρ
1	-5.4932(0.0295)	-0.0018(0.0006)	0.9998(0.0000)	-0.0931(0.0659)
2	-5.4583(0.0281)	-0.0019(0.0006)	0.9998(0.0000)	-0.0939(0.0655)
3	-5.5025(0.0291)	-0.0018(0.0006)	0.9998(0.0000)	-0.0931(0.0660)
4	-5.4828(0.0280)	-0.0022(0.0006)	0.9998(0.0000)	-0.0988(0.0695)
5	-5.5055(0.0284)	-0.0018(0.0006)	0.9998(0.0000)	-0.0879(0.0625)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	0.3930(0.1405)	0.3229(0.3170)	108.9778(106.5437)	0.0017(0.0013)
2	0.4914(0.1523)	0.4120(0.4032)	82.6660(70.8930)	0.0026(0.0017)
3	0.3864(0.1050)	0.3980(0.3804)	121.6428(104.3103)	0.0016(0.0008)
4	0.4061(0.1426)	0.5249(0.5141)	167.9159(157.1644)	0.0018(0.0013)
5	0.5050(0.1337)	0.3265(0.3230)	60.8695(57.1545)	0.0026(0.0014)

M.c				
block	a	b	ϕ_1	ρ
1	-2.2395(0.0257)	-0.0018(0.0003)	1.9838(0.0001)	-0.4079(0.2414)
2	-2.2285(0.0253)	-0.0018(0.0003)	1.9838(0.0001)	-0.2980(0.1819)
3	-2.2583(0.0257)	-0.0017(0.0003)	1.9837(0.0001)	-0.3844(0.2652)
4	-2.2480(0.0258)	-0.0021(0.0003)	1.9837(0.0001)	-0.4218(0.2458)
5	-2.2506(0.0255)	-0.0018(0.0003)	1.9837(0.0001)	-0.2695(0.1853)
block	$\Lambda_{\eta\eta}$	K	θ	Γ
1	1.1004(0.4179)	0.7874(0.9055)	159.0989(344.6171)	0.0114(0.0089)
2	1.6217(0.4182)	0.7148(0.8310)	20.0492 (41.3737)	0.0244(0.0130)
3	1.3807(0.5094)	0.9672(1.1904)	174.5443(381.1318)	0.0183(0.0136)
4	1.2184(0.3931)	1.2101(1.2709)	141.2490(310.3958)	0.0131(0.0097)
5	1.6600(0.5043)	0.7131(0.9826)	33.2070(104.6666)	0.0268(0.0153)

Model Type	Sum of Squared Errors of Prediction					MSEP
	block 1	block 2	block 3	block 4	block 5	
Correct β	1.4051	0.9099	0.4696	1.1807	0.6236	0.0918
PCR	1.7382	1.0936	1.1063	0.5547	1.3942	0.1177
M.a	2.2548	0.6712	1.5867	0.3681	1.2565	0.1227
M.b	2.7987	0.9825	1.6738	0.6236	1.2353	0.1463
M.c	8.1014	3.0987	2.2136	2.5771	1.7363	0.3545

Table 6.7: MSEP of the models for the artificial data.

Figure 6.10: Original β and the Rao-Blackwellised estimates of β



almost ignorable for these parameters. The values of b are not very different for M.a, M.b, and M.c, while a strongly depends on the assumption we chose for the correlation matrix of X_t. The posterior variances of these four hyperparameters are very small. The posterior variance of ϕ_1 in M.c is small because its parameter space has been limited to a small range, while the parameter spaces of the other three parameters are not particularly limited. Since the posterior variances of them are so small, the posterior densities of other parameters in the random regression are like the posterior results for modelling the controlled regression $Y_t|X_t, \beta, \Gamma, \theta, \rho, K, \Lambda, a, b, \tau, \phi_1$ given very strong priors on a, b, τ and ϕ_1.

The estimate of β should decide whether a model can predict well, and Γ, θ, ρ, K and $\Lambda_{\eta\eta}$ are parameters related to the distribution of the prediction errors. The mean of the full conditional distribution of β is

$$\hat{\beta}_{fc} = [(1+\theta)Q_{xx} + {X_t}^t X_t]^{-1}[(1+\theta)Q_{x\eta} + {X_t}^t Y_t].$$

When $(1+\theta)Q_{xx}$ and $(1+\theta)Q_{x\eta}$ are large, $\hat{\beta}_{fc}$ is close to $Q_{xx}^{-1}Q_{xy}$, the prior mean of β. When $(1+\theta)Q_{xx}$ and $(1+\theta)Q_{x\eta}$ are small, the posterior mean has more weight on data. The prior mean for β somehow has little ability in predicting Y. ${X_t}^t X_t$, ${X_t}^t Y_t$ and the MCMC estimates of $(1+\theta)Q_{xx}$ and $(1+\theta)Q_{x\eta}$ for the five models for the real example are shown in figures E.1 to E.9 in appendix E. We can see that for M.c, which predicts the worst, the mean $(1+\theta)Q_{xx}$ is very close to ${X_t}^t X_t$ and the magnitude of the mean $(1+\theta)Q_{x\eta}$ is even greater than ${X_t}^t Y_t$, while mean $(1+\theta)Q_{xx}$ and mean $(1+\theta)Q_{x\eta}$ are much closer to 0 for the models which predict better. This is mainly due to the value of a. Although θ is also a parameter in the posterior mean of β, a is more influential on $\hat{\beta}_{fc}$. The posterior means for β in M.a for the real example and the artificial data are both very close to the PCR estimates, the estimates which only contain information from data, and provide the best predictive performance among our Bayesian models.

In the analysis for the artificial data with M.c, the posterior density functions of some parameters appear to be bimodal. This does not happen, or is not so obvious for the real data. Since the 50 artificial $Y - X\beta$ are generated from a fixed normal distribution given σ^2, a fixed sum of a random sample of Γ and a random

sample of Φ, the data provide information about σ^2, a mixture of Γ and Φ, i.e. the information for Γ is mixed up with the information for Φ. Therefore, it is reasonable to have the two modes in the the posterior distribution of some parameters. However, this is not so obvious in our real example. This may simply because of the process that generates the natural spectra is not a normal distribution given fixed parameters.

The parameter Φ is the variance of the latent variable α in the non-conjugate model. One may question whether it is necessary to include the non-zero α in the regression model. We have considered the non-conjugate model because it is believed to be a truer model. Whether this model is better than a conjugate model in term of its predictive performance has not been investigated in this thesis.

According to our results for both real data and artificial data, we found that the predictive performance of the models are not good although the assumption for the covariance structure of the variables is closer to the the truth. This is because the posterior density function of β has more information from the prior belief about β, which is an implied result of the prior assumption for the covariance matrix of all the variables. Even though the covariance structural assumption for X is good, the Wishart assumption for Σ does not imply a sensible prior assumption for β and hence the predictive performance is bad. A more flexible prior assumption should be considered. A prior such that the posterior mean of β should not be strongly influenced by the value of a may be considered as the next attempt so that the influence of the covariance structure or extra error may clearly be expressed.

Chapter 7

Bayesian Discrimination with Many Variables

7.1 Introduction

Quantitative methods for allocating an object to one of several categories (groups, classes, or populations) on the basis of an observed feature vector are sometimes referred to as discriminant analysis. Some characteristics of an object are to be observed quantitatively, and in one formulation of the problem they are considered as random variables with different probability distributions in different populations. Distinguishing between normal populations is probably the most frequently considered case in discriminant analysis. In this chapter, we investigate the case of allocating a sample to one of several high dimensional multivariate normal populations using a Bayesian approach.

The distributions of these populations are usually unknown but can be learnt from observed data whose population identities are known. For a multivariate normal distribution, there are two parameters: mean and covariance matrix, usually denoted as μ_i and Σ_i respectively in population i. The main problem in discriminant analysis with many variables is similar to that of regression analysis with many variables. In the classical approach, the number of items observed in each population needs to be greater then the number of variables; otherwise, the

MLS or ML estimates of Σ_i are singular and the population i has a degenerate estimated density function. In the Bayesian approach, the number of observations is not a barrier for model inference if we have proper prior distributions. However, care must be taken in setting up a model when the number of observations is smaller than the number of variables. When the number of training samples is much greater the the number of variables, the concept of 'let the data speak for themselves' is usually implemented by assigning non-informative prior density functions to the unknown random quantities in the model. These non-informative prior density functions are frequently improper. The use of improper priors very possibly yields improper posterior density functions for these parameters. When a model is complicated, it is usually difficult or even impossible to check whether the posterior density functions are well-defined or not. Therefore, one should avoid the use of improper prior distributions because using proper priors automatically prevents the possibility of producing improper posterior results. The natural conjugate priors for the means and the covariance matrices of the multivariate normal distributions are frequently used proper priors. By making the prior more diffuse, the information contained in the proper prior is reduced.

We again follow Brown [23]'s framework for modelling multivariate normal random vectors and assume that the expected covariance matrices are matrix functions with a small number of hyperparameters so that the unknown quantities in the models can be limited to a small number. When we do not have deterministic information about these hyperparameters or we do not want to make deterministic assumptions about them, a hierarchical structural model can be applied by assigning hyper prior distributions to the hyperparameters.

In this chapter, several types of covariance structures which have been considered in chapter 5 and chapter 6 will again be considered in the context of discriminant analysis. We apply our modelling framework to an NIR example. Brown *et al.* [24] have considered using the identity and AR(1) correlation structures with equal expected variances for each variable given the hyperparameters. We consider the hierarchical model with more complex correlation structures than

in Brown *et al.* [24]. We also allow the expected variance to be different for every variable. Posterior distributions of the parameters are estimated using MCMC samples. The same simulation and numerical techniques that have been applied in regression analysis will again be applied in this chapter.

Consider the problem of discrimination with g finite populations labelled from 1 to g. Let Π be an indicator variable which indicates which population a sample comes from. Suppose Z is the measurement on the sample we want to classify, and Z is q-variate. If the object belongs to the i^{th} population, the probability of the sample being observed to have value Z is $p(Z|\Pi = i)$. Suppose we have a prior belief that the probability of any sample being from the i^{th} population is $p(\Pi = i)$ or π_i. According to Bayes formula, the probability of the sample being in the i^{th} popuation after being observed as Z is

$$p(\Pi = i|Z) = \frac{\pi_i p(Z|\Pi = i)}{\sum_{j=1}^{g} \pi_j p(Z|\Pi = j)}. \tag{7.1}$$

We call $p(\Pi = i|Z)$ the ith group membership probability (GMP) of Z, and $p(\Pi = 1|Z), p(\Pi = 2|Z), \ldots, p(\Pi = g|Z)$ form the group membership distribution. Suppose the cost of misclassification is equal for all types of errors. According to Bayes procedure, a sample should be assigned to the group with the largest group membership probability. Thus, the risk of misclassification is minimised (see Anderson [2]). In Bayesian discrimination, the Bayes formula has to be applied again in learning the posterior distributions of the parameters in the sampling distribution of a population. For each i, let Y_i be a matrix of training data such that each row of Y_i is an independent observation from the i^{th} population. The predictive probability of Z given $\Pi = i$ and Y_i is $p(Z|\Pi = i, Y_i)$, and the posterior GMP of Z given $\Pi = i$ is

$$p(\Pi = i|Z, Y_i) = \frac{\pi_i p(Z|\Pi = i, Y_i)}{\sum_{j=1}^{g} \pi_j p(Z|\Pi = j, Y_j)}. \tag{7.2}$$

The following notation will be useful for Bayesian inference. Suppose the size of Y_i is n_i by q for $i = 1, \ldots, g$. Let $\overline{Y}_i$ (1 by q) be the sample mean of the observations from the i^{th} population. Define $S_i = (Y_i - 1_{n_i}\overline{Y}_i)^t(Y_i - 1_{n_i}\overline{Y}_i)$, where 1_{n_i} is a n_i by 1 column vector of 1's and and S_i is q by q. Define $S = \sum_{j=i}^{g} S_j$, which

is also q by q. If the mean of each group is known and denoted as $\mu_1, \mu_2, \ldots, \mu_g$ (all 1 by q), then let $T_i = (Y_i - 1_{n_i}\mu_i)^t \times (Y_i - 1_{n_i}\mu_i)$ and $T = \sum_{j=1}^{g} T_j$, where T_i and T are q by q matrices. Let $Y = \{Y_1, \ldots, Y_g\}$.

7.2 Normal Populations

Suppose the mean of the j^{th} population is $\mu_j(1 \times q)$, and the covariance is $\Sigma_j(q \times q)$. With Dawid's notation, we denote the distribution of Y_j as

$$Y_j \sim 1_j\mu_j + \mathcal{N}(I_{n_j}, \Sigma_j) \tag{7.3}$$

for every j, where $1_{n_j}(n_j \times 1)$ is a column vector of ones, I_{n_j} is an n_j^{th} order identity matrix.

Since q is large, the use of improper prior distributions for μ_j and Σ_j selected by formal rules is very likely to result in improper posterior distributions for parameters. Therefore, we assume the natural conjugate prior distributions for μ_j and Σ_j, which are

$$\mu_j \sim m_j + \mathcal{N}(h_j^{-1}, \Sigma_j), \tag{7.4}$$

$$\Sigma_j \sim \mathcal{IW}(\delta_j; Q_j), \tag{7.5}$$

where m_j is 1 by q, $h_j > 0$ is 1 by 1, $\delta_j > 0$ is 1 by 1, and $Q_j > 0$ is q by q.

It is frequently the case that not much information is available about Σ_j. However, in many situations it is reasonable to assume that the Σ_j are exchangeable, i.e. $\delta_j = \delta$ and $Q_j = Q$ for all possible j. Thus, the number of parameters can be reduced. Another even stronger assumption frequently made is that the covariance matrix of each population is equal, i.e. $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$ and $h_1 = h_2 = \cdots = h_g = h$. As a result, the only difference between the g groups is their means m_j. The number of parameters required by the model with a common group covariance matrix is greatly reduced, since each covariance matrix is a very large matrix when the number of variables is large.

The multivariate normal populations model is one of the most frequently used models. The prior-posterior derivation of this family of models can easily be

achieved due to the conjugate prior assumption, and its applications can easily be found in the literature. In this thesis, we further consider the hierarchical model based on the conjugate model. The following section contains a brief review of some important posterior results of the conjugate model.

In the previous chapter, we have used a non-conjugate sampling model to avoid the deterministic property for infinite regression proved by Dawid [43]. Fang and Dawid [44] also show that as the number of variables tends to infinity, the probability of correct discrimination tends to 1 under a non-hierarchical normal-inverse-Wishart model assumption for observations. This is somehow not realistic in practice. They suggest that in such cases one should either avoid using a normal-inverse-Wishart model or make an assumption so that

$$\gamma_q = (m_i - m_j)Q^{-1}(m_i - m_j)^t < \infty$$

when q (the number of variables) is infinity in the cases with common Σ. We will however continue with the conjugate distribution for the spectra, and have not yet considered whether they satisfy the condition $\gamma_\infty < \infty$ or not. More work has to be done in order to make theoretical non-perfect discrimination.

7.2.1 Unequal Covariance matrices

Consider the case when the the covariance matrices of different populations are exchangeable and the prior distributions of μ_j and Σ_j are

$$\begin{aligned} \mu_j &\sim m_j + \mathcal{N}(h_j^{-1}, \Sigma_j), \\ \Sigma_j &\sim \mathcal{IW}(\delta; Q). \end{aligned}$$

Let $\Theta = \{(\mu_j, \Sigma_j)|j = 1, \cdots g\}$. The posterior joint density function of Θ is

$$\begin{aligned}
& p(\Theta|Y, m_1, \ldots, m_g, h_1, \ldots, h_g, \delta, Q) \\
\propto & \prod_{j=1}^{g} p(Y_j|\mu_j, \Sigma_j) p(\mu_j|m_j, h_j, \Sigma_j) p(\Sigma_j|\delta, Q) \\
\propto & \prod_{j=1}^{g} |\Sigma_j|^{-\frac{n_j}{2}} \exp -\frac{1}{2}\mathrm{tr}\Sigma_j^{-1}(Y_j - 1_{n_j}\mu_j)^t(Y_j - 1_{n_j}\mu_j) \\
& \times \prod_{j=1}^{g} |\Sigma_j|^{-\frac{1}{2}} \exp -\frac{1}{2}\mathrm{tr}\Sigma_j^{-1}(\mu_j - m_j)^t h_j(\mu_j - m_j) \\
& \times \prod_{j=1}^{g} |\Sigma_j|^{-\frac{\delta+2q}{2}} |Q|^{\frac{\delta+q-1}{2}} \exp -\frac{1}{2}\Sigma_j^{-1}Q \\
= & \prod_{j=1}^{g} |\Sigma_j|^{-\frac{\delta+n_j+2q}{2}} |Q|^{\frac{\delta+q-1}{2}} \exp -\frac{1}{2}\mathrm{tr}\Sigma_j^{-1}(Q + S_j) \\
& \times \prod_{j=1}^{g} |\Sigma_j|^{-\frac{1}{2}} \exp -\frac{1}{2}\mathrm{tr}\Sigma_j^{-1}(\mu_j - \hat{\mu_j})^t(n_j + h_j)(\mu_j - \hat{\mu_j})
\end{aligned}$$

where

$$\hat{\mu_j} = \frac{n_j\overline{Y}_j + h_j m_j}{n_j + h_j},$$

Given Q, δ, h_j, and m_j, the posterior group means are conditionally independent of each other given the group covariance matrices, and the posterior group covariance matrices are also independent of each other. Therefore, the posterior distributions of μ_j and Σ_j can be written separately

$$\mu_j|Y, \Sigma_j \sim \hat{\mu_j} + \mathcal{N}\left(\frac{1}{n_j + h_j}, \Sigma_j\right) \tag{7.6}$$

$$\Sigma_j|Y \sim \mathcal{IW}(\delta_j^*; Q_j^*), \tag{7.7}$$

where $\delta_j^* = \delta + n_j$ and $Q_j^* = Q + S_j$. The posterior mean of μ_j given Σ_j is a weighted average of the sample mean and the prior mean, with weights n_j and h_j respectively.

Now consider $h_j \to 0$ for every j, leading to a vague prior distribution for each μ_j. Then,

$$\lim_{h_j \to 0} \hat{\mu_j} = \overline{Y}_j,$$

and (7.6) becomes

$$\mu_j - \overline{Y}_j|\Sigma_j \sim \mathcal{N}(n_j^{-1}, \Sigma_j),$$

and the posterior distribution of Σ_j remains unchanged. The joint posterior distribution of Θ is therefore a properly defined distribution, although the prior distributions of μ_j are improper as $h_j \to 0$. When $h_j \to 0$, the posterior mean of μ_j is the sample mean whatever m_j is.

Suppose Z is a future observation from the j^{th} population, then given (μ_j, Σ_j),

$$Z - \mu_j | \Sigma_j \sim \mathcal{N}(1, \Sigma_j). \tag{7.8}$$

Summing equation (7.6) and equation (7.8) results in

$$Z - \hat{\mu}_j | \Sigma_j \sim \mathcal{N}\left(1 + \frac{1}{n_j + h_j}, \Sigma_j\right).$$

The conditioning on Σ_j is removed using the posterior distribution of Σ_j (7.7). Thus, the predictive distribution of Z when it is from the jth population is a matrix-t distribution

$$Z | (\Pi = j, Y) \sim \hat{\mu}_j + \mathcal{T}(\delta_j^*; a_j, Q_j^*), \tag{7.9}$$

where $a_j = 1 + 1/(n_j + h_j)$, and the posterior probability of Z being in group j is

$$p(\Pi = j | Z, Y) \propto \pi_j g_j^* a_j^{\delta_j^*/2} |Q_j^*|^{-\frac{1}{2}} \{a_j + (Z - \hat{\mu}_j) Q_j^{*-1} (Z - \hat{\mu}_j)^t\}^{-(\delta_j^* + q)/2}$$

where $g_j^* = \Gamma((\delta_j^* + q)/2)/\Gamma(\delta_j^*/2)$.

Consider $h_j \to 0$, the predictive distribution becomes

$$Z | (\Pi = j, Y) \sim \overline{Y}_j + \mathcal{T}(\delta_j^*; a_j^*, Q_j^*), \tag{7.10}$$

where $a_j^* = 1 + 1/n_j$ and the posterior GMP of Z being in group j is

$$p(\Pi = j | Z, Y) \propto \pi_j g_j^* a_j^{*\delta_j^*/2} |Q_j^*|^{-\frac{1}{2}} \{a_j^* + (Z - \overline{Y}_j) Q_j^{*-1} (Z - \overline{Y}_j)^t\}^{-(\delta_j^* + q)/2}.$$

These results can also be found in Brown [23] and Brown *et al.* [24].

Suppose the mean of each population μ_j is actually known. Let $\Theta = \{\Sigma_i | i = 1 \ldots g\}$, the posterior density function of Θ is

$$\begin{aligned} p(\Theta | Y) \propto & \quad \prod_j \quad |\Sigma_j|^{-\frac{n_j}{2}} \exp -\frac{1}{2} \mathrm{tr} \Sigma_j^{-1} (Y_j - 1_j \mu_j)^t (Y_j - 1_j \mu_j) \\ & \times \prod_j \quad |\Sigma_j|^{-\frac{\delta + 2q}{2}} |Q|^{\frac{\delta + q - 1}{2}} \exp -\frac{1}{2} \Sigma_j^{-1} Q \\ = & \quad \prod_j \quad |\Sigma_j|^{-\frac{\delta + n_j + 2q}{2}} |Q|^{\frac{\delta + q - 1}{2}} \exp -\frac{1}{2} \mathrm{tr} \Sigma_j^{-1} (Q + T_j). \end{aligned}$$

That is,

$$\Sigma_j \sim \mathcal{IW}(\delta_j^*; Q_j^{**}),$$

where $Q_j^{**} = Q + T_j$. The predictive distribution for Z being in the j^{th} population is

$$Z - \mu_j \sim \mathcal{T}(\delta_j^*; 1, Q_j^{**}).$$

The variance of Z is less than when it has the predictive distribution (7.9). The posterior GMP for Z being in group j is now

$$p(\Pi = j|Z, Y, \mu_j) \propto \pi_j g_j^* |Q_j^{**}|^{-\frac{1}{2}} \{1 + (Z - \mu_j) Q_j^{**-1} (Z - \mu_j)^t\}^{-(\delta^* + q)/2}.$$

7.2.2 Equal Covariances

Suppose the means of populations are different but the covariance matrices of different populations are the same. The model is denoted as

$$\begin{aligned} Y_j &\sim 1_j \mu_j + \mathcal{N}(I, \Sigma), \\ \mu_j &\sim m_j + \mathcal{N}(h_j^{-1}, \Sigma), \\ \Sigma &\sim \mathcal{IW}(\delta; Q). \end{aligned}$$

Let $\Theta = (\mu_1, \mu_2, \cdots, \mu_g) \cup \Sigma$. The joint posterior density function of Θ given m_j Q, δ and h_j is

$$\begin{aligned} & p(\Theta|Y, Q, \delta, h_1, \ldots, h_g, m_1, \ldots, m_g) \\ = \; & p(\Sigma|\delta, Q) \prod_{j=1}^{g} p(Y_j|\mu_j, \Sigma) p(\mu_j|m_j, h_j, \Sigma) \\ = \; & |\Sigma|^{-\frac{\delta+n+2q}{2}} |Q|^{\frac{\delta+q-1}{2}} \exp\left[-\frac{1}{2}\text{tr}\Sigma^{-1}(Q+S)\right] \\ & \times \prod_{j=1}^{g} |\Sigma|^{-\frac{1}{2}} \exp -\frac{1}{2}\text{tr}\Sigma^{-1}(\mu_j - \hat{\mu_j})^t (n_j + h_j)(\mu_j - \hat{\mu_j}). \end{aligned}$$

That is

$$\begin{aligned} \mu_j - \hat{\mu_j}|Y, \Sigma &\sim \mathcal{N}\left(\frac{1}{n_j + h_j}, \Sigma\right) \\ \Sigma|Y &\sim \mathcal{IW}(\delta + n; Q + S) \end{aligned}$$

The predictive distribution is

$$Z|(\Pi = j, Y_j) \sim \hat{\mu}_j + \mathcal{T}(\delta^*; a_j, Q^*)$$

with $\delta^* = \delta + n$, $a_j = 1 + 1/(n_j + h_j)$, $Q^* = Q + S$, and $n = \sum n_j$.

Now, we assume $h_j \to 0$ so that the prior distribution of μ_j becomes a vague prior. The posterior distribution of μ_j becomes

$$\mu_j - \overline{Y}_j \sim \mathcal{N}(\frac{1}{n_j}, \Sigma).$$

and the predictive distribution is

$$Z|(\Pi = j, Y) \sim \overline{Y}_j + \mathcal{T}(\delta^*; a_j^*, Q^*) \tag{7.11}$$

where $a_j^* = 1 + 1/n_j$. Therefore, the posterior group membership probability of Z being in group j is

$$p(\Pi = j|Z, Y) \propto \pi_j a_j^{*\delta^*/2} |Q^*|^{-\frac{1}{2}} \{a_j^* + (Z - \overline{Y}_j)^t Q^{*-1} (Z - \overline{Y}_j)\}^{-(\delta^*+q)/2}.$$

These results can also be found in Brown [23] and Brown *et al.* [24].

Suppose the mean of each population is actually known, then we have

$$\begin{aligned} p(\Sigma|Y) \propto & \left\{ \prod_j |\Sigma|^{-\frac{n_j}{2}} \exp\left[-\frac{1}{2} \mathrm{tr} \Sigma^{-1} (Y_j - 1_j \mu_j)^t (Y_j - 1_j \mu_j) \right] \right\} \\ & \times |\Sigma|^{-\frac{\delta+2q}{2}} |Q|^{\frac{\delta+q-1}{2}} \exp\left(-\frac{1}{2} \Sigma^{-1} Q \right) \\ = & |\Sigma|^{-\frac{\delta+n+2q}{2}} |Q|^{\frac{\delta+q-1}{2}} \exp\left[-\frac{1}{2} \mathrm{tr} \Sigma^{-1} (Q + T) \right]. \end{aligned}$$

Therefore,

$$\Sigma \sim \mathcal{IW}(\delta^*; Q^{**}),$$

where $Q^{**} = Q + T$. The predictive distribution for Z being in the j^{th} population is

$$Z \sim \mu_j + \mathcal{T}(\delta_j^*; 1, Q^{**}).$$

and the posterior group membership probability for Z being in group j is

$$p(\Pi = j|Z, Y) \propto \pi_j |Q^{**}|^{-\frac{1}{2}} \{1 + (Z - \mu_j) Q^{**-1} (Z - \mu_j)^t\}^{-(\delta^*+q)/2}.$$

7.3 Hierarchical Normal Discrimination

7.3.1 Hyperparameters

Suppose $h_j \to 0$, δ is given, m_j and Q are unknown. Since $h_j \to 0$, the posterior distribution of μ_j and Σ_j are functions independent of m_j. Whatever the distribution of m_j is, the posterior distributions of μ_j and Σ_j remain the same as does the predictive distribution of the sample. Therefore, we only need to consider the distribution of Q. Suppose Q is a function of some parameters θ. The prior distribution of μ_j and Σ_j should be considered as functions of θ instead of functions of Q. It can be shown that the posterior of Σ and Σ_j are inverse-Wishart, therefore, Σ and Σ_j can be marginalised in the posterior models, and the posterior models involve only the parameter θ.

Given θ, the predictive distribution of a future observation is:

1. Unequal Covariance

$$Z|(Y, \Pi = j, \theta) \sim \overline{Y}_j + \mathcal{T}(\delta_j^*; a_j^*, Q_j^*(\theta)).$$

2. Equal Covariance

$$Z|(Y, \Pi = j, \theta) \sim \overline{Y}_j + \mathcal{T}(\delta^*; a_j^*, Q^*(\theta)),$$

where

$$Q_j^*(\theta) = Q(\theta) + S_j, a_j^* = 1 + 1/n_j, \delta^* = \delta + n, Q^*(\theta) = Q(\theta) + S, \text{ and } \delta_j^* = \delta + n_j$$

Both predictive distributions are independent of m_j in the case when $h_j \to 0$.

Suppose the population means are known, the predictive distribution of a future observation given θ is

1. Unequal Covariance

$$Z|(Y, \Pi = j, \theta) \sim \mu_j + \mathcal{T}(\delta_j^*; 1, Q_j^{**}(\theta)).$$

2. Equal Covariance

$$Z|(Y, \Pi = j, \theta) \sim \mu_j + \mathcal{T}(\delta^*; 1, Q^{**}(\theta)),$$

where $Q_j^{**}(\theta) = Q(\theta) + T_j$ and $Q^{**}(\theta) = Q(\theta) + T$.

The predictive density function of Z with the hierarchical structure with non-deterministic θ is

$$\begin{aligned} p(Z|\Pi = j, Y) &= \int_\theta p(Z|Y, \Pi = j, \theta)\, p(\theta|Y)\, d\theta, \\ &= E_{\theta|Y}(p(Z|Y, \Pi = j, \theta)) \end{aligned}$$

where $p(\theta|Y)$ is the posterior density function of θ. The density function of $\theta|Y$ can in general only be estimated using MCMC samples.

7.3.2 Group Membership Probability

The next step is to estimate the posterior group membership density function of Z using the predictive density function of Z. The probability of an item with observation Z being in the i^{th} group is

$$p(\Pi = i|Z) = \frac{\pi_i p(Z|\Pi = i)}{\sum_{j=1}^{g} \pi_j p(Z|\Pi = j)}.$$

Replacing $p(Z|\Pi = i)$ by $p(Z|\Pi = i, Y)$, the posterior group membership frequency function of Z is

$$p(\Pi = i|Z, Y) = \frac{\pi_i p(Z|\Pi = i, Y)}{\sum_{j=1}^{g} \pi_j p(Z|\Pi = j, Y)} \quad \forall i = 1 \ldots g. \tag{7.12}$$

Since

$$p(Z|\Pi = i, Y) = E_{\theta|Y}\{p(Z|\Pi = i, \theta, Y)\},$$

therefore,

$$p(Z|\Pi = i, Y) = \frac{\pi_i E_{\theta|Y}\{p(Z|\Pi = i, \theta, Y)\}}{\sum_{j=1}^{g} \pi_j E_{\theta|Y}\{p(Z|\Pi = j, \theta, Y)\}}, \quad \forall i = 1 \ldots g.$$

Suppose $\{\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(m)}\}$ is a sequence of MCMC samples generated from the posterior distribution of θ, $E_{\theta|\text{Data}}\{p(Z|\Pi = i, \theta, Y)\}$ can be estimated by its MCMC sample average

$$\frac{1}{m} \sum_{k=1}^{m} p(Z|\Pi = i, \theta^{(k)}, Y)$$

Table 7.1: Parameters in M.a, M.b, and M.c for discrimination analysis

Note: $\Omega = \{\phi_1 | -1 < \phi_2 < 1, \phi_2 + \phi_1 < 1, \phi_2 - \phi_1 < 1\}$

Model	Parameter spaces
M.a	$a \in \mathrm{R}$, $b < 0$
M.b	$a \in \mathrm{R}$, $b < 0$, $0 \leq \tau < 1$
M.c	$a \in \mathrm{R}$, $b < 0$ $\phi_1 \in \{\phi_1 \| \phi_1 \in \Omega, \phi_2 = -\phi_1 + 0.99985\}$

and $p(\Pi = i|Z, Y)$ is then estimated by

$$\hat{p}_i(Z) = \frac{\sum_{k=1}^{m} \pi_i p(Z|\Pi = i, \theta^{(k)}, Y)}{\sum_{j=1}^{g} \sum_{k=1}^{m} \pi_j p(Z|\Pi = j, \theta^{(k)}, Y)}.$$

The most probable group that Z belongs to is then the group with largest $\hat{p}_i(Z)$.

7.4 Example

Consider the example for discriminant analysis we have described in section 3.5.2. The 292 samples of unground wheat had their NIR spectra measured using a Infratec Grain Analyzer which measures transmission through the wheat sample of radiation at 100 wavelengths from 850 to 1048 nm in steps of 2 nm. There are 9 varieties of wheat, and the group identity of each wheat sample is known. These data were split randomly within groups so that 80% of samples are assigned to the training set and remaining 20% of the samples are in the validation set. Table 3.1 shows the number of samples of each variety in the training and validation sets. Figure 3.7 shows the spectra of 234 samples in the training set, while figure 3.8 shows the spectra of the samples in the validation set. The training data and the validation data are identical to the training data and validation data in Fearn *et al.* [56] so that we can compare our result with theirs.

Our analysis is based on the assumption that the population covariance matrices for the 9 varieties are the same, and means are known (using the sample means). The covariance matrix Σ is marginalised, hence the joint posterior density function of the hyperparameter θ with hyper prior $\pi(\theta)$ is

$$|Q(\theta) + T|^{-\frac{\delta+n+q-1}{2}} |Q(\theta)|^{\frac{\delta+q-1}{2}} \pi(\theta). \tag{7.13}$$

Here, $q = 100$, $n = 234$. The same numerical techniques introduced in section 6.5 will be applied.

Derivative spectra have not been considered because derivation eliminates the effects caused by the physical properties that make the difference between different varieties. The models we consider are the M.a, M.b and M.c in chapter 5 and 6 with three different correlation structures, i.e. the identity matrix, AR(1) and AR(2) correlation functions for R (100×100), where $Q = \Lambda R \Lambda$, Λ (100×100) is a diagonal matrix function of scalar parameters a and b and the i^{th} diagonal element is

$$\sqrt{|\mu_i|}\exp[a + b(i - \bar{i})],$$

where μ_i is the overall sample mean of the i^{th} variable. R is an identity matrix in M.a, a matrix function of scalar τ in M.b, and a matrix function of scalar ϕ_1 and ϕ_2 with a constraint $\phi_2 = -\phi_1 + 0.99985$ in M.c (refer to section 5.7). These hyperparameters have the parameter spaces shown in table 7.1. The hyper prior distribution for a is $N(0, 100)$, and the hyper prior density for b is proportional to the density of $N(0, 100)$ restricted to $b < 0$. These priors for a and b are relatively very diffuse and provide little prior information. For τ, we assign a flat prior with parameter space $[0, 0.9998]$ so that the MCMC sequences of τ will not converge to 1 to produce near singular samples of Q_{xx}. The right bound 0.9998 is a bit greater than our prior bound for ρ in section 5.7. The hyper prior density for ϕ_1 is also a flat density. According to our preliminary study, MCMC sequences for the discrimination model M.c converge better than MCMC sequences for the regression M.c. Unlike in the regression analysis, the MCMC sequences for ϕ_1 do not converge to 2. However, using the entire parameter space for ϕ_1 in table 7.1 still causes very slow convergence. Therefore, we use uniform(1.982,1.984), which has a larger domain than the uniform distribution for ϕ_1 we used for regression model M.c and in section 5.7). The full conditional density functions of the parameters in each model can be derived using the joint density function (7.13), by replacing θ with the hyperparameters in the models.

We also vary the shape parameter in the prior distribution of the covari-

ance matrix of the variables. We suppose also that for every $p(\Pi = i)$, the prior probability of any sample coming from the i^{th} population is known. In practice, $p(\Pi = i)$ is estimated by the proportion of the training samples from the i^{th} group in the whole training data set.

7.5 PCA and Logistic Discriminant Analysis

We also provide the results from the linear discriminant analysis with PCA and the logistic discriminant analysis in Fearn *et al.* [56] as reference. In this section, the methods for discriminant analysis using PCA and the logistic discriminant analysis in Fearn *et al.* are briefly introduced.

7.5.1 PCA

PCA has been introduced in section 3.7.4 as a method for compressing data into fewer variables in order to reduce the number of variables in the model. The original training data from 9 groups are $Y_1, Y_2, \ldots, Y_9$, where Y_i is an n_i by q matrix, n_i is the number of training samples from the i^{th} group and q is the number of variables. The PCA score matrix of Y_i is denoted by G_i, which is an n_i by k matrix where k is the number of PC's applied in the models. The PCA is based on total sample covariance, so G_1 represents the PC which has the largest total variance among all PC's. The discriminant analysis is now based on the PCA data $G_1, G_2, \ldots, G_9$. The sample mean of G_i is used to estimate the mean of PCA data from the i^{th} group, and the common covariance matrix of the distributions of the PCA data from the 9 groups is also estimated using the $G_1, G_2, \ldots, G_9$. The distributions of the PCA samples of the 9 groups are then obtained. The 9 by 58 group membership probabilities can then be calculated using the formula (7.1) by plugging in the estimates for the group means and the covariance matrix.

7.5.2 Logistic Discriminant Analysis

Suppose $p_j(Z)$, $j = 1, \ldots, g$, to be the probability of a sample being observed with value Z $(1 \times q)$ in the j^{th} population, and the prior probability of any sample being in the j^{th} group is π_j. Logistic discriminant analysis considers a regression approach to predict the group membership probability of a new observation. According to Bayes formula, the posterior group membership probability of the item with measurement Z is

$$p_j(Z) = \frac{\pi_j p_j(Z)}{\sum_{i=1}^{g} \pi_i p_i(Z)}.$$

Logistic discriminant analysis uses the logistic regression model

$$\log\left[\frac{\pi_j p_j(Z)}{\pi_g p_g(Z)}\right] = \beta_{0j}^0 + Z\beta_j \quad j = 1, \ldots g-1, \tag{7.14}$$

where $\beta_{0j}^0 > 0$ is a scalar and β_j is a q by 1 vector of regression coefficients. Equation (7.14) is equivalent to the equation

$$\log\left[\frac{p_j(Z)}{p_g(Z)}\right] = \beta_{0j} + Z\beta_j \quad j = 1, \ldots g-1, \tag{7.15}$$

where $\beta_{0j} = \beta_{0j}^0 - \log(\pi_j/\pi_g)$ and the g^{th} group has been arbitrarily chosen as a reference. The posterior group membership probability of any future observation being in group j can then be predicted using this logistic regression model trained by the training data.

Fearn *et al.* [56] provided a comparison of two Bayesian logistic discriminant analyses: 1, plugging the Bayes estimates for β_{0j} and β_j which maximise their posterior densities into equation (7.15) to classify new cases; 2, predicting the posterior GMP of a new observation being in each group using the Laplace approximation. The data they used is the same data set we use in this chapter. We use the training and the tuning data set in Fearn *et al.* as the training set in this chapter while the validation data set is the same here and in Fearn *et al.* so that the results are comparable. In [56], the data have been compressed using PCA to reduce the number of variables.

7.6 Results and Diagnostics

Our hierarchical models were fitted using ARMS within Gibbs sampling, with 2000 iterations. The first half of each Markov chain was discarded. Convergence diagnostics for the MCMC and the histograms of the MCMC samples with different parameters in M.a, M.b and M.c are shown in appendix D. Table D.1 shows that the square roots of $\widehat{R}_G$, $\widehat{R}_C$ and $\widehat{R}_{interval}$ (see section 4.5.2) of the parameters of the three models are all very close to 1. Figures D.1, D.2 and D.3 show that the second half of the 4 independent MCMC sequences for each model setting. According to these graphs, there is no obvious evidence to suggest that the 4 MCMC sequences converge to different target distributions. With these variance-ratios and the time-series plots, we accept the use of the MCMC samples in the second half of each chain to estimate our posterior distributions of parameters.

The MCMC estimates for parameters and their standard deviations are shown in table 7.2. The number of correct classifications (C.C.) for each model is presented in table 7.3. According to the posterior joint density function, it is expected that the standard deviations of the hyperparameters will be smaller when δ increases. Denote the estimates of a in M.a, M.b, and M.c as $\widehat{a}_{\text{M.a}}$, $\widehat{a}_{\text{M.b}}$ and $\widehat{a}_{\text{M.c}}$ respectively. From table 7.2, we observe that $\widehat{a}_{\text{M.a}} < \widehat{a}_{\text{M.b}} < \widehat{a}_{\text{M.c}}$. As we have discussed in section 6.10, the level of posterior a depends strongly on the correlation assumption for the variables. According to table 7.3, model M.c, provides the worst prediction among all models, while the correct-classification rates of M.a and M.b are better. However, there is no clear pattern how δ affects the C.C. rates. The histograms of MCMC samples in figure D.4, D.5 and D.6 in appendix D indicate the shape of the marginal distribution of the parameters in the models. They show that the posterior marginal distribution of ϕ_1 in M.c with $\delta = 3$ is very different from the posterior marginal distribution of ϕ_1 in M.c with $\delta = 250$ and $\delta = 500$. According to the figures, the posterior marginal probability density of ϕ_1 increases as ϕ_1 decreases for $\delta = 3$ within the range shown in the figure, while for ϕ_1 with $\delta = 250$ and $\delta = 500$, the probability density decreases quickly as ϕ_1 decreases. The histograms for parameters also show that when δ is

high (the prior for the covariance matrix of the spectra is stronger), the MCMC sample variances of parameters are generally smaller.

One may worry whether the low C.C. rate is due to overfitting or not. In classical analysis, the problem of overfitting frequently exists while modelling with many variables. The numbers for correct classifications on the training data for the three models are shown in table 7.4. They indicate that although the correct classification rates for training data are better than the correct classification rates for validation data, they do not show signs of extreme overfitting. The C.C. rates on training data show that M.b is a slightly better model than M.a, while M.c is slightly worse than M.a.

The calibration of the GMP's for the validation data predicted by several models are summarised in table 7.5. These 9×58 posterior GMP's have been put into bins of width 0.2. The columns labelled with n are the frequency of GMP's in each bin. The columns labelled with c are the actual number of correct groups of the validation data in each bin. The columns labelled with e represent the expected numbers of correct groups in each bin (sum of the GMP's in each bin) for each of the models. A model provides good calibration when c is close to e. According to the table, M.b is generally better than M.a and M.c in our cases, although our three models seem to be over-optimistic because the expected correct groups in high probability bins are much higher than the actual number of correct groups in high probability bins. On the other hand, the expected correct groups in low probability bins are much less than the actual number of correct groups in low probability bins. In general, models with high δ are worse than the models with low δ.

7.7 Remark

Table 7.5 (a) shows that when the number of PC's included in the model increases, there appears to be be more certainty about assignment to groups. However, this seems to be over-optimistic because when the number of PC's increases, the actual correct classification rates are not better, whilst the differences between the ex-

pected correct classifications and the actual number of correct classifications with GMP's between 0.8 and 1 are larger. For a non-hierarchical Bayesian discriminant analysis, Dawid and Fang [44] provided the conclusion for a conjugate normal analysis that when the number of variables increases to infinity, an item is classified to a group with probability one, which might not be appropriate in practice. Although we have not varied the number of variables in our hierarchical model in order to compare models with different number of variables, our practical case result shows with large number of variables the hierarchical model also yields high probabilities for assignment to groups. When the number of variables is fixed, our analysis indicates that when δ increases, the decision tends to be more certain. However, the evidence shows that increase of certainty does not guarantee the increase in C.C rate. When δ is large it means that we are more sure about our prior belief, while our belief may not be that true. We have found that our models did not have the sign of overfitting because their correct classification rates on training data are not really very large. When the posterior a is large, the covariance matrix of the posterior models for 9 groups is greater.

Our Bayesian models with more realistic covariance structure for the variables are not better than the PCA method or Fearn *et al.*'s logistic models. In this example, the sample means of the training data from 9 populations are actually very close and their variances are not small. Many validation spectra cannot be clearly distinguished from the training samples in other populations according to the means of the population distribution. As a result, the within group variance and the autocovariance structure for the spectra in each population may provide more information for discriminating a future observation than the means of population distributions. An assumption that the observations in different populations with different covariance matrices would probably be more appropriate for our data.

Table 7.2: MCMC estimates of parameters and their standard deviations (in parentheses)

M.a			
δ	a	b	
3	-9.2675 (0.0110)	-0.0005 (0.0003)	
250	-8.1853 (0.0081)	-0.0021 (0.0003)	
500	-7.6396 (0.0081)	-0.0028 (0.0003)	
M.b			
δ	a	b	τ
3	-5.0453 (0.0143)	-0.0004 (0.0003)	0.9998 (0.0000)
250	-4.1114 (0.0079)	-0.0023 (0.0002)	0.9998 (0.0000)
500	-3.6799 (0.0071)	-0.0029 (0.0002)	0.9998 (0.0000)
M.c			
δ	a	b	ϕ_1
3	-1.6408 (0.0193)	-0.0007 (0.0002)	1.9829 (0.0006)
250	-0.6305 (0.0075)	-0.0017 (0.0001)	1.9839 (0.0001)
500	-0.2306 (0.0070)	-0.0019 (0.0001)	1.9839 (0.0001)

Table 7.3: Number of correct classifications (C.C.) on 58 validation samples

The left table contains the number of C.C.s using M.a, M.b, and M.c. The middle table shows the value of C.C.s using the linear discriminant approach with different numbers of PC's as variables. The right table shows the C.C. of the logistic model in Fearn *et al.*

δ	M.a	M.b	M.c
3	38	38	35
250	38	41	29
500	42	41	29

No of PC's	
10	36
30	41
50	36

Logistic	
plug-in	38
predictive	37

Table 7.4: Number of correct classifications (C.C.) on 234 training samples

δ	M.a	M.b	M.c
3	215	216	213
250	211	212	205
500	204	206	195

Table 7.5: Frequency tables of the GMP's

Frequency tables of the 9×58 group membership probabilities of different models including PCA models with first 10, first 30, and first 50 PC's, the plug-in and the predictive versions of the logistic discriminant analysis in Fearn *et al.*, and our Bayesian hierarchical models M.a, M.b, and M.c with $\delta = 3,\quad 250 \quad$ and $\quad 500$.

(a) PCA models

No of PC's	10			30			50		
π	n	c	e	n	c	e	n	c	e
0.0-0.2	438	15	11.4	455	14	4.1	451	19	2.8
0.2-0.4	30	9	8.4	10	3	2.9	13	2	3.3
0.4-0.6	20	10	10.0	4	1	2.0	5	1	2.6
0.6-0.8	13	7	9.1	7	2	5.0	9	5	6.6
0.8-1.0	21	17	19.1	46	38	44.0	44	31	42.7

(b) Logistic models in Fearn *et al.*

	plug-in			predictive		
π	n	c	e	n	c	e
0.0-0.2	439	17	10.4	439	17	12.2
0.2-0.4	29	4	8.1	30	5	8.9
0.4-0.6	14	6	7.0	15	7	7.2
0.6-0.8	21	15	14.9	23	16	16.9
0.8-1.0	19	16	17.6	15	13	13.7

Table 7.5 (*continued*)

(c) M.a

δ	3			250			500		
π	n	c	e	n	c	e	n	c	e
0.0-0.2	452	19	2.9	455	18	1.8	459	15	1.9
0.2-0.4	12	1	3.5	4	1	1.2	4	1	1.3
0.4-0.6	2	1	1.0	11	3	5.4	5	1	2.5
0.6-0.8	12	4	8.3	5	1	3.6	3	2	2.1
0.8-1.0	44	33	42.3	47	35	46.0	51	39	50.2

(d) M.b

δ	3			250			500		
π	n	c	e	n	c	e	n	c	e
0.0-0.2	450	18	3.3	451	15	2.6	453	15	1.6
0.2-0.4	12	2	3.4	14	2	3.8	10	2	2.9
0.4-0.6	9	2	4.8	5	1	2.6	6	2	3.0
0.6-0.8	13	5	9.5	7	3	5.0	6	2	4.4
0.8-1.0	38	31	37.1	45	37	44.0	47	37	46.1

(e) M.c

δ	3			250			500		
π	n	c	e	n	c	e	n	c	e
0.0-0.2	448	19	5.5	448	22	4.5	449	23	3.8
0.2-0.4	16	4	4.3	17	6	4.9	15	6	4.5
0.4-0.6	10	6	4.9	8	3	4.2	8	3	4.0
0.6-0.8	10	4	7.3	11	3	7.9	8	4	5.8
0.8-1.0	38	25	36.0	38	24	36.5	42	22	39.9

Chapter 8

Summary and Discussion

8.1 Summary

In this thesis, we have considered regression analysis and discriminant analysis with many variables using a Bayesian approach. Other research in multivariate analysis has mainly been focused on variable selection using computationally convenient models, which are usually not realistic. We have attempted to use more realistic prior information rather than using prior assumptions that are favoured for their computational advantage. We used all variables for modelling. When computing was expensive, selecting and compressing information were important issues because it was necessary to reduce the cost of computing. Nowadays, it is possible to store a large amount of information at a low cost and fitting complicated models can be finished in a short period of time. Therefore, investigation of modelling with all variables is more important.

We used the wheat NIR data as examples for our regression analysis and discriminant analysis. Empirically, linear models are able to provide good predictions for most NIR applications. Therefore, we are justified in assuming that the NIR data could be modelled by a linear model. These NIR spectra are usually assumed to be multivariate normally distributed. Based on this framework, we constructed our sampling distribution for our NIR spectral data. One important property of NIR spectra is that the measurements at different wavelengths are

highly correlated, and in fact these spectra are very smooth. That is, the amount of information carried by each single variable is small. Deleting variables causes the loss of important information. In chemometrics, keeping most information in a small number of new variables has been an important way to model such problems.

Three correlation structures: the identity structure, AR(1) correlation function and AR(2) correlation functions were considered for the examples in this thesis. The identity structure is the most common one which has been considered by other authors due to the simplicity in computation. We consider the AR(2)-type model a more realistic model. ARMA-type structures have been suggested by Brown [23] for the property of structural coherence. For regression analysis, we considered a non-conjugate normal model instead of the conjugate normal model which had been proved to have the property of determinism by Dawid [43]. This property is believed to be inappropriate for the NIR calibration model. For discriminant analysis, we simply assumed that the samples in 9 populations were multivariate normally distributed. These models were evaluated mainly by a cross-validation method for regression analysis and validation set for discriminant analysis.

The models were fitted by the MCMC approach. We employed ARMS within Gibbs sampler as our sampling scheme because the Gibbs sampler is generally the best strategy to sample from high dimensional distributions. Since the posterior models are complicated, we ran multiple chains in order to detect whether the Markov chains do not converge to the same target distributions. Variance ratio methods have been used for convergence checking. Extensive graphical presentations have also been produced as an aid to check the output of MCMC.

8.2 Prior Information

Formalising prior information is not easy in multivariate cases. First of all, the real situation is frequently so complicated that it is very difficult to understand the relationship between the variables involved, and there may even be no model that we know which can describe the events properly. Secondly, there are only

a small number of known parametric distribution families. Multivariate normal distributions have been frequently used as sampling distributions in multivariate analysis, even though the real distribution of the multivariate observations may not really be a multivariate normal distribution. Moreover, we have only a few choices for the prior distributions for the mean vectors and the covariance matrices of the multivariate normal distributions. Fortunately, the prior model can be updated gradually by the data to a posterior model that is closer to the true process that generated the data if the structure of the prior model does not prevent it. The more data we have, the closer the posterior model should be to the right model.

Since we would like to consider the effect of using a more 'realistic' prior assumption, we have to distinguish between what is 'realistic' and what is not. In this thesis we used a subjective method to distinguish between them. Some spectra were drawn from the prior models. We judged whether a model was a more 'realistic model' according to our prior knowledge about the typical pattern of the NIR spectra, their covariance matrix and their correlation matrix. This prior knowledge was learned from other cases (see for example [29], [39], and [100]). The generated spectra, their sample covariance matrix and sample correlation matrix were plotted. If the patterns of the NIR spectra, their covariance matrix and their correlation matrix were close to the typical patterns, we concluded that the model was a realistic model.

NIR spectra are usually considered to be smooth spectra. However, whether they are theoretically smooth (differentiable) or not is unknown and may not be a sensible question. Although we know the AR(1) process is non-differentiable and the 2^{nd} derivative of an AR(2) process does not exist, we consider models M.c and M.e are very similar to the real process that produce the NIR spectra in comparison with M.a and M.d because M.a and M.d are simply white noise processes. One of the considerations in using the ARMA-type models is still the convenience in computation. The analytical inverses and determinants of their covariance matrices are available. They are also recommended by Brown [23] for the property of structural coherence.

8.3 Correlation Structure

There are many other random processes with alternative covariance structures that we might have considered as possible models. The ARMA models we considered in this thesis involve a second order stationarity assumption, and they are short memory processes. Somehow, these may not necessarily be appropriate assumptions for our spectral data. It has been suggested in Brown *et al.* [24] for example that the NIR spectra may be long-memory processes.

A typical example of a long memory process is the stationary fractional ARIMA process (see Brockwell and Davis [19]). Fractional ARIMA models can also be non-stationary processes (see Beran [7]). The covariance structure of a fractional ARIMA is more complicated, and it may not have nice algebraic properties for efficient computing. Intrinsic random functions (IRF) [33] are a more general class of processes that we may also consider. This has been widely applied in geostatistics and spatial-temporal modelling. The variogram is a function that is used to define an IRF. By checking the sample variogram one may be able to get information which cannot be easily seen in sample correlation. Chilés and Delfiner [33] suggest several variogram models which apply to models with different covariance structures. Haslett [78] suggests the use of variogram in fitting non-stationary time series. We might also apply the variogram in modelling our spectra.

According to figure 5.3 (a.1), the spectra have many crossovers at wavelength 960nm. Suppose each spectrum were to be cut at 960nm into two sections, one can see that the autocorrelation within each section is higher and the cross-correlation between these two sections is lower. The sample correlation in figure 5.3 (a.3) also demonstrates this. There may be a parametric correlation structure that can describe this non-stationary situation. Such correlation structure would be much closer to the true structure. According to the results in chapter 6 and 7, however, a structure closer to the true state does not seem to be able to provide better predictive performance under the inverse-Wishart assumption for the covariance matrix for Y and X.

8.4 Regression

In the regression analysis, we have fitted 5 Bayesian models and compared them with PCR models. We considered M.c and M.e the models that are the closest to the realistic situation. However, the predictive performance shown in chapter 7 does not support M.c and M.e as better models. We also found that their poorer performance was not due to overfitting. One of the advantages of Bayesian modelling is that it prevents a fitted model from overfitting. This is automatically achieved by introducing a suitable prior distribution for parameters.

Let Y denote the response variable and X denote the vector of the regressors. The normal inverse-Wishart models for such high dimensional data are usually chosen for computational convenience. It is known such models lack flexibility since in real cases data are rarely normally distributed. If the true joint model of X and Y is known, one should be able to make a good prediction for Y using X under the true model. However, a model we can choose is seldom a true model but just an approximation. That is, the chosen model can never perfectly describe the event we are interested in. One can only choose a model according to some criteria. For regression analysis, good estimates for the regression coefficients are what we are looking for. In the maximum likelihood approach, we choose the model that maximises the likelihood function as the best model. In the Bayesian approach, the posterior distribution of the regression coefficients rather then one fixed solution is obtained

Suppose the likelihood function of our hierarchical Bayesian random regression model is $l(\beta, \Gamma, \Sigma_{xx}, \theta | X_t, Y_t)$ or also denoted as $p(X_t, Y_t | \beta, \Gamma, \Sigma_{xx}, \theta)$, where (X_t, Y_t) are training data for (X, Y), β is the matrix if regression coefficients, Σ_{xx} is the covariance matrix of X, and $(1 + \theta)\Gamma$ is the variance of the residual $Y - X\beta$. Let us discuss the problem in a maximum likelihood context here. In the controlled regression case, getting the maximum likelihood estimates of all parameters is equivalent to getting estimates that minimise the sum of squares of $Y_t - X_t\beta$. The solution optimises the ability to predict the response variable using the regressors for the training data. However, it is not the case in our random

regression model. The maximum likelihood estimates of the parameters for the random regression model do not yield a model which makes the best prediction for Y using X but describes the joint behaviour of X_t and Y_t, the best in the sense of maximising their joint likelihood function. A similar explanation applies to our Bayesian estimates in chapter 6. Bayesian estimators which explain the distribution of (X, Y) well do not necessarily minimise the Y residuals.

The likelihood function of the controlled regression model in our non-conjugate framework is $p(Y_t - X_t\beta|\beta, \Gamma, \theta)$, which is the factor (a) (see the equation below) of the likelihood function of the random regression model such that

$$p(X_t, Y_t|\beta, \Gamma, \Sigma_{xx}, \theta) = \underbrace{p(Y_t - X_t\beta|\beta, \Gamma, \theta, X_t)}_{\text{(a)}} \underbrace{p(X_t|\Sigma_{xx})}_{\text{(b)}}, \tag{8.1}$$

where the factor (b) $p(X_t|\Sigma_{xx})$ is the marginal likelihood function of X. In a low dimensional case when (a) dominates the likelihood function (8.1), good Bayesian estimators for the parameters in (8.1) may also be good estimators for the controlled regression model. When the number of regressors is high, the likelihood function (8.1) is then dominated by factor (b). Bayesian estimators which are good for (8.1) are not good enough for the controlled regression model. Therefore, in the case that we emphasise on the use of a more realistic prior for the regressors in a high dimensional case, other estimators rather than the estimators good for the joint model of X and Y should be used. For example, one can use the decision approach, by introducing a loss function such as $(Y_t - X_t\beta)^t(Y_t - X_t\beta)$, then choose the β which minimises the loss function. Further improvement can possibly be achieved by using a more flexible sampling distribution and prior distribution rather than the normal inverse-Wishart model.

Given the expectation, Q, of the covariance matrix, Σ, of (Y, X), it is known that the posterior mean of the regression coefficient vector is $(X^tX+Q_{xx})^{-1}(X^tY+Q_{xy})$ under the conjugate analysis. In Brown *et al.* [25, 28]), the covariance matrix of all variables has been assumed implicitly to be a multiple of the identity matrix. The posterior mean of the regression coefficient vector given the mean of this covariance structure is then the same as the typical ridge estimator $(X^tX+kI)^{-1}X^tY$ (see section 3.6.2). In some applications (eg our example), the ridge estimator for

the regression coefficients may be inappropriate because it shrinks important information but keeps the unimportant information (some of which is associated with large eigenvalues of X^tX) in the explanatory variables. As a result, ridge regression may produce a worse result than other approaches can provide (see Fearn [55]). An interest in this thesis is whether an alternative structure for the covariance matrix may preserve the important information in the explanatory variables rather than shrink it. However, a theoretical study of whether this purpose can be achieved by using a more complicated covariance structure rather than a diagonal matrix is not attempted in this thesis, and our practical results do not show that either the AR(1) or AR(2) structures we used could achieve this aim.

We have used the non-conjugate regression model involving a random error independent of the explanatory variables in the response variable. In this thesis, we have not actually compared a conjugate model with the non-conjugate one in the data analysis. One may query what the difference in the results for the two models would be. The mean of full conditional distribution for β is

$$\hat{\beta}_{non-con} = [(1+\theta)Q_{xx} + X_t{}^tX_t]^{-1}[(1+\theta)Q_{x\eta} + X_t{}^tY_t].$$

As we have mentioned in section 6.10, the hyparameters relating to Q_{xx} are strongly dominated by the information from X^tX alone. We can expect this to be the same in the conjugate case. Since $(1+\theta)$ is always greater than one, the existence of the extra error term actually reduces the importance of the data, which contain the real information for β, and enhances the importance of the prior covariance structural assumption, which is rather arbitrary.

In order to improve the model, a more flexible prior assumption might help. One suggestion is to apply the generalized inverse-Wishart distribution [26] for Σ. We can consider a different prior distribution for β, which keeps the correlation structure of X and Y. The original prior distribution of β is $Q_{xx}^{-1}Q_{x\eta} + \mathcal{N}(Q_{xx}^{-1}, \Gamma)$, where $Q_{xx} = \Lambda_{xx}R_{xx}\Lambda_{xx}$. Since a in the diagonal matrix Λ_{xx} strongly affects the posterior mean of β and a is almost decided by X only, we introduce a new diagonal matrix V_{xx}, which is not a function of a but some new parameters, to substitute Λ_{xx}, so that a will not strongly affect the posterior density function of

β. V_{xx} might be able to cancel out the effect of θ in the mean of β so that θ could focus on explaining the regression error.

8.5 Discrimination

In the discriminant analysis, we are to fit a model that describes the joint behaviour of the variables well and use it to derive the posterior group member probability. However, the normal inverse-Wishart assumption for our high dimensional problem still does not provide a good correct-classification rate for our validation data. In our analysis, the difference between models M.a and M.b is not very much in term of the correct-classification rate. Considering the correct-classification rate for the training data, we also find that the models M.b and M.c are not greatly different. However, the correct-classification rates on the validation data for M.c with different levels of δ are obviously less that the rates of the other two models.

The mean spectra of the 9 populations in the example are in fact quite close. The validation data in one particular population frequently can not be clearly distinguished from the training samples belong to other populations according to the means of the populations. The within group variances and the autocovariance structure for the spectra in a population may be much more important for discriminating a future observation. In our model, we assume the samples in different populations have the same covariance matrix. This has prevented us from extracting the covariance information of each population. Therefore, the covariance structure of the samples in each population becomes more important in helping discriminate between samples. A discriminant model assuming different prior covariance matrices for different populations may be a better model. However, the extra cost would have to be paid in using more parameters.

8.6 Model Assessment and MCMC

Cross-validation is the major tool we used to check and compare the models. Ideally, a leave-one-out validation should be the best way because the number of

observations we have is very small. However, due to the cost of computing, in the regression analysis we only used the leave-one-block-out validation. In the example in chapter 7 for the discriminant analysis, we only use the splitting data validation that we test the model trained by the training data on the validation data. In the leave-one-out validation, the score (MSEP or correct-classification rate) is unique for a fixed data set we observe, while the leave-one-block-out approach and the splitting data method, the score varies according to the way we partition the data. When the difference between two models is not much, leave-one-block-out approach and the splitting data method may not be a good way for model selection. However, these two methods should be able to indicate a significantly different model. In our examples, they singled out the models with AR(2) correlation function as having a much worse predictive performance.

In chemometrics, the leave-one-block-out approach and the splitting data approach are the most popular way for model selection because the leave-one-out method requires much computing effort. Cross-validation can only indicate the predictive ability and possibly indicate outliers. It provides no information on how one can improve a model. In a univariate case, one might be able to obtain information for improving a model from graphical tools such as residual plots. We have been using the 3-D visualised covariance and correlation plot to check our prior assumption about the covariance matrix and the correlation. However, it is very difficult to check whether the multivariate normal distribution is a good distribution for our data. As we have mentioned, we do not have many choices for sampling distribution. Perhaps the predictive performance is the most valuable tool to assess models.

Sensitivity analysis has been mentioned as a necessary model checking procedure because a good model should not be sensitive to small changes in the prior model. However, sensitivity analysis is time consuming. In a high dimensional problem, there are a great many possible combinations of prior settings we ought to test. In our thesis, we did not do much sensitivity analysis because we mainly focused on the effect of the covariance structures. We did find that the posterior

distributions of variables are quite sensitive to the covariance structure we chose, which means that the normal inverse-Wishart distribution is not ideal. Perhaps the general inverse-Wishart distribution proposed by Brown is a better choice for β, Γ and Σ_{xx} because it is more flexible than the standard inverse-Wishart distribution.

We employed the ARMS within Gibbs sampling and ran multiple chains. This is the easiest method we know that can handle a high dimensional situation such as ours. The convergence checking seemed all right for us. The variance ratios of the MCMC simulations for θ and Γ are rather large in some sequences, while the other ratios for the other parameters are generally much closer to one. This should be due to the special reciprocal pattern of the posterior joint distribution of θ and Γ. Perhaps running longer sequences would make their variance ratios closer to one. According to our experience, however, it would not change their marginal posterior distributions of them very much.

8.7 Conclusion

We have proposed a method for carrying out high dimensional regression and discriminant analyses. We have tried to incorporate realistic prior distributions into our models, so that the model does at least generate data that visually resemble our real data. Structural covariance matrices were suggested to reduce the number of parameters in the model, although some models may be impossible to describe with a small number of parameters. It is quite reasonable to use them in our cases. ARMS within Gibbs sampling is an efficient way of sampling from high dimensional distributions. Running multiple chains provides an easy way to check for convergence. The posterior models were assessed by the cross-validation method and other simple approaches.

We consider that models with ARMA-type correlation functions are more realistic models for the NIR spectra than those normally used. However, under the normal inverse-Wishart distribution assumption, the predictive performance of these 'realistic' models is not better than the performance of the less-realistic models. For regression analysis, we can see both with natural spectra and with

artificial spectra that the AR(2)-type models yield a posterior distribution of β with less weight on the data and more weight on the prior information, which may be the reason they provide worse predictions. On the other hand, the model M.a in chapter 6 with an unrealistic covariance structure yields a posterior for β that contains more information from the data and makes better predictions. A more flexible prior such as the generalised inverse-Wishart distribution for Σ might be helpful to improve this situation. Another suggestion is to introduce a loss function so that fitting $Y - X\beta$ is more important than fitting X. For discriminant analysis, it might be worthwhile remodelling the problem with a model that assumes the covariance matrices of the distributions of observations in different populations are different.

The traditional method such as PCR (used in this thesis) or PLS (not used in this thesis but typically rather similar to PCR) for regression provide reliable and cheap results. For discriminant analysis, linear discriminantion with PCA that has been considered in this thesis is a cheap method with a correct-classification rate similar to our more complicated models. Our models have, unfortunately, not been able to provide a better predictive performance. More investigation has to been done before the hierarchical Bayesian model considering the structural information of the explanatory variables can be applied in practice. Future work will be focused on some of the issues for improving predictive performance that have been mentioned in this section. For regression analysis, a model with a more flexible prior for Σ should be implemented. For discriminant analysis, we will consider the model with different covariance matrices for different populations. Then, we may continue to develop models with different covariance structures.

Bibliography

[1] J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response dat. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

[2] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, 2nd edition, 1984.

[3] K. E. Atkinson. *An Introduction to Numerical Analysis.* John Wiley & Sons, New York, 2nd edition, 1989.

[4] J. E. Baker, F. E. Dowell, and J. E. Throne. Detection of parasitized rice weevils in wheat kernels with Near-Infrared spectroscopy. *Biological Control*, 16:88–90, 1999.

[5] J. Barnard, R. McCulloch, and X. L. Meng. Modelling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1312, 2000.

[6] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:330–418, 1763. Reprinted in *Biometrika* 45, 293-315, 1958.

[7] J. Beran. Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models. *Journal of the Royal Statistical Society B*, 57:659–672, 1995.

[8] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[9] J. O. Berger and J. M. Bernardo. On the development of reference priors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 35–60. Clarendon Press, Oxford, 1992.

[10] J. M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society B*, 41:113–147, 1979.

[11] J. M. Bernardo. An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *The Statistician*, 47:101–135, 1998.

[12] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* John Wiley & Sons, Chichester, 1994.

[13] J. Besag. Markov chain Monte Carlo for statistical inference. Working Paper 9, CSSS, University of Washington, 2000.

[14] J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society B*, 55:25–37, 1993.

[15] J. Besag and D. Higdon. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society B*, 61:691–717, 1999.

[16] G. E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, 143:383–430, 1980.

[17] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis.* Prentice-Hall, Englewood Cliffs, 1994.

[18] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, London, 1973.

[19] P. J. Brockwell and R. A. Davis. *Time Series Analysis: Theory and Methods.* Springer-Verlag, New York, 1986.

[20] L. D. Broemeling. *Bayesian Analysis of Linear Models.* Marcel Dekker, London, 1985.

[21] S. P. Brooks and A. Gelman. Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1998.

[22] S. P. Brooks and G. O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335, 1998.

[23] P. J. Brown. *Measurement, Regression, and Calibration.* Clarendon Press, Oxford, 1993.

[24] P. J. Brown, T. Fearn, and M. S. Haque. Discrimination with many variables. *Journal of the American Statistical Association*, 94(448):1320–1329, 1999.

[25] P. J. Brown, T. Fearn, and M. Vannucci. The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika*, 86(3):635–648, 1999.

[26] P. J. Brown, N. D. Le, and J. V. Zidek. Inference for a covariance matrix. In P. R. Freeman and A. F. M. Smith, editors, *Aspects of Uncertainty: a Tribute to D. V. Lindley.* Wiley, Chichester, 1993.

[27] P. J. Brown and T. Mäkeläinen. Regression, sequenced measurements and coherent calibration. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 97–108. Clarendon Press, Oxford, 1992.

[28] P. J. Brown, M. Vannucci, and T. Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society B*, 60(3):627–641, 1998.

[29] D. A. Burns and E. W. Ciurczak, editors. *Handbook of Near-Infrared Analysis.* Marcel Dekker, New York, 1992.

[30] C. E. Byrne, G. Downey, D. J. Troy, and D. J. Buckley. Non-destructive prediction of selected quality attributes of beef by Near-Infrared reflectance spectroscopy between 750 and 1098 nm. *Meat Science*, 49(4):339–409, 1998.

[31] K. Chaloner and R. Brant. A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 75(4):651–659, 1988.

[32] C. F. Chen. Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society B*, 41(2):235–248, 1979.

[33] J. Chilés and P. Delfiner. *Geostatistics: Modeling Spatial Uncertianty.* John Wiley & Sons, Chichester, 1999.

[34] J. Cohen, D. Nagin, G. Wallstorm, and L. Wasserman. Hierarchical Bayesian analysis of arrest rates. *Journal of the American Statistical Association*, 93(444):1260–1270, 1998.

[35] P. Congdon and N. Best. Small area variation in hospital admission rates: Bayesian adjustment for primary care and hospital factors. *Journal of the Royal Statistical Society C*, 49:207–226, 2000.

[36] M. K. Cowels and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–905, 1996.

[37] M. J. Daniels. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, 27(3):567–578, 1999.

[38] M. J. Daniels and R. E. Kass. Nonconjugate Bayesian estimation of covariance matrix and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448):1254–1263, 1999.

[39] A. M. C. Davies and R. Giangiacomo, editors. *Near Infrared Spectroscopy: Proceedings of the 9th International Conference.* NIR Publications, Chichester, 2000.

[40] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, 41:1–31, 1979. (with discussion).

[41] A. P. Dawid. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68:265–274, 1981.

[42] A. P. Dawid. Invariant prior distributions. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Science*, pages 228–236. John Wiley, New York, 1983.

[43] A. P. Dawid. The infinite regression and its conjugate analysis. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 3*, pages 95–110. Clarendon Press, Oxford, 1988.

[44] A. P. Dawid and B. Q. Fang. Conjugate Bayes discrimination with infinitely many variables. *Journal of Multivariate Analysis*, 41:27–42, 1992.

[45] A. P. Dawid and J. Pueschel. Hierarchical modelling for DNA profiling using heterogeneous databases. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 187–212. Clarendon Press, Oxford, 1998.

[46] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society B*, 35:189–233, 1973. (with discussion).

[47] B. de Finetti. *Theory of Probability I.* John Wiley & Sons, New York, 1974.

[48] S. de Jong and H. A. L. Kiers. Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems*, 14:155–164, 1992.

[49] Luc Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.

[50] D. K. Dey, A. E. Gelfand, T. B. Swartz, and P. K. Vlachos. Simulation based model checking for hierarchical model. *Test*, 7:325–346, 1998.

[51] J. M. Dickey, D. V. Lindley, and S. J. Press. Bayesian estimation of the dispersion matrix of a multivariate normal distribution. *Communication in Statistics*, 14:1019–1034, 1985.

[52] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B*, 57:45–97, 1995.

[53] D. Draper. Comment: Utility, sensitivity analysis, and cross-validation in Bayesian model-checking. *Statistica Sinica*, 6:760–767, 1996.

[54] B. Q. Fang and A. P Dawid. Nonconjugate Bayesian regression on many variables. Techinical Report 175, Dept. of Statistical Science, UCL, London, 1996.

[55] T Fearn. A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Applied Statistics*, 32(1):73–79, 1983.

[56] T. Fearn, P. J. Brown, and M. S. Haque. Logistic discrimination with many variables. *Rev. R. Acad. Cienc. Exact. Fis. Nat.*, 93(3):337–342, 1999.

[57] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[58] S. Geisser. Posterior odds for multivariate normal classification. *Journal of the Royal Statistical Society B*, 26:69–76, 1964.

[59] S. Geisser. Bayesian estimation in multivariate analysis. *Annals of Mathematical Statistics*, 36:150–159, 1965.

[60] S. Geisser. Predictive discrimination. In P. R. Krishnaiah, editor, *Multivariate Analysis*, pages 149–163. Academic Press, New York, 1966.

[61] S. Geisser. Estimation associated with linear discriminants. *Annals of Mathematical Statistics*, 38:807–817, 1967.

[62] A. E. Gelfand, D. K. Dey, and H. Chang. Model determination using predictive distributions with implementation via sammpling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 147–167. Clarendon Press, Oxford, 1992.

[63] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

[64] A Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, 1996.

[65] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Champman and Hall, London, 1995.

[66] A. Gelman, X. L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

[67] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

[68] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[69] J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 169–193. Clarendon Press, Oxford, 1992.

[70] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, 1991.

[71] C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral analysis. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

[72] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4):455–472, 1995.

[73] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London, 1996.

[74] W. R. Gilks, G. O. Robert, and E. I. George. Adaptive direction sampling. *The Statistician*, 42:179–189, 1994.

[75] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

[76] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[77] P. J. Green and D. J. Murdoch. Exact sampling for Bayesian inference: Towards general purpose algorithms. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 95–110. Clarendon Press, Oxford, 1999.

[78] J. Haslett. On the sample variogram and the sample autocovariance for noon-stationary time series. *Statistician*, 46:475–485, 1997.

[79] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[80] J. S. Hodges. Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society B*, 60(3):497–536, 1998.

[81] A Höskuldsson. A combined theory for PCA and PLS. *Journal of Chemometrics*, 9:91–123, 1995.

[82] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, 186:453–461, 1946.

[83] V. E. Johnson. A model for segmentation and analysis of noisy images. *Journal of the American Statistical Association*, 89(425):230–241, 1994.

[84] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[85] R. E. Kass and D. Steffey. Approximate Bayesian inference in conditionally independent hierarchical models. *Journal of the American Statistical Association*, 84(407):717–726, 1989.

[86] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[87] G. Kiskó, K. Kaffka, and J. Farkas. Estimation of mouldiness of paprika powder by Near Infrared spectroscopy. In A. M. C. Davies and R. Giangiacomo, editors, *Near Infrared Spectroscopy: Proceedings of the 9th International Conference*, pages 455–461. NIR Publications, Chichester, 2000.

[88] M. Lavine and M. West. A Bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20(4):451–461, 1992.

[89] T. Leonard and J. S. J. Hsu. Bayesian inference for a covariance matrix. *Annals of Statistics*, 20:1669–1696, 1992.

[90] D. V. Lindley. The Bayesian approach. *Scandinavian Journal of Statistics*, 5:1–26, 1972. (with discussion).

[91] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society B*, 34:1–42, 1972.

[92] C. J. Liu, J. Liu, and D. B. Rubin. A control variable for assessment the convergence of the of the Gibbs sampler. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 74–78, 1993.

[93] T. Mäkeläinen and P. J. Brown. Coherent priors for ordered regressions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 3*, pages 677–696. Clarendon Press, Oxford, 1988.

[94] H. Martens and T. Næs. *Multivariate Calibration.* Chichester, Wiley, 1989.

[95] G. J. McLachlan. *Discriminat Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.

[96] K. L. Mengersen, C. P. Robert, and C Guihenneuc-Jouyaux. MCMC convergence diagnostics: A reviewww. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 6*, pages 415–440. Clarendon Press, Oxford, 1999.

[97] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087–1091, 1952.

[98] P. Mykland, L. Tierney, and B. Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241, 1995.

[99] J. M. Olinger and P. R. Griffiths. Theory of diffuse reflectance in NIR region. In D. A. Burns and E. W. Ciurczak, editors, *Handbook of Near-Infrared Analysis*, pages 383–432. Marcel Dekker, New York, 1992.

[100] B. G. Osborne, T. Fearn, and P. H. Hindle. *Practical NIR Spectroscopy.* Longman Scientific & Technical, Harlow, 2nd edition, 1993.

[101] L. I. Pettit. Diagnostics in Bayesian model choice. *The Statistician*, 35:183–190, 1986.

[102] L. I. Pettit and K. D. S. Young. Measuring the effect of observation on Bayes factor. *Biometrika*, 77:455–466, 1990.

[103] S. J. Press. *Applied Multivariate Analysis.* Holt, Rinehart and Winston, London, 1972.

[104] M. B. Priestley. *Spectral Analysis and Time Series.* Academic Press, London, 1981.

[105] J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[106] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society B*, 59(4):731–792, 1997.

[107] R. A. Rigby. A credibility that a new observation belongs to one of two multivariate normal populations. *Journal of the Royal Statistical Society B*, 44(2):212–220, 1982.

[108] R. A. Rigby. Bayesian discrimination between two multivariate normal populations with equal covariance. *Journal of the American Statistical Association*, 92(439):1151–1154, 1997.

[109] B. D. Ripley. *Stochastic Simulation.* John Wiley & Sons, Chichester, 1987.

[110] G. O. Robert and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization. *Journal of the Royal Statistical Society B*, 59(2):291–317, 1997.

[111] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.

[112] B. Schmeiser and M. H. Chen. On hit-and-run Monte Carlo sampling for evaluating multidimensional integrals. Technical Report 91-39, Dept. Statistics, Purdue University, 1991.

[113] G. A. F. Seber. *Linear Regression Analysis.* John Wiley & Sons, New York, 1977.

[114] J. S. Shenk, J. J. Workman Jr., and M. O. Westerhaus. Application of NIR spectroscopy to agricultural products. In D. A. Burns and E. W. Ciurczak, editors, *Handbook of Near-Infrared Analysis*, pages 383–432. Marcel Dekker, New York, 1992.

[115] L. Sørensen and R. Jepsen. Assessment of sensory properties of cheese by Near Infrared spectroscopy. *International Dairy Journal*, 8:863–871, 1998.

[116] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS 0.6 Manual.* MRC Biostatistics Unit, Cambridge, 1995.

[117] D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS Version 0.5:Bayesian Inference Using Gibbs Sampling.* MRC Biostatistics Unit, Cambridge, 1995.

[118] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

[119] M. Stone. Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71(353):114–125, 1976.

[120] M. Stone and R. J. Brooks. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society B*, 52:237–269, 1990.

[121] R. Sundberg. Continuum regression and ridge regression. *Journal of the Royal Statistical Society B*, 55:653–659, 1993.

[122] The MathWorks, Inc. *Using MATLAB: Version 5.* The MathWorks, Inc., Natick, 1996.

[123] G. C. Tiao and G. E. P. Box. Bayesian analysis of a three-component hierarchical design model. *Biometrika*, 54:109–125, 1967.

[124] G. C. Tiao and A. Zellner. On the Bayesian estimation of multivariate regression. *Journal of the Royal Statistical Society B*, 26:277–285, 1964.

[125] R. Weiss. An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society B*, 58(4):739–750, 1996.

[126] M. West. Bayesian regression analysis in the "large p, small n" paradigm. Working Paper 00-22, ISDS, Duke University, 2000.

[127] M. West, J. R. Nevins, J. R. Marksm, R. Spang, C. Blanchette, and H. Zuzan. DNA microarray data analysis and regression modelling for genetic expression profiling. Working Paper 00-15, ISDS, Duke University, 2000.

[128] A. M. Yaglom. *An introduction to the theory of stationary random functions.* Prentice-Hall, Englewood Cliffs, N.J., 1962.

[129] R. Yang and J. O. Berger. Estimation of covariance matrix using the reference prior. *Annals of Statistics*, 22(3):1195–1211, 1994.

[130] B. Yu and P. Mykland. Looking at the Markov sampler through cusum path plots: A simple diagnostic idea. *Statistics and Computing*, 8:275–286, 1998.

[131] A. Zellner. *An Introduction to Bayesian Inference in Econometrics.* John Wiley & Sons, London, 1971.

[132] L. Zhu and B. P. Carlin. Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19:2265–2278, 2000.

Appendix A

Matrix Distributions

Matrix Normal Distribution

Suppose $Z \sim \mathcal{N}(\Lambda, \Sigma)$, $\Lambda > 0$, and $\Sigma > 0$. The pdf of Z is

$$f(Z) = (2\pi)^{-pq/2}|\Lambda|^{-q/2}|\Sigma|^{-p/2} \exp(-1/2)tr(\Lambda^{-1}Z\Sigma^{-1}Z^t).$$

Wishart Distribution

Suppose $\Psi \sim \mathcal{W}(\nu; \Sigma)$, $\nu > 0$, and $\Sigma > 0$. The pdf of Ψ is

$$f(\Psi) = c(q, \nu)|\Psi|^{(\nu-q-1)/2}|\Sigma|^{-\nu/2} \exp -1/2tr(\Sigma^{-1}\Psi)$$

where $c(q, \nu) = 2^{-q\nu/2}/\Gamma_q(\nu/2)$ and

$$\Gamma_p(t) = \pi^{q(q-1)/4} \prod_{i=1}^{q} \Gamma[t - (i-1)/2].$$

Inverse-Wishart Distribution

Suppose $\Phi \sim \mathcal{IW}(\delta; \Sigma)$, $\delta > 0$, and $\Sigma > 0$. The pdf of Φ is

$$f(\Phi) = c(q, \delta)|\Sigma|^{(\delta+q-1)/2}|\Phi|^{-(\delta+2q)/2} \exp -1/2tr(\Phi^{-1}\Sigma).$$

Matrix-t Distribution

Suppose $T \sim \mathcal{T}(\delta; \Lambda, \Sigma)$, $\delta > 0$, $\Lambda > 0$, and $\Sigma > 0$. The pdf of T is

$$f(T) = c(p, q, \delta)|\Lambda|^{(\delta+p-1)/2}|\Sigma|^{-p/2}|\Lambda + T\Sigma^{-1}T^t|^{-(\delta+p+q-1)/2}$$

where

$$c(p,q,\delta) = \pi^{-pq/2}\Gamma_q[(\delta+p+q-1)/2]/\Gamma_q[(\delta+q-1)/2].$$

Matrix F Distribution

Suppose $U \sim \mathcal{F}(\nu,\delta;K)$, $\nu > 0$, $\delta > 0$, and $K > 0$. The pdf of U is

$$f(U) = \frac{\Gamma_q((\delta+\nu+q-1)/2)}{\Gamma_q(\nu/2)\Gamma_q((\delta+q-1)/2)}|K|^{(\delta+q-1)/2}|U|^{(\nu-q-1)/2}|U+K|^{-(\nu+\delta+q-1)/2}.$$

Appendix B

Results for Regression Models

Table B.1: Variance ratio for models M.a-M.e with five blocks of data

The figures in the row of Variance ratio (VR) method 1 are the $\sqrt{\widehat{R}_G}$; in the row of VR method 2 are the $\sqrt{\widehat{R}_C}$; in the row of VR method 3 are the $\widehat{R}_{interval}$ with $\alpha = 0.05$.

M.a								
Block	VR method	a	b	ρ	θ	$\Lambda_{\eta\eta}$	Γ	K
1	1	1.00	1.00	1.00	1.02	1.02	1.02	1.00
	2	1.00	1.00	1.00	1.02	1.05	1.07	1.00
	3	1.00	1.00	1.01	1.00	0.96	0.93	1.00
2	1	1.00	1.00	1.00	1.00	1.01	1.02	1.00
	2	1.00	1.00	1.00	1.01	1.04	1.08	1.00
	3	1.00	1.00	1.00	1.02	1.03	1.02	0.99
3	1	1.00	1.00	1.00	1.03	1.03	1.03	1.00
	2	1.00	1.00	1.00	1.04	1.06	1.09	1.01
	3	1.01	1.00	1.00	1.05	1.10	1.10	1.01
4	1	1.00	1.00	1.00	1.00	1.00	1.01	1.00
	2	1.00	1.00	1.00	1.00	1.01	1.02	1.00
	3	0.99	0.99	1.00	1.01	1.03	1.07	1.01
5	1	1.00	1.00	1.00	1.03	1.05	1.05	1.00
	2	1.00	1.00	1.00	1.05	1.12	1.18	1.00
	3	1.00	1.00	1.00	1.03	1.02	1.04	1.01

Table B.1 (a)

M.b									
Block	VR method	a	b	τ	ρ	θ	$\Lambda_{\eta\eta}$	Γ	K
1	1	1.00	1.00	1.00	1.00	1.05	1.06	1.05	1.00
	2	1.00	1.00	1.00	1.00	1.08	1.09	1.08	1.01
	3	0.98	1.00	0.99	1.01	1.06	1.03	1.01	1.01
2	1	1.00	1.00	1.00	1.00	1.03	1.05	1.05	1.00
	2	1.00	1.00	1.00	1.00	1.07	1.08	1.09	1.01
	3	0.99	0.99	1.01	1.00	1.10	1.07	1.05	0.98
3	1	1.02	1.00	1.00	1.00	1.02	1.00	1.00	1.00
	2	1.00	1.00	1.01	1.00	1.26	1.02	1.02	1.03
	3	1.00	1.00	0.99	0.99	0.82	1.03	1.04	1.02
4	1	1.00	1.00	1.00	1.00	1.02	1.02	1.02	1.00
	2	1.00	1.00	1.00	1.00	1.04	1.04	1.05	1.01
	3	1.01	1.00	0.99	0.99	1.01	1.02	1.01	0.98
5	1	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.00
	2	1.00	1.00	1.00	1.00	1.04	1.04	1.08	1.01
	3	1.00	1.00	0.99	1.00	1.02	1.00	0.97	1.00

Table B.1 (b)

M.c									
Block	VR method	a	b	ϕ_1	ρ	θ	$\Lambda_{\eta\eta}$	Γ	K
1	1	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.00
	2	1.00	1.00	1.00	1.00	1.03	1.02	1.02	1.01
	3	0.98	0.99	1.00	1.00	0.98	1.01	1.01	0.98
2	1	1.00	1.00	1.00	1.01	1.04	1.05	1.05	1.01
	2	1.00	1.00	1.00	1.01	1.09	1.07	1.06	1.03
	3	1.00	1.00	1.00	1.00	1.05	1.03	1.03	0.97
3	1	1.00	1.00	1.00	1.01	1.03	1.02	1.01	1.01
	2	1.00	1.00	1.00	1.01	1.12	1.03	1.02	1.02
	3	1.00	1.01	1.00	1.02	1.11	1.04	1.03	1.02
4	1	1.00	1.00	1.00	1.01	1.02	1.03	1.04	1.00
	2	1.00	1.00	1.00	1.01	1.05	1.05	1.07	1.01
	3	0.99	1.00	1.00	1.01	1.00	1.07	1.07	1.01
5	1	1.00	1.00	1.00	1.02	1.06	1.09	1.10	1.01
	2	1.00	1.00	1.00	1.03	1.12	1.13	1.14	1.03
	3	1.00	1.00	1.00	1.03	1.10	1.07	1.10	1.05

Table B.1 (c)

M.d							
Block	VR method	a	ρ	θ	$\Lambda_{\eta\eta}$	Γ	K
1	1	1.00	1.00	1.04	1.02	1.02	1.00
	2	1.00	1.01	1.22	1.04	1.05	1.00
	3	1.00	1.00	0.96	1.03	1.01	0.98
2	1	1.00	1.00	1.03	1.04	1.03	1.00
	2	1.00	1.00	1.18	1.05	1.04	1.00
	3	1.00	1.00	0.93	1.03	1.01	1.00
3	1	1.00	1.00	1.02	1.01	1.00	1.00
	2	1.00	1.00	1.07	1.03	1.03	1.00
	3	1.00	1.01	1.07	1.05	1.03	0.97
4	1	1.00	1.00	1.00	1.01	1.01	1.00
	2	1.00	1.00	1.14	1.02	1.03	1.00
	3	1.01	1.01	0.79	1.01	1.01	1.02
5	1	1.00	1.01	1.09	1.11	1.10	1.01
	2	1.00	1.01	1.17	1.15	1.15	1.02
	3	1.00	1.00	1.09	1.10	1.11	1.02

Table B.1 (d)

M.e								
Block	VR method	a	ϕ_1	ρ	θ	$\Lambda_{\eta\eta}$	Γ	K
1	1	1.00	1.00	1.00	1.01	1.00	1.00	1.00
	2	1.00	1.00	1.01	1.03	1.01	1.01	1.00
	3	1.00	1.01	0.99	1.05	1.01	1.01	0.99
2	1	1.00	1.00	1.00	1.03	1.01	1.01	1.01
	2	1.00	1.00	1.00	1.17	1.03	1.03	1.02
	3	1.00	1.00	1.01	0.93	1.06	1.07	0.92
3	1	1.00	1.00	1.01	1.00	1.01	1.01	1.00
	2	1.00	1.00	1.02	1.02	1.03	1.04	1.01
	3	1.01	1.00	1.00	1.03	1.01	0.99	0.99
4	1	1.00	1.00	1.00	1.05	1.03	1.03	1.00
	2	1.00	1.00	1.01	1.15	1.04	1.03	1.01
	3	0.99	1.01	1.01	1.07	1.02	1.03	1.02
5	1	1.00	1.00	1.00	1.03	1.05	1.05	1.00
	2	1.00	1.00	1.00	1.05	1.12	1.18	1.00
	3	1.00	1.00	1.00	1.03	1.02	1.04	1.01

Table B.1 (e)

Figure B.1: MCMC output of regression model M.a (4 chains plotted together)

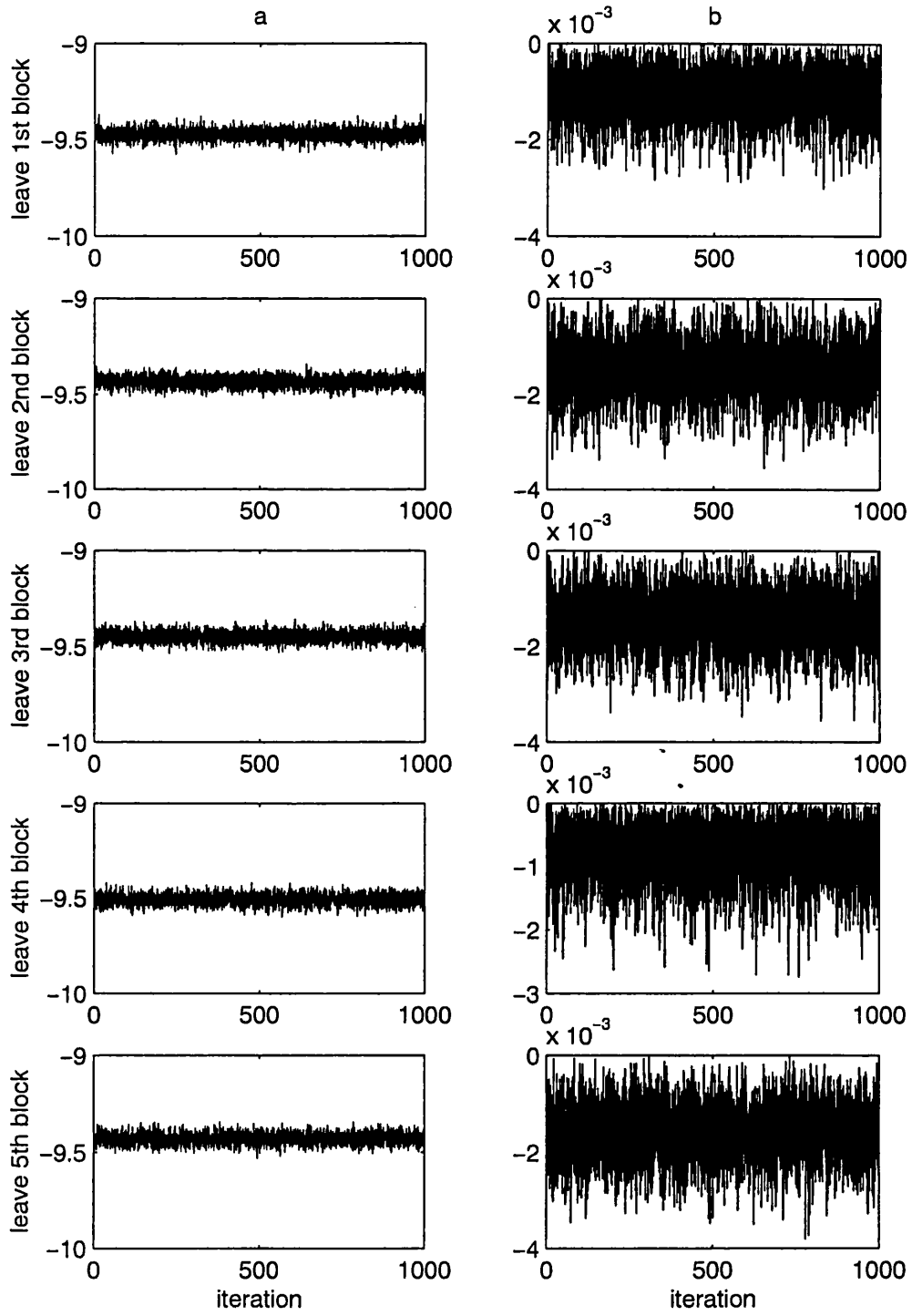


Figure B.1 (a)

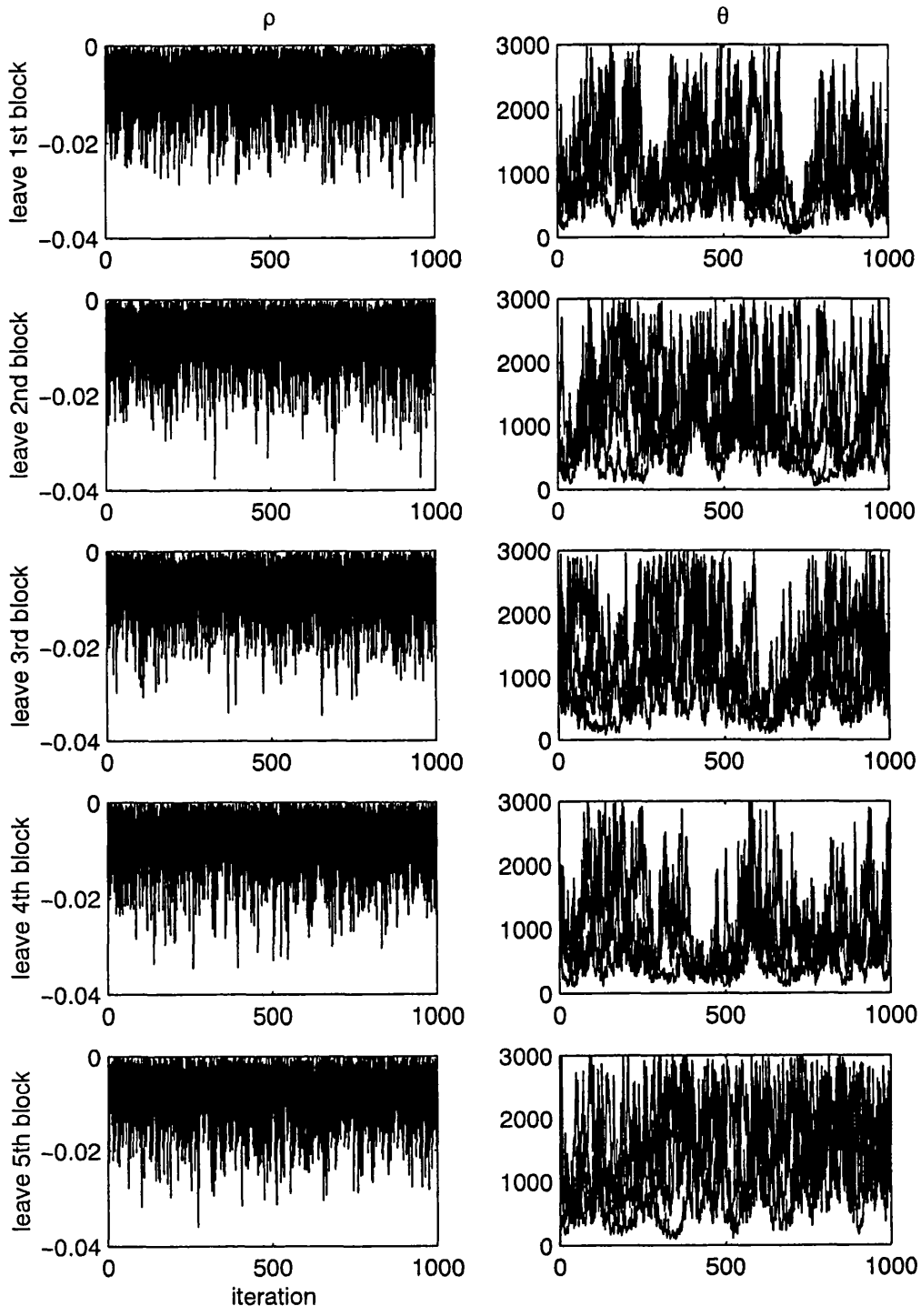

ρ
θ
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0
-0.02
-0.04
3000
2000
1000
0
500
1000
iteration

Figure B.1 (b)

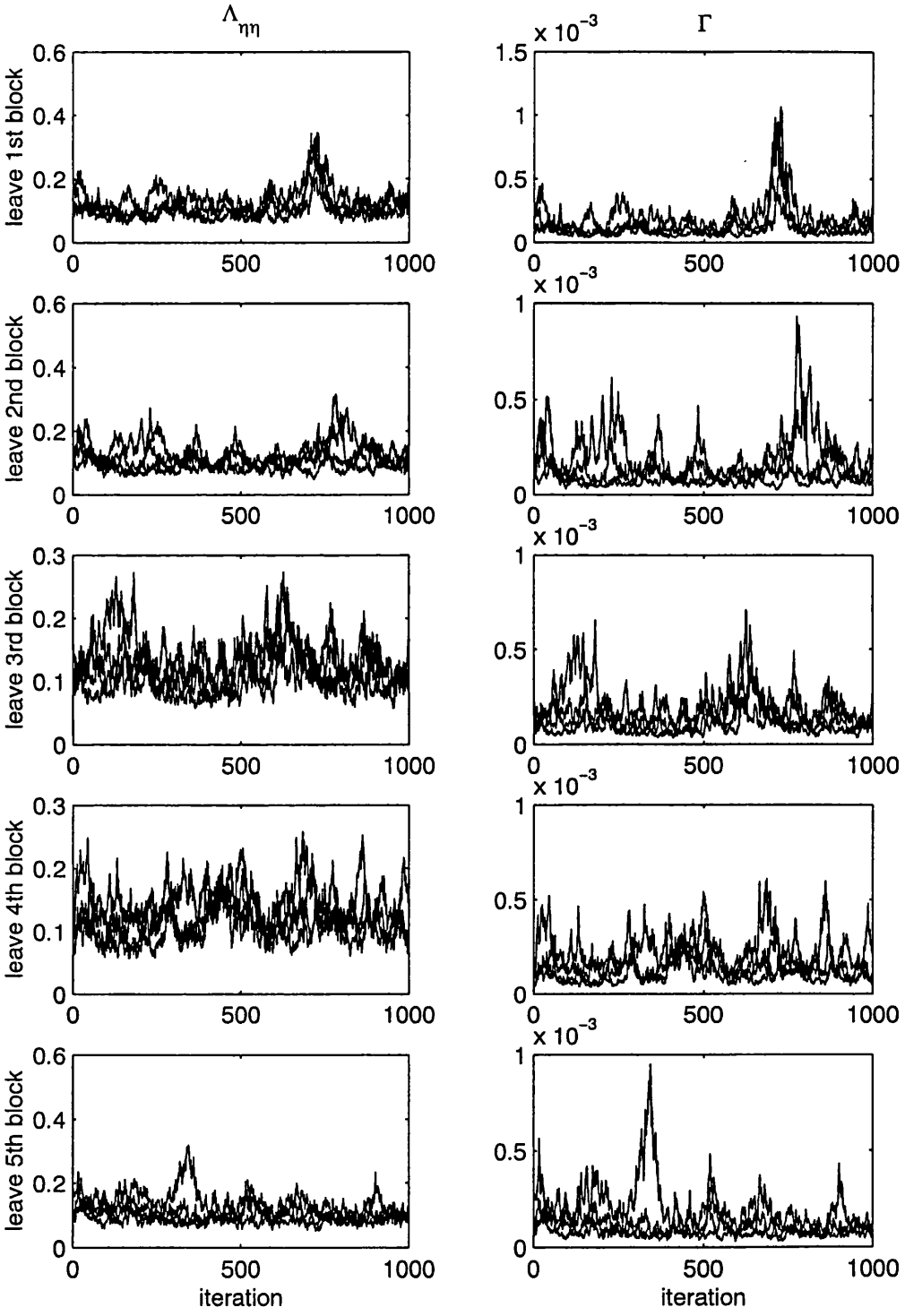

$\Lambda_{\eta\eta}$
Γ
x 10^-3
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
iteration
iteration

Figure B.1 (c)

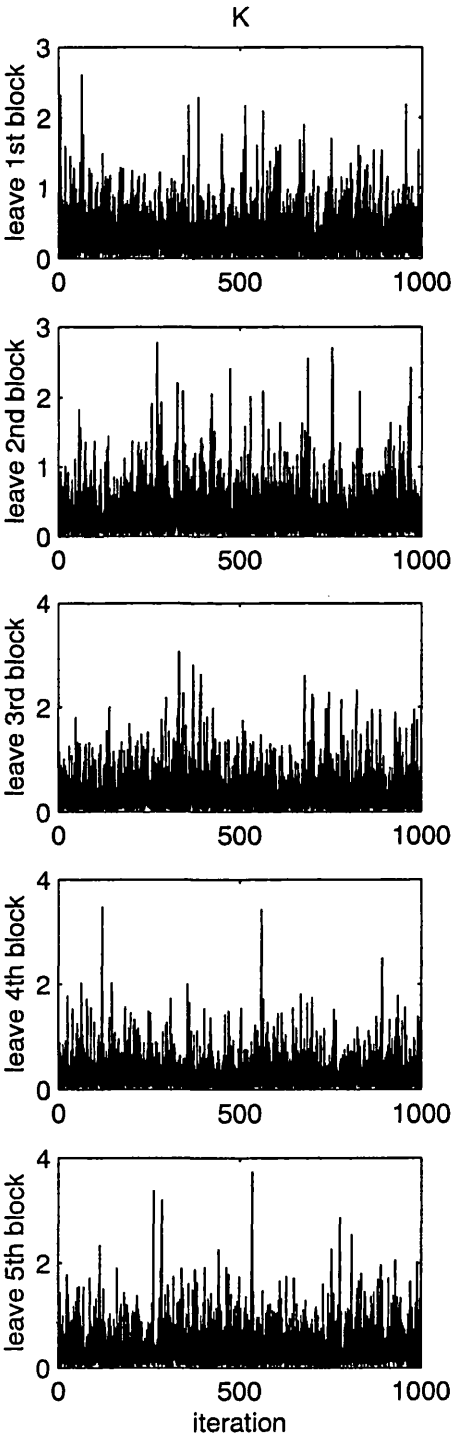

K
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0
500
1000
iteration

Figure B.1 (d)

Figure B.2: MCMC output of regression model M.b (4 chains plotted together)

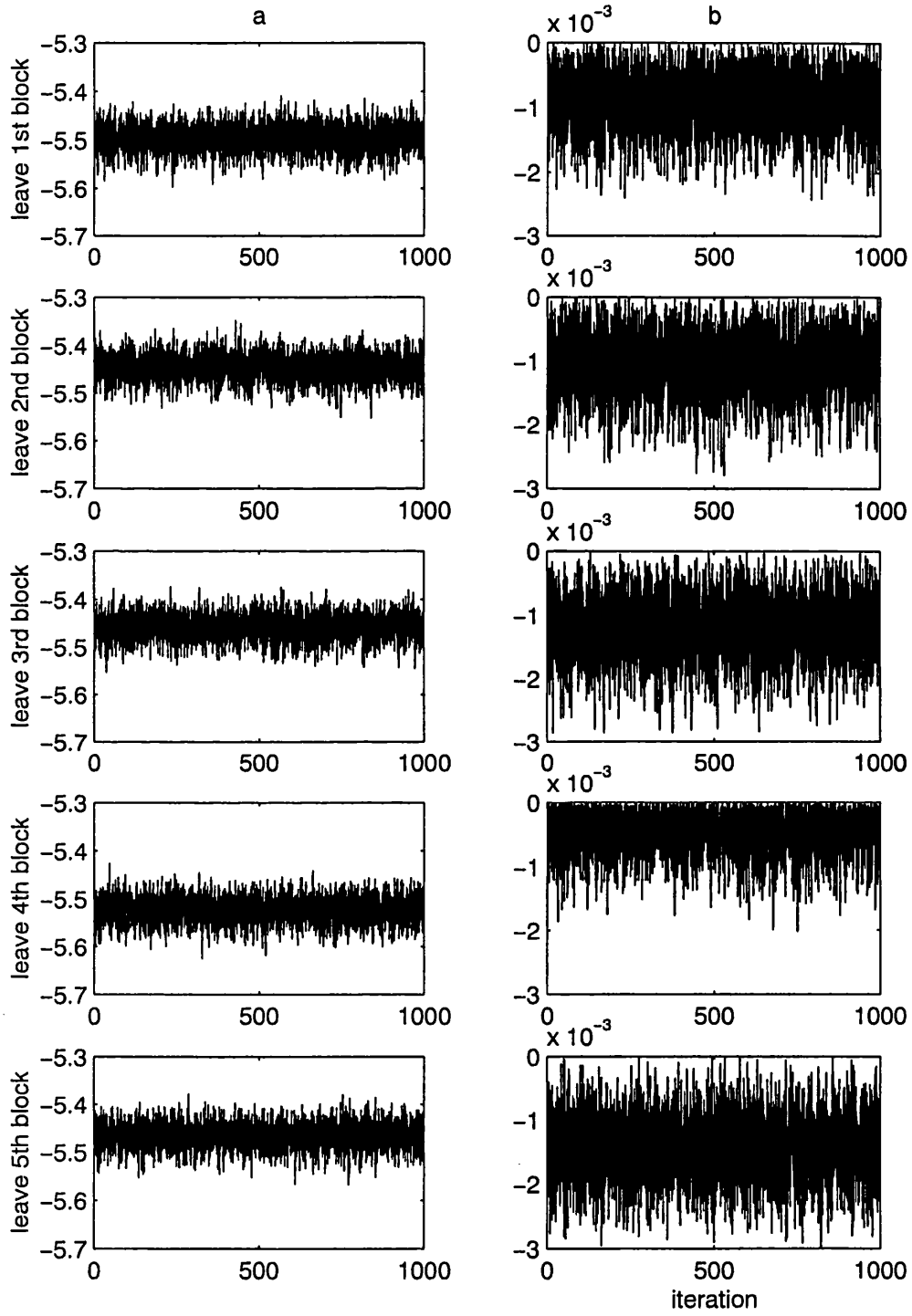


Figure B.2 (a)

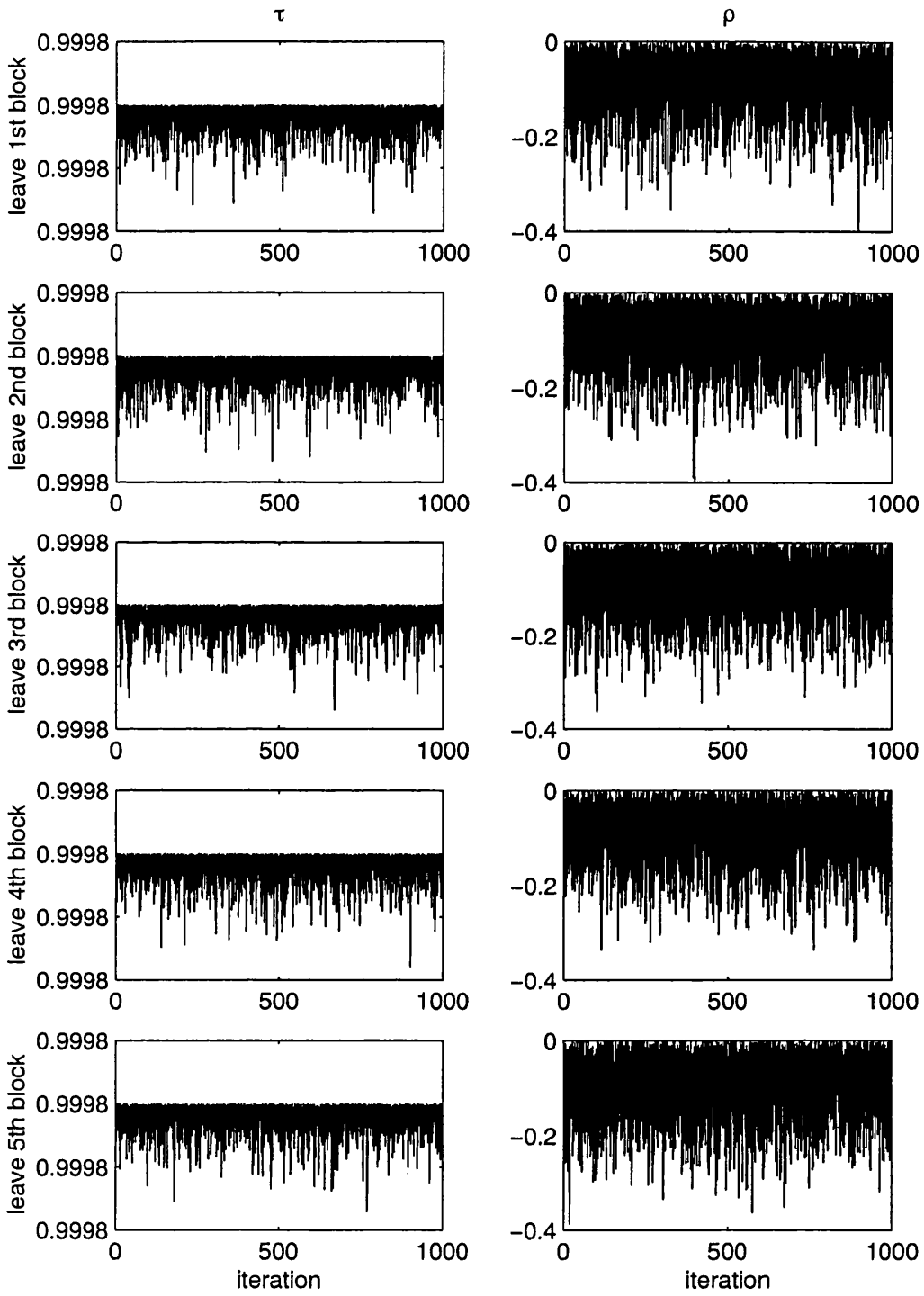

τ
ρ
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0.9998
0.9998
0.9998
0.9998
0
-0.2
-0.4
0
500
1000
iteration

Figure B.2 (b)

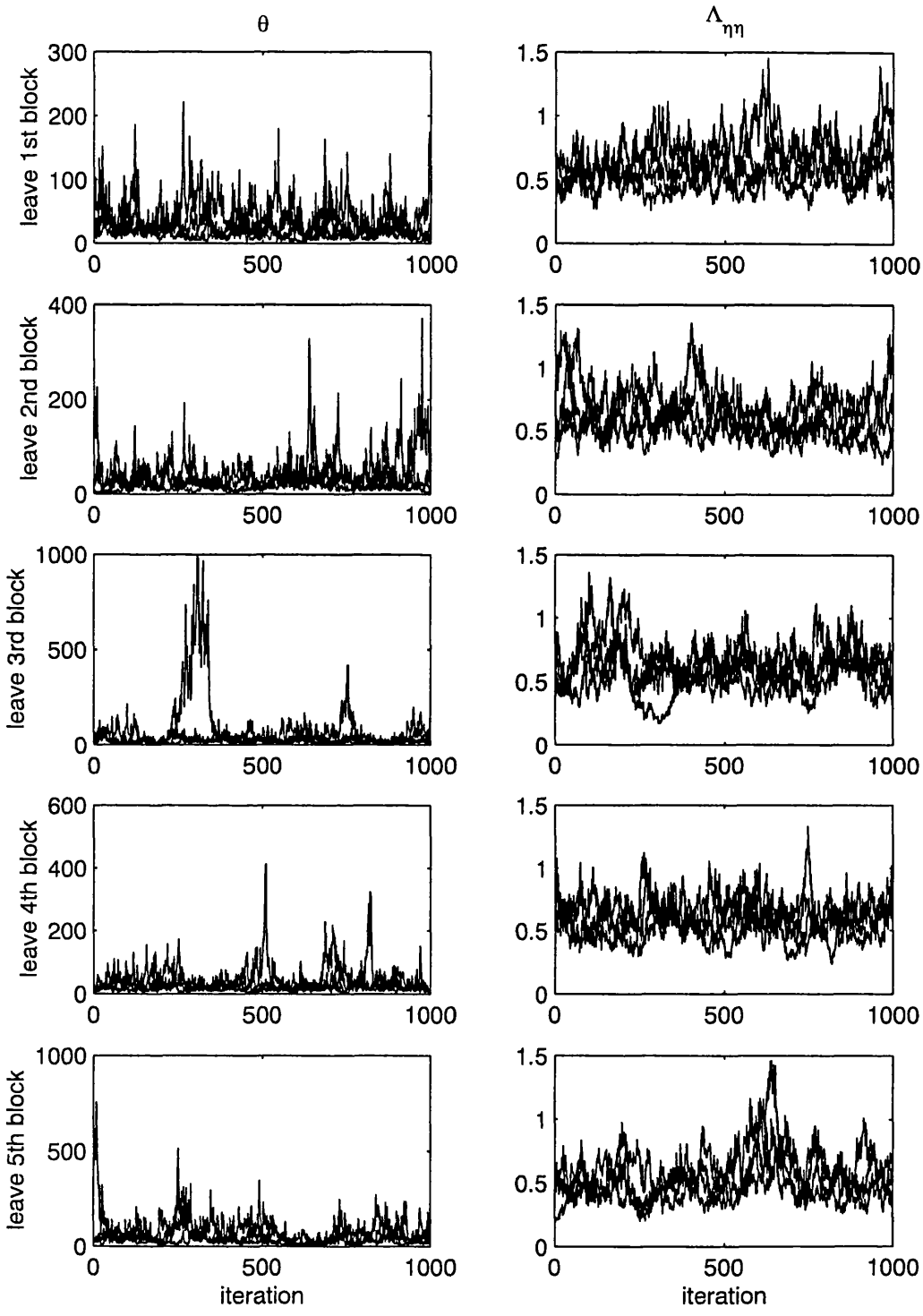

θ
$\Lambda_{\eta\eta}$
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
iteration
iteration

Figure B.2 (c)

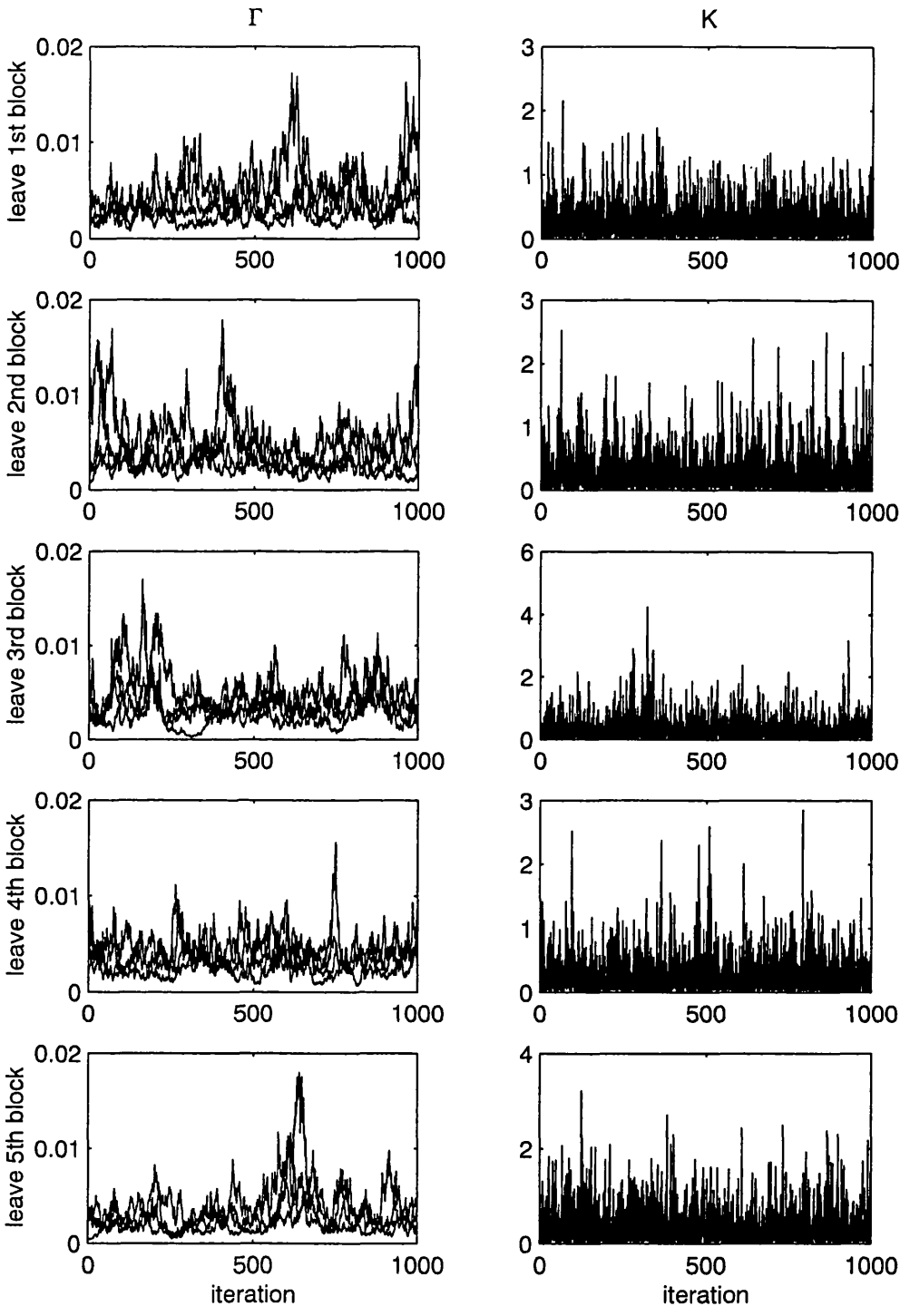

Γ
K
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0.02
0.01
0
3
2
1
6
4
500
1000
iteration

Figure B.2 (d)

Figure B.3: MCMC output of regression model M.c (4 chains plotted together)

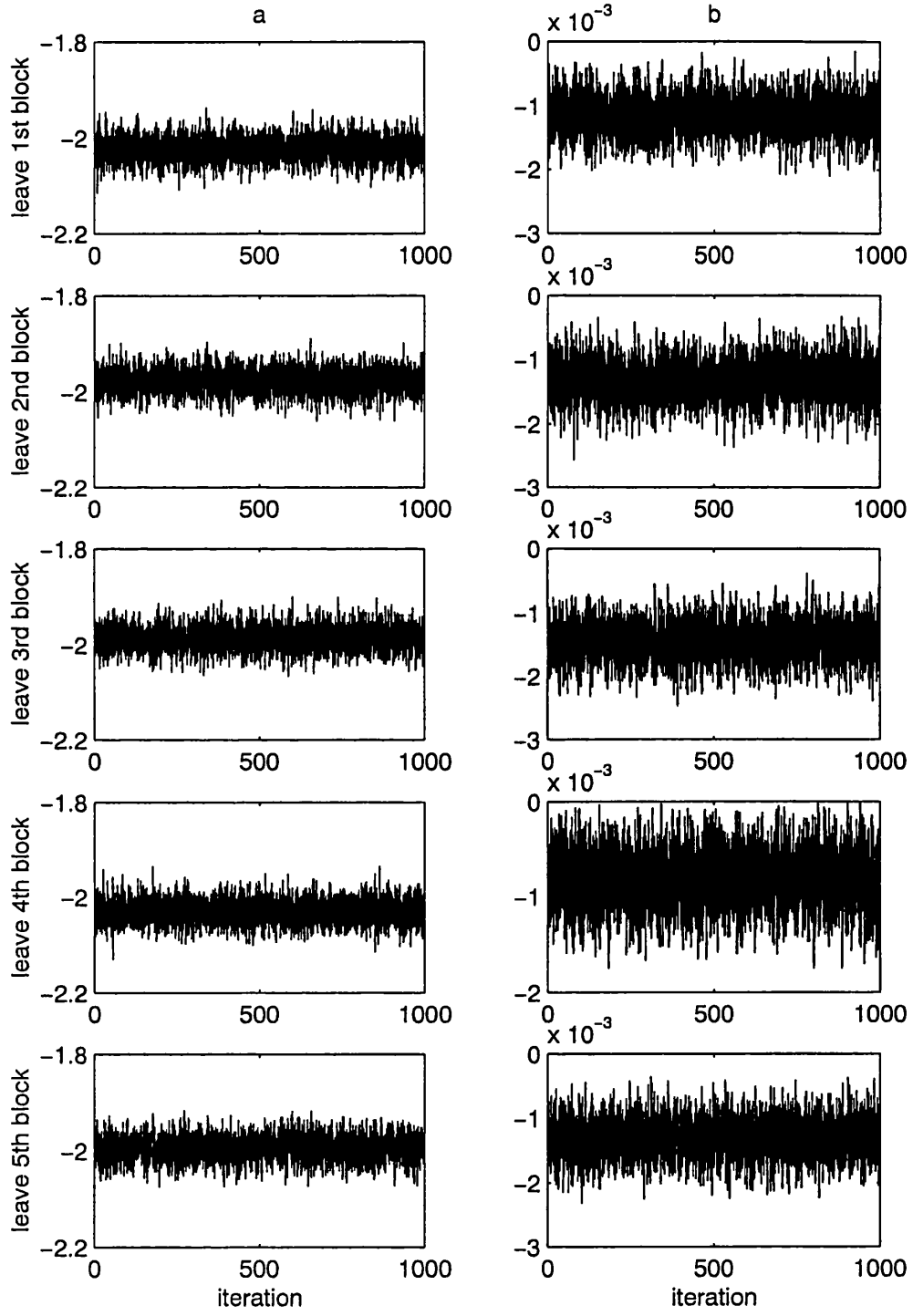


Figure B.3 (a)

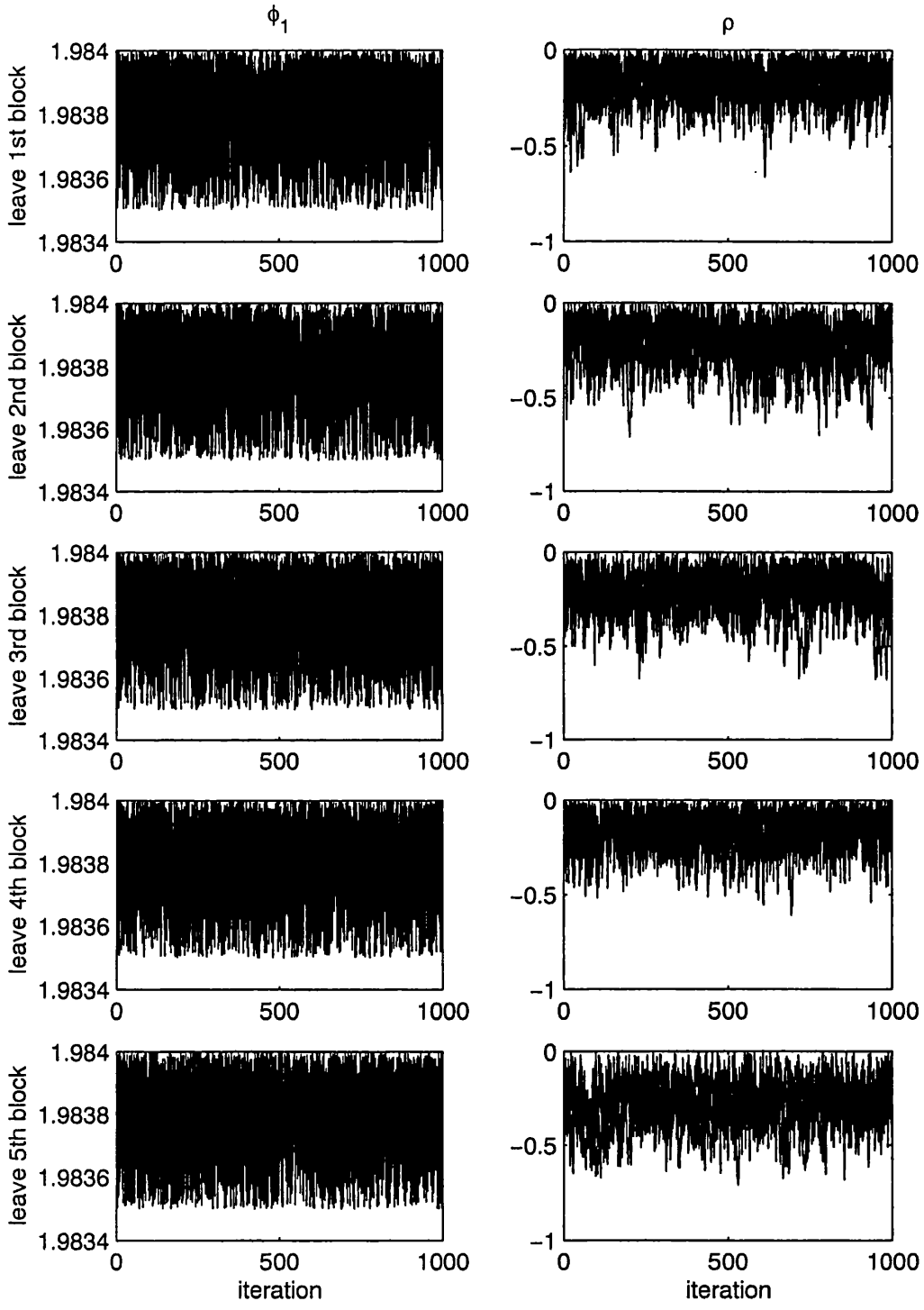

ϕ_1
ρ
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
1.984
1.9838
1.9836
1.9834
0
-0.5
-1
0
500
1000
iteration

Figure B.3 (b)

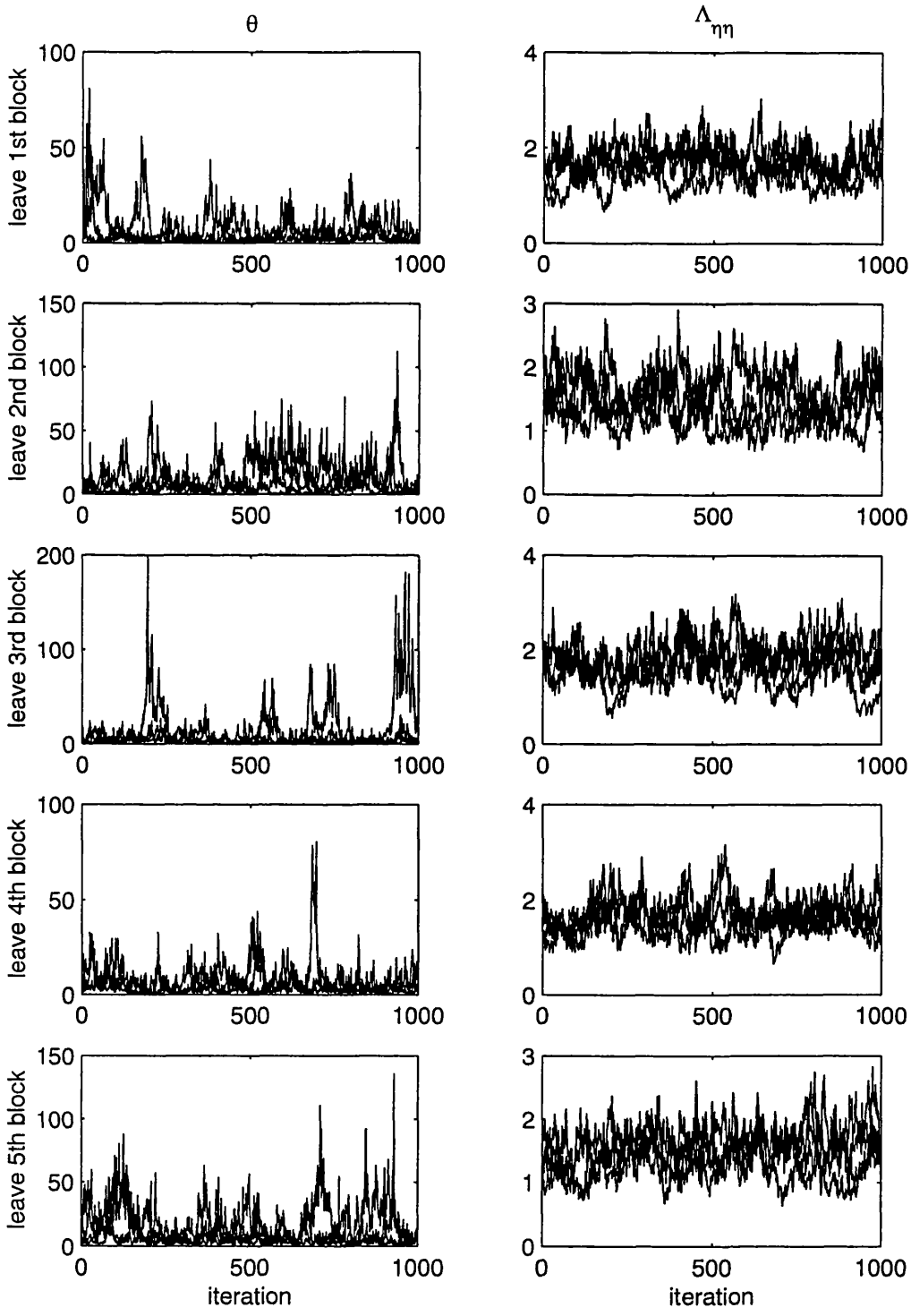

θ
$\Lambda_{\eta\eta}$
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
iteration
iteration

Figure B.3 (c)

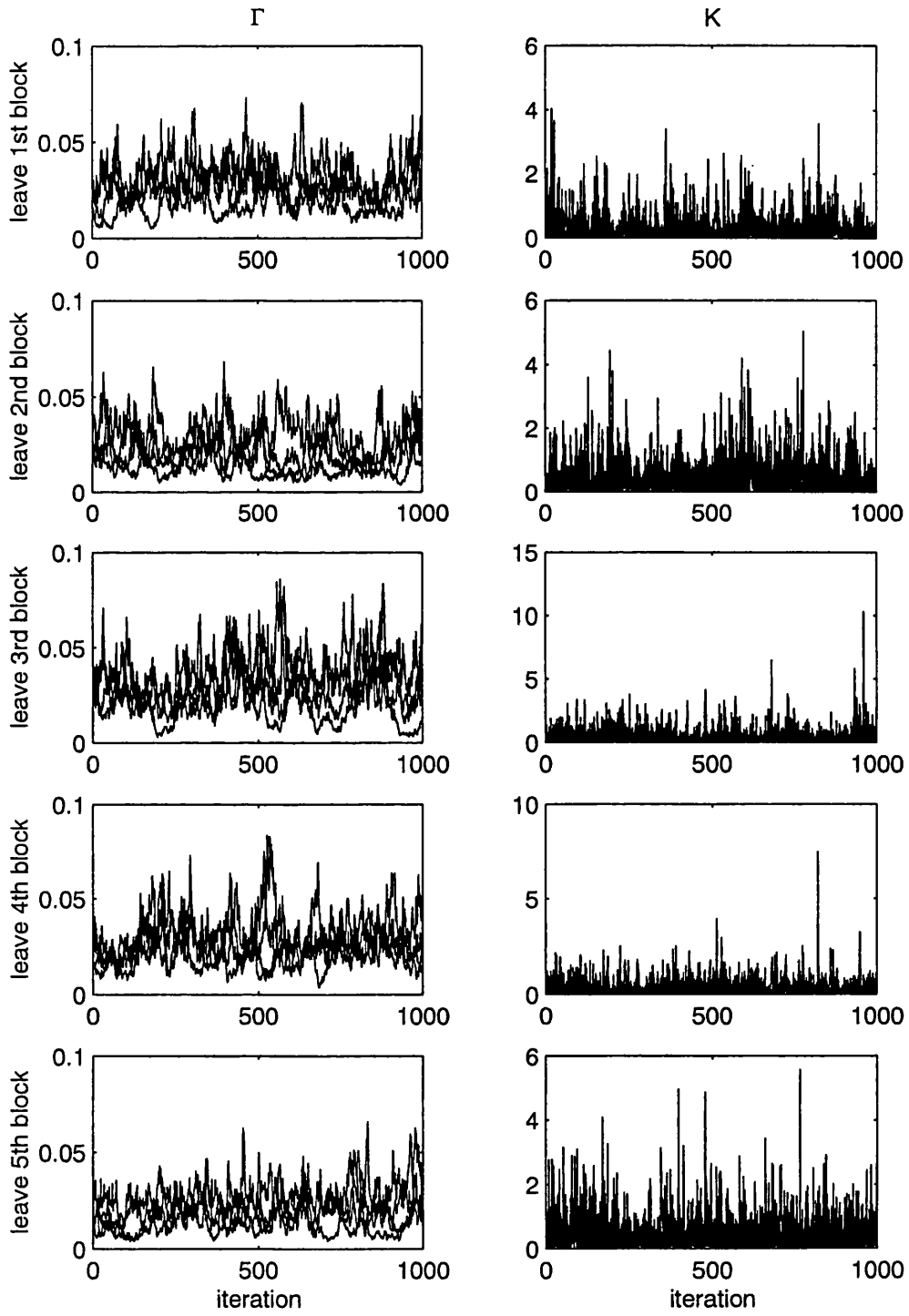

Γ
K
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0.1
0.05
0
6
4
2
15
10
5
0
500
1000
iteration

Figure B.3 (d)

Figure B.4: MCMC output of regression model M.d (4 chains plotted together)

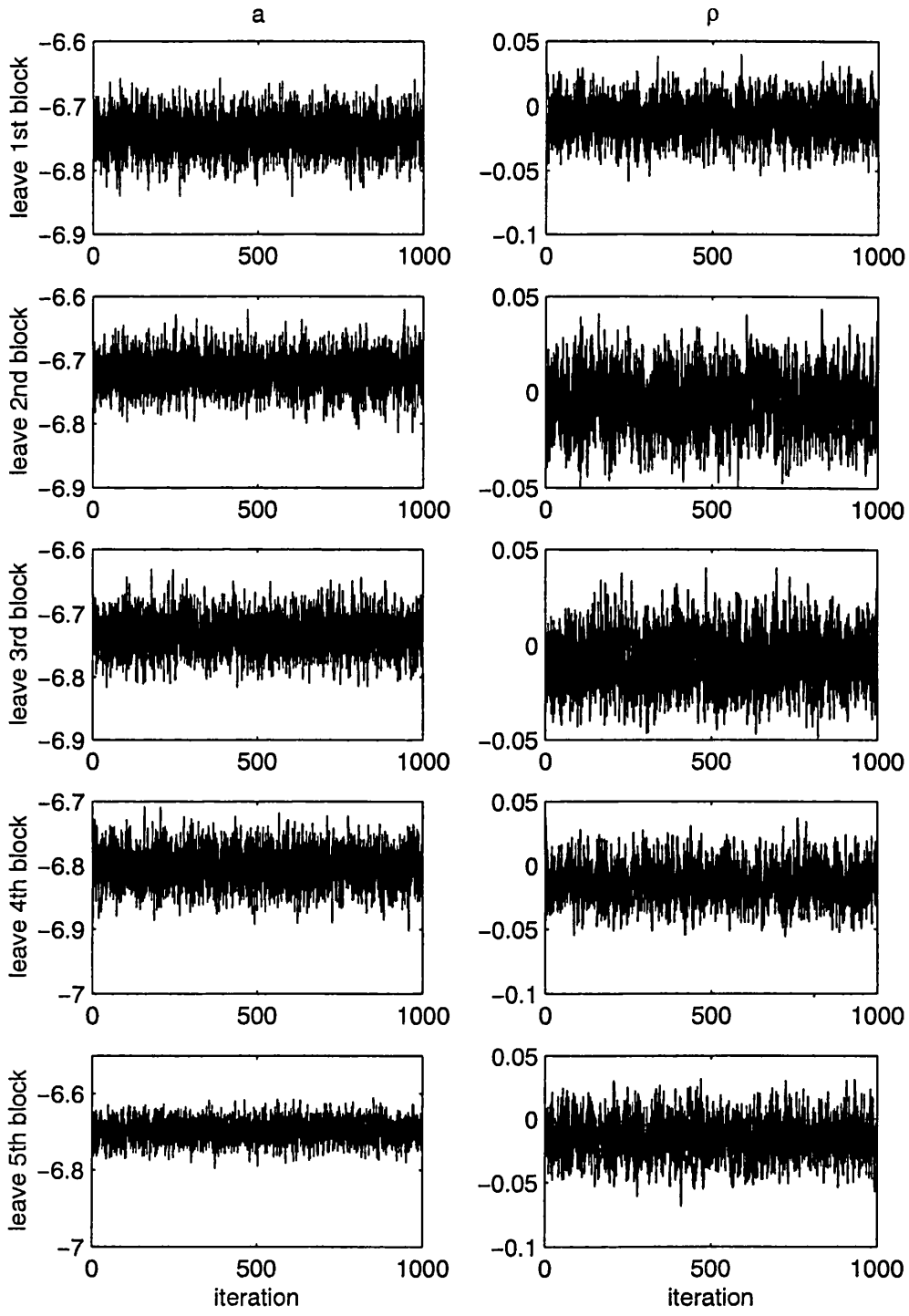


Figure B.4 (a)

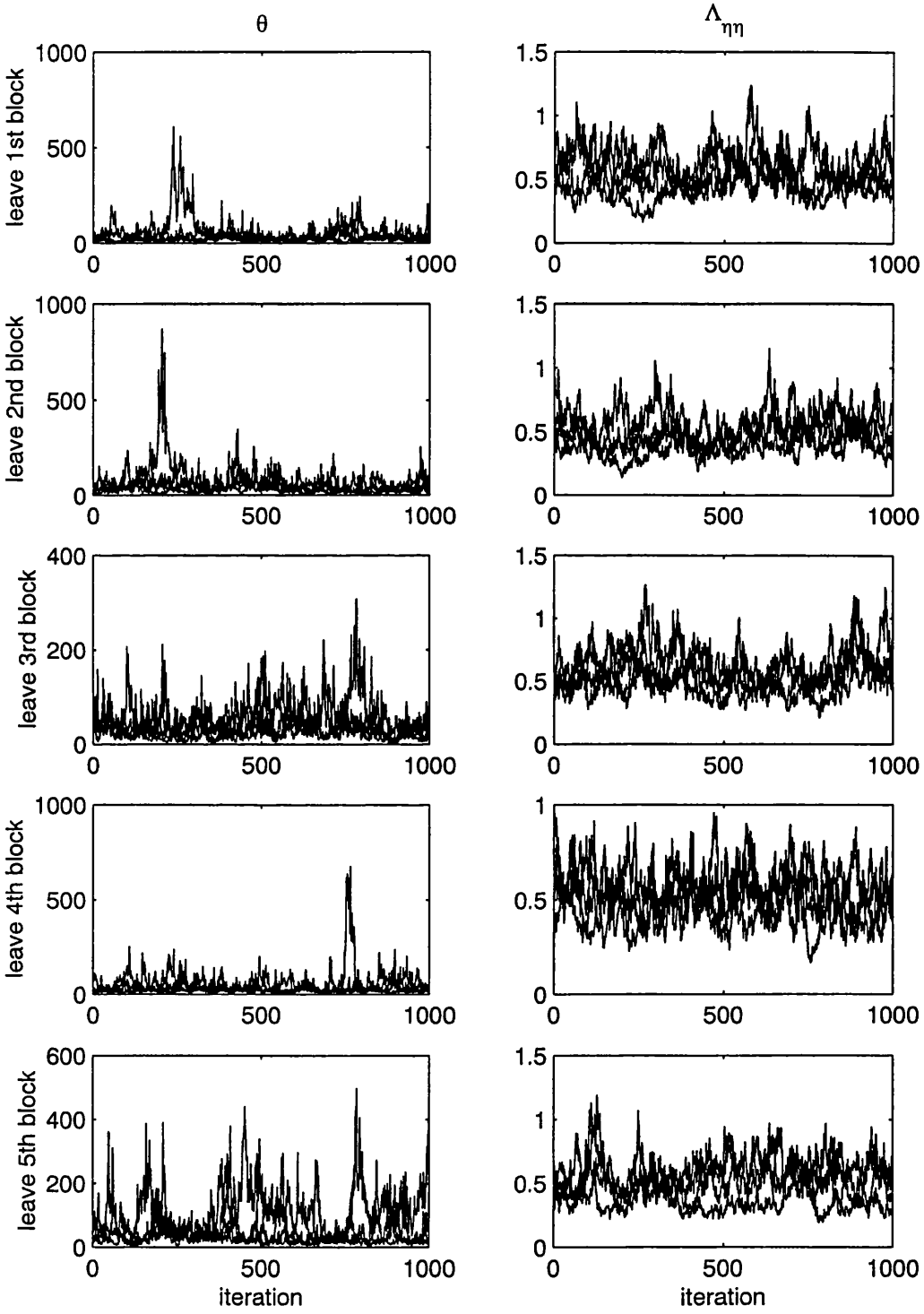

θ
$\Lambda_{\eta\eta}$
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
iteration
iteration

Figure B.4 (b)

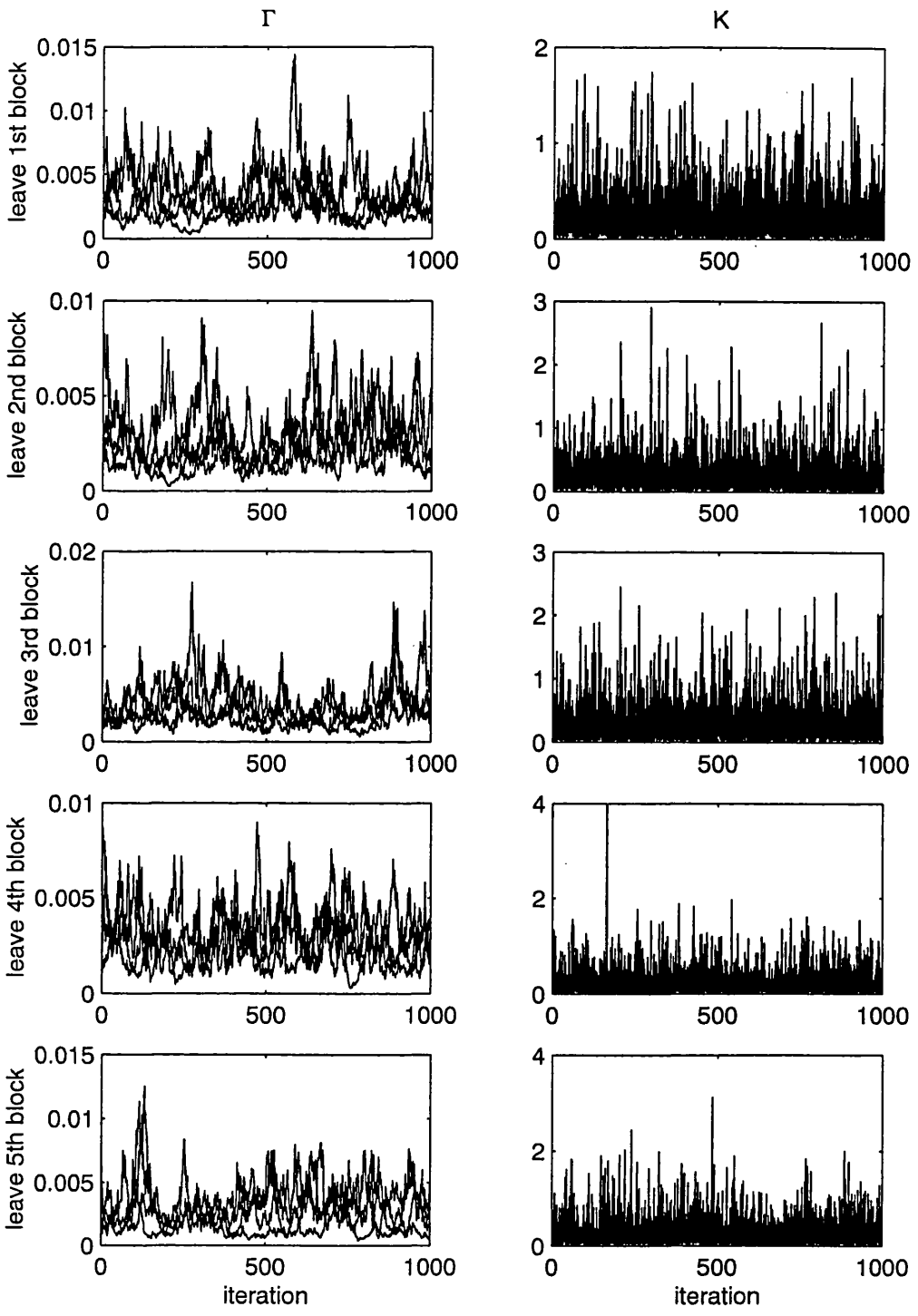

Γ
K
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0.015
0.01
0.005
0
0.02
2
1
3
4
0
500
1000
iteration

Figure B.4 (c)

Figure B.5: MCMC output of regression model M.e (4 chains plotted together)

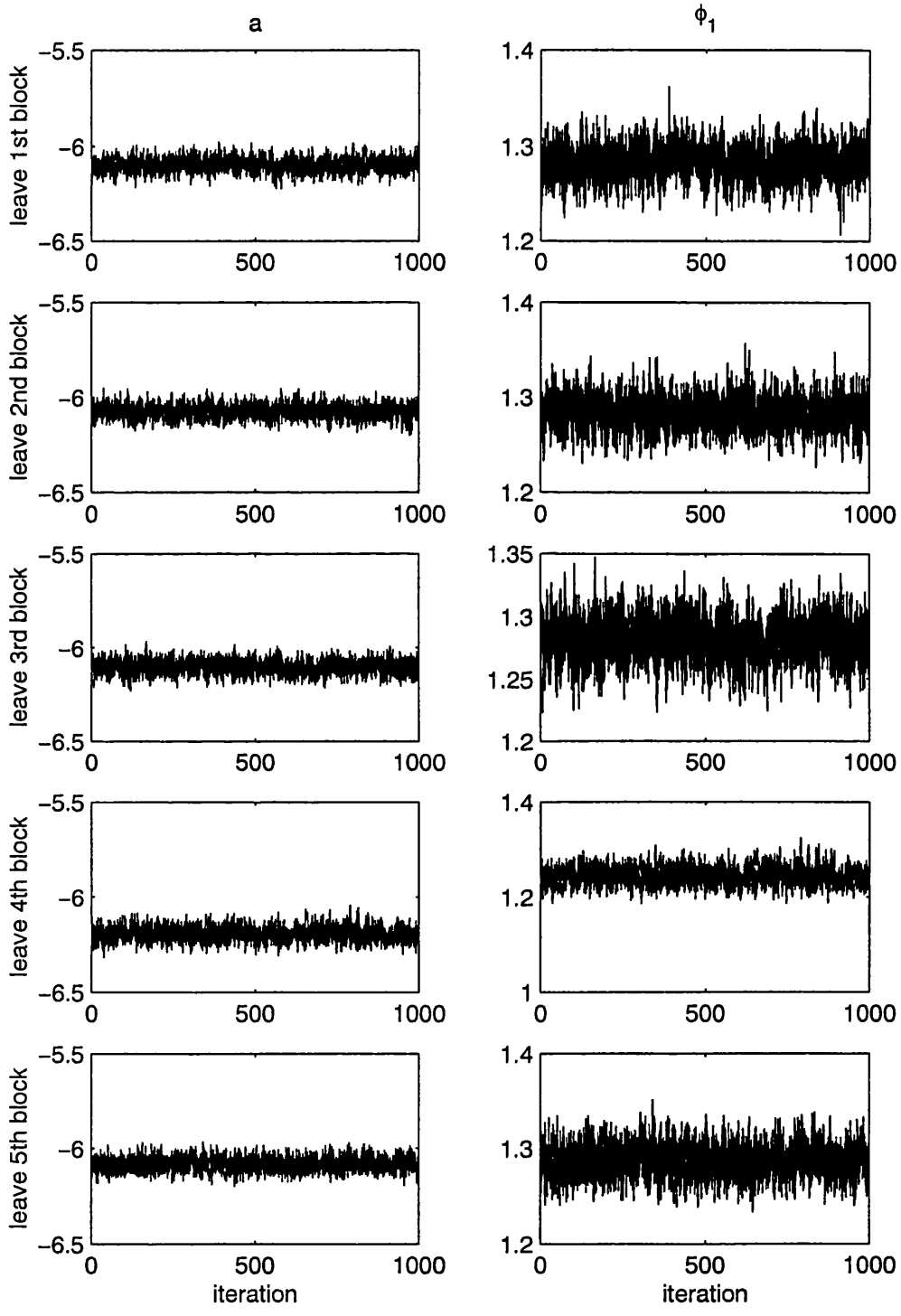


Figure B.5 (a)

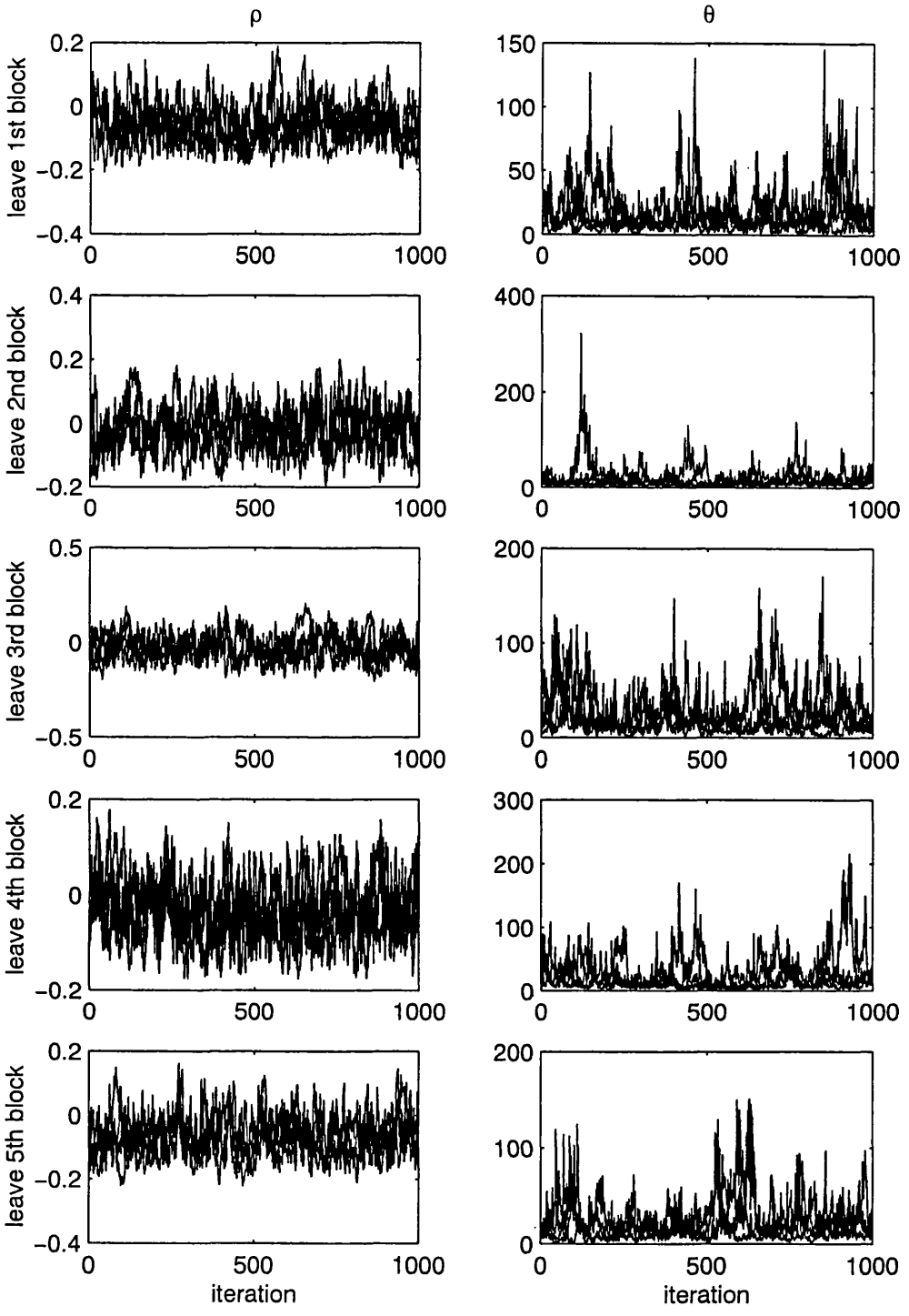

ρ
θ
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0.2
0
-0.2
-0.4
150
100
50
0.4
400
200
0.5
-0.5
300
0
500
1000
iteration

Figure B.5 (b)

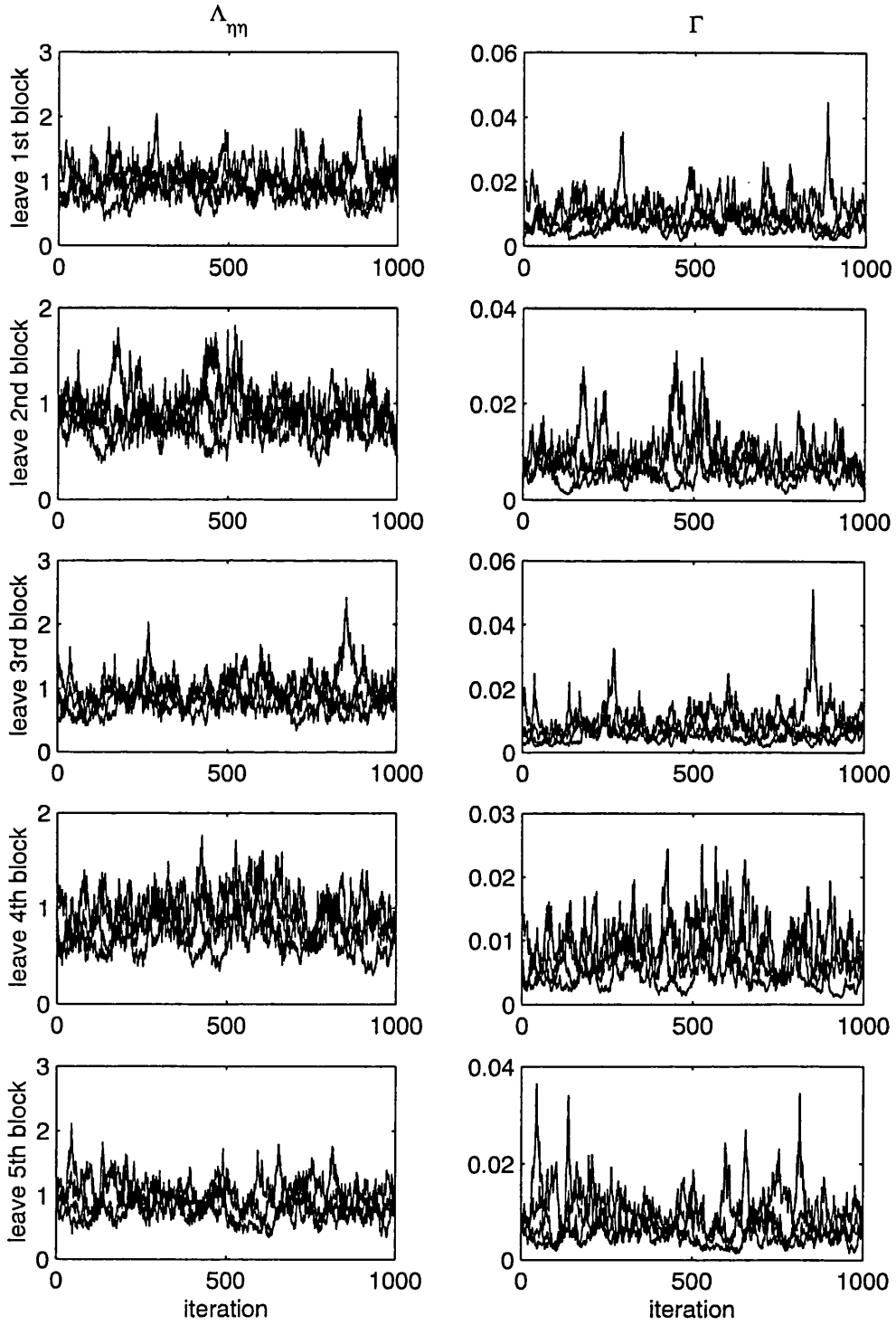

$\Lambda_{\eta\eta}$
Γ
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
iteration
iteration

Figure B.5 (c)

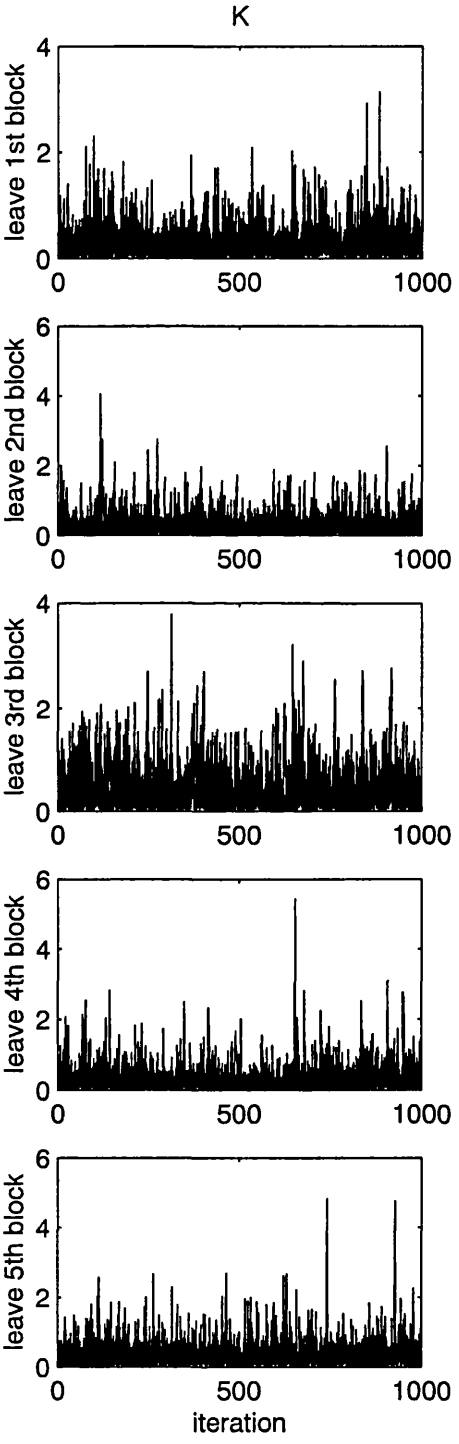

K
leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
0
500
1000
2
4
6
iteration

Figure B.5 (d)

Figure B.6: Histograms of MCMC samples for the parameters in regression model M.a

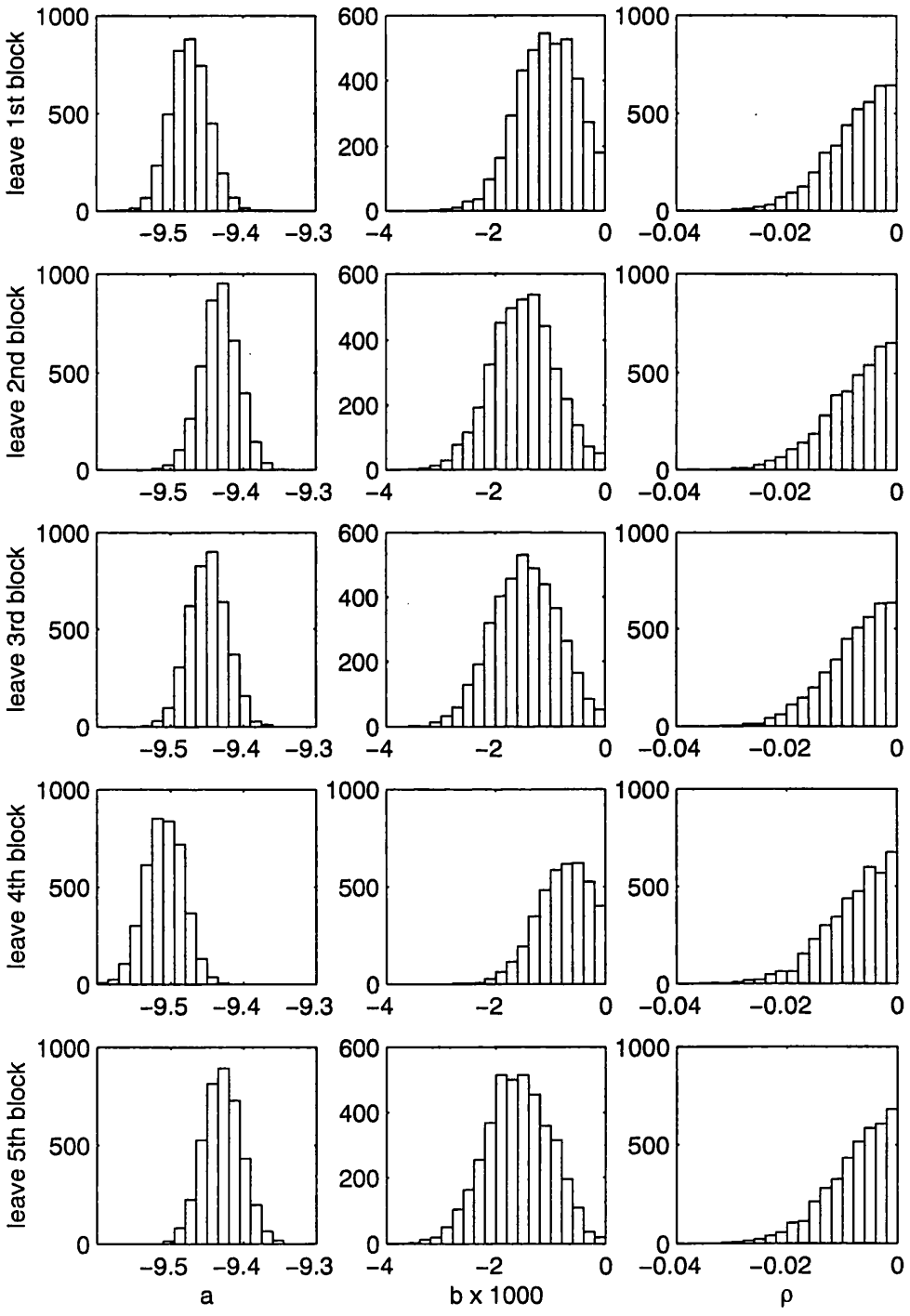


Figure B.6 (a)

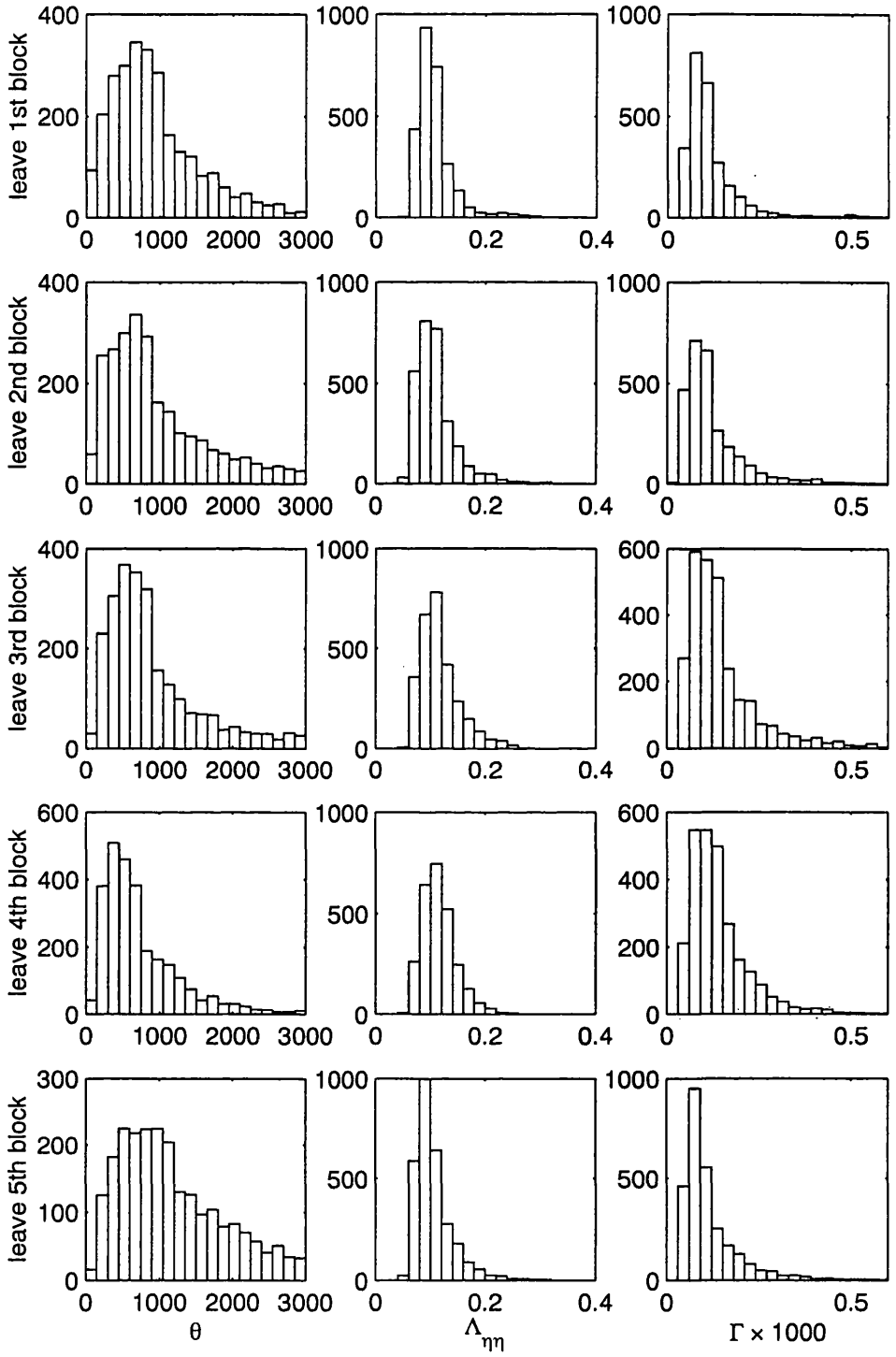

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
θ
$\Lambda_{\eta\eta}$
Γ × 1000

Figure B.6 (b)

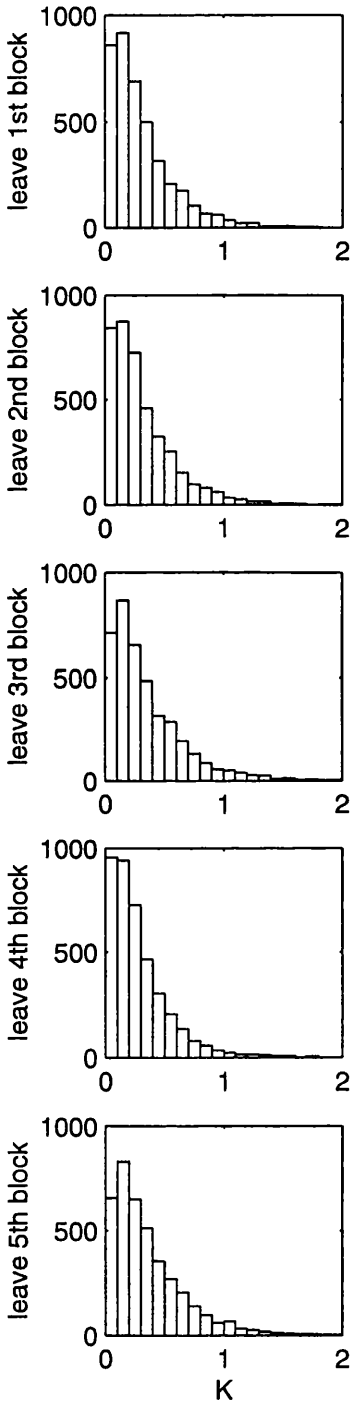

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
1000
500
0
0
1
2
K

Figure B.6 (c)

Figure B.7: Histograms of MCMC samples for the parameters in regression model M.b

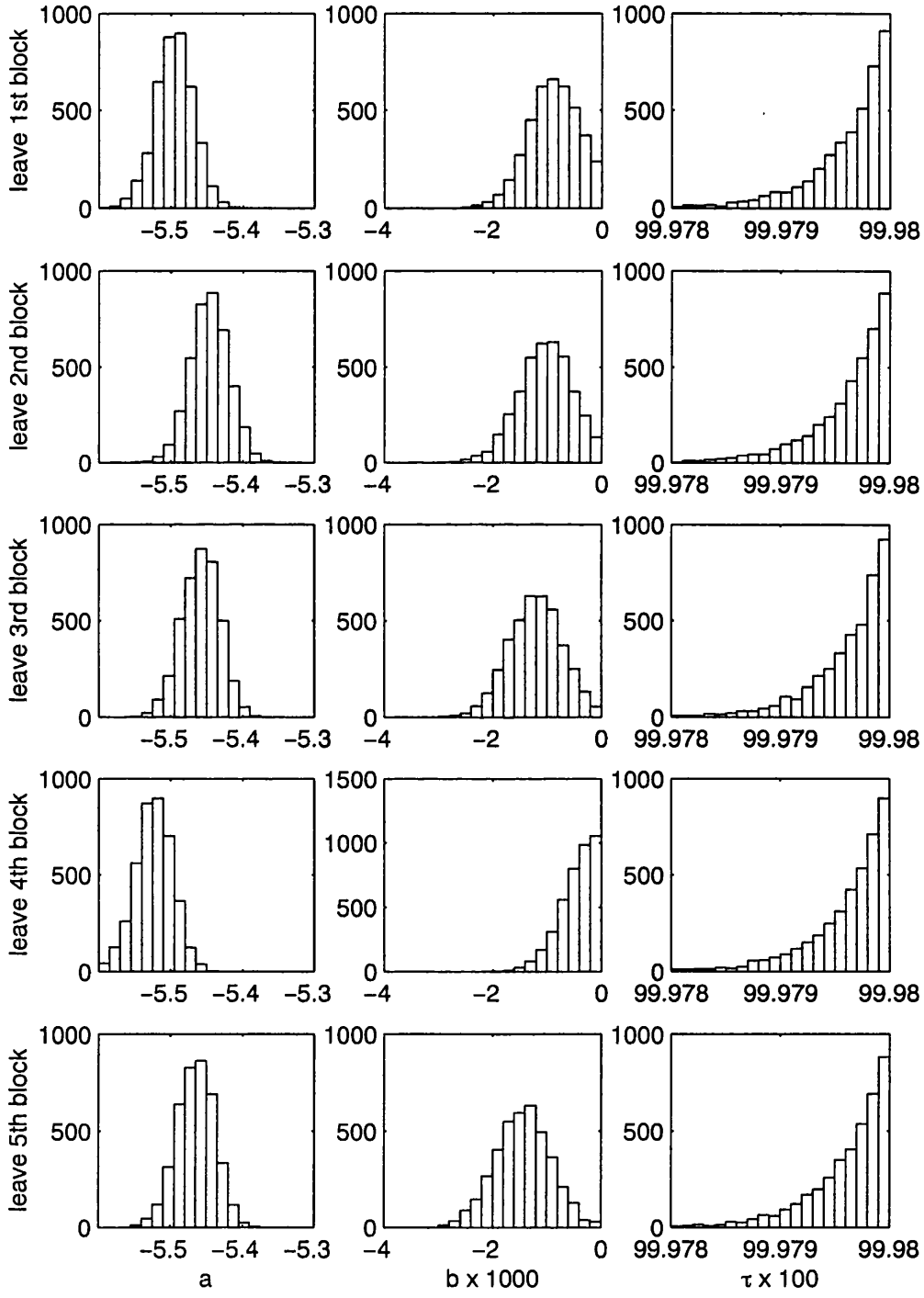


Figure B.7 (a)

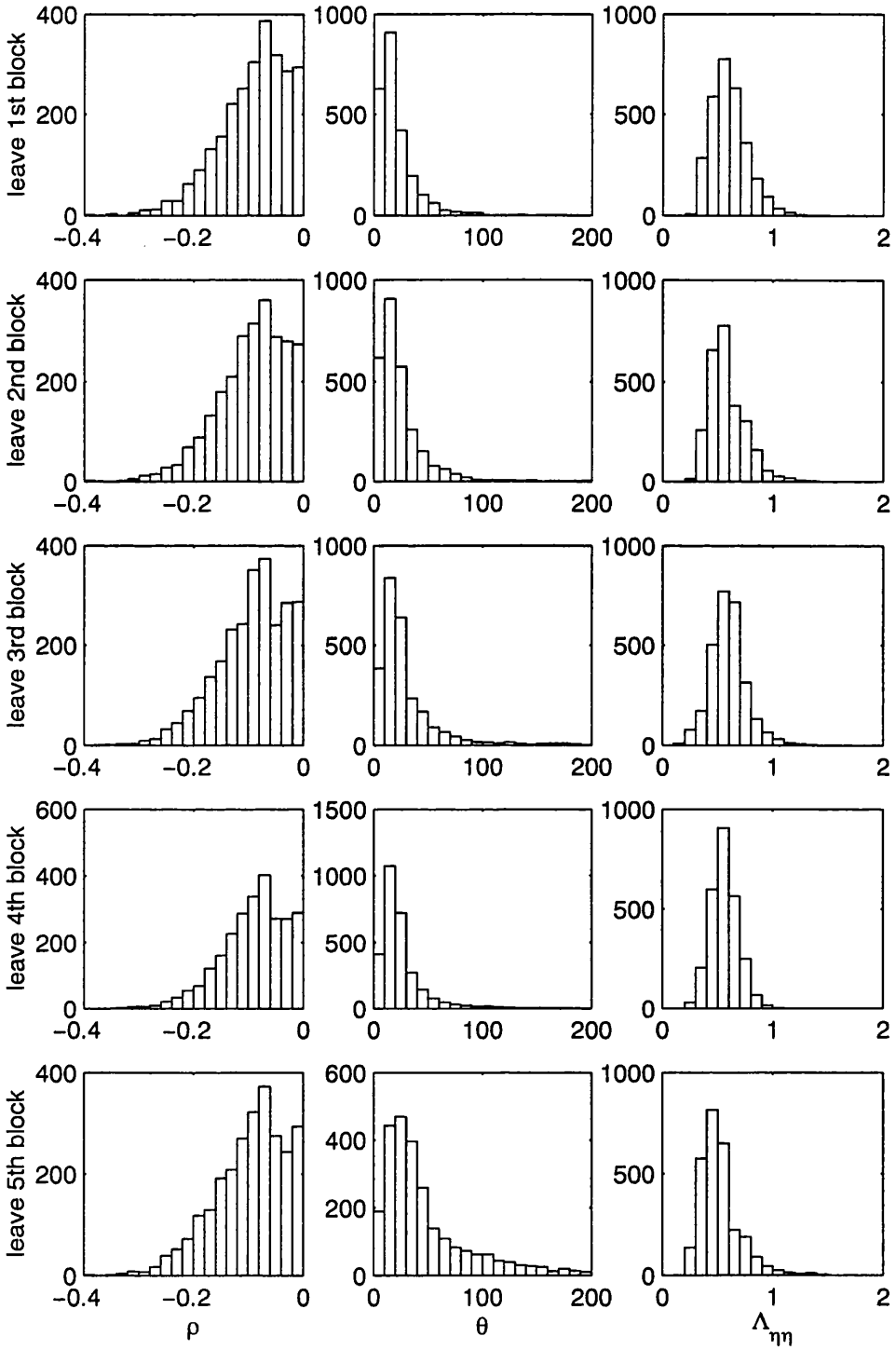

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
ρ
θ
$\Lambda_{\eta\eta}$

Figure B.7 (b)

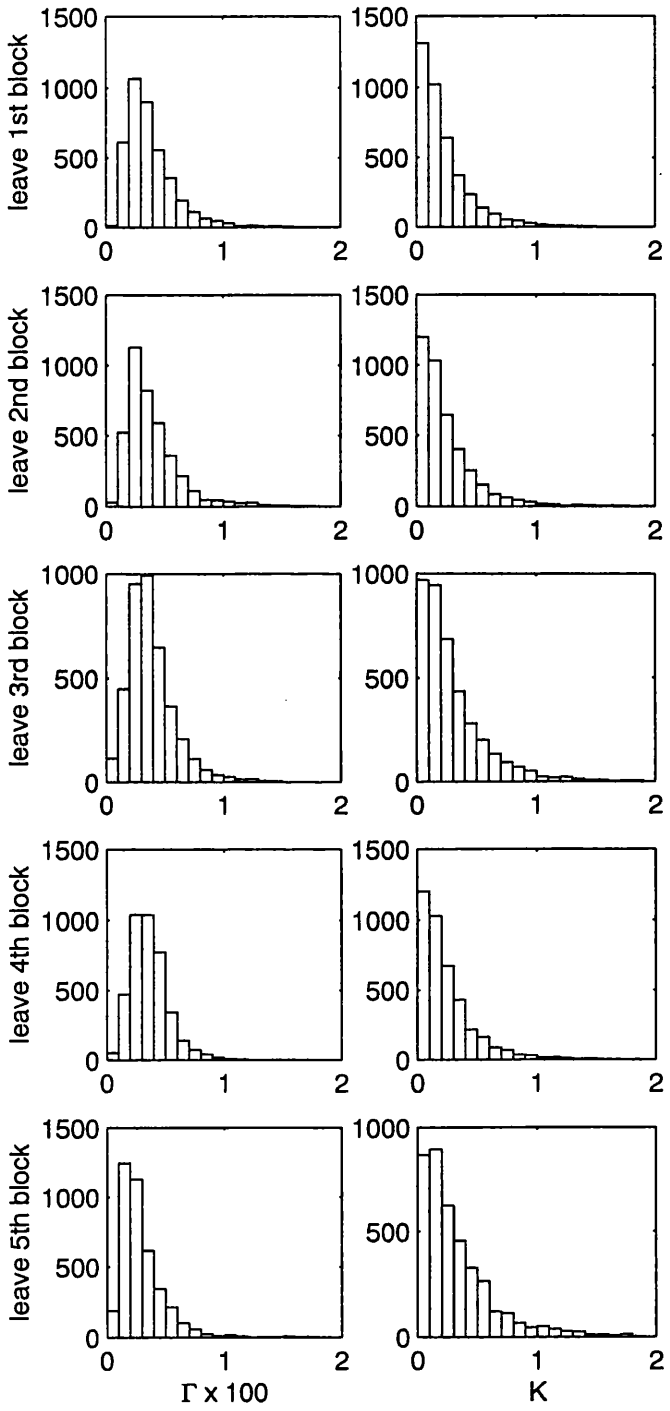

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
Γ x 100
K

Figure B.7 (c)

Figure B.8: Histograms of MCMC samples for the parameters in regression model M.c

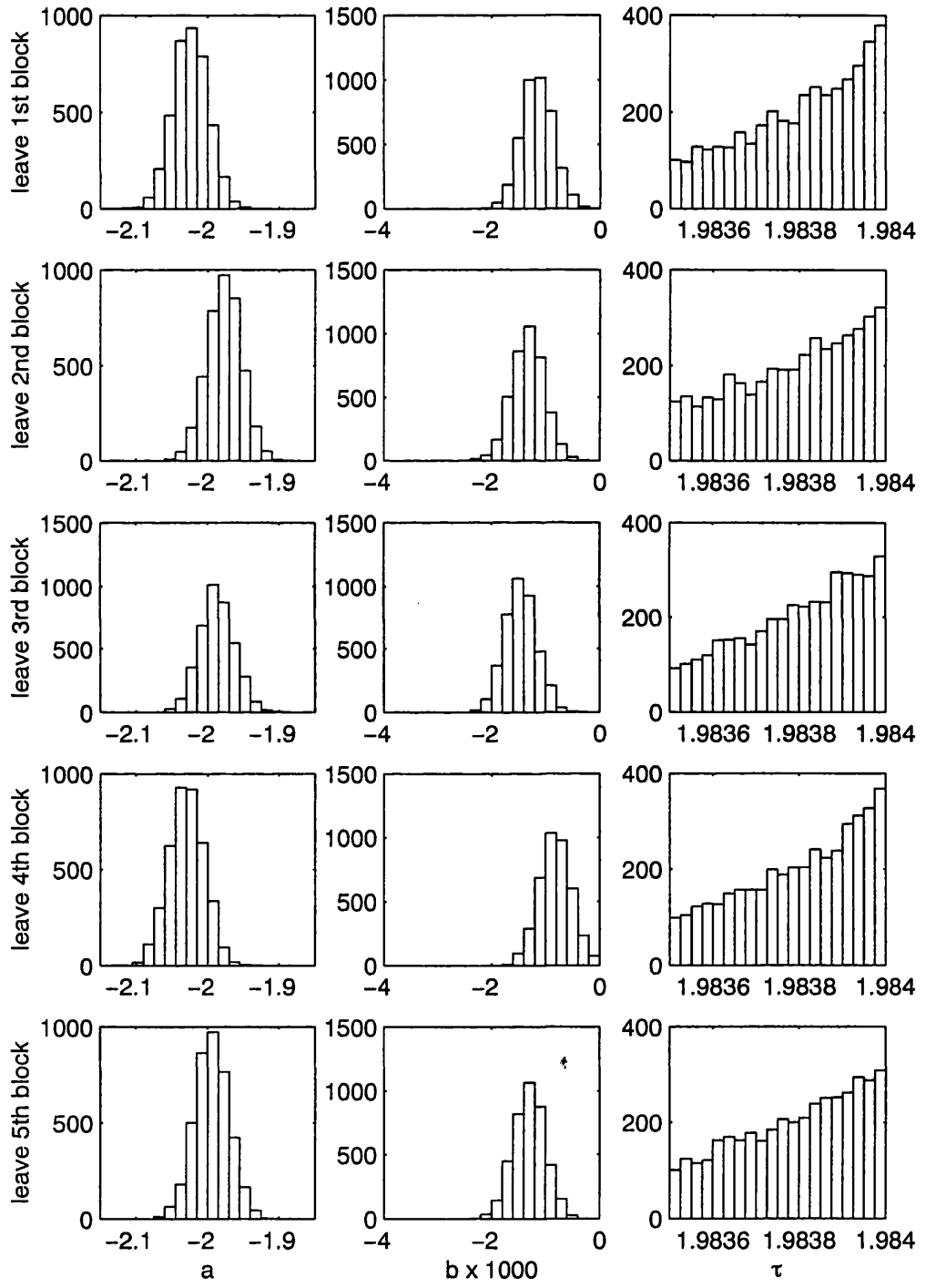


Figure B.8 (a)

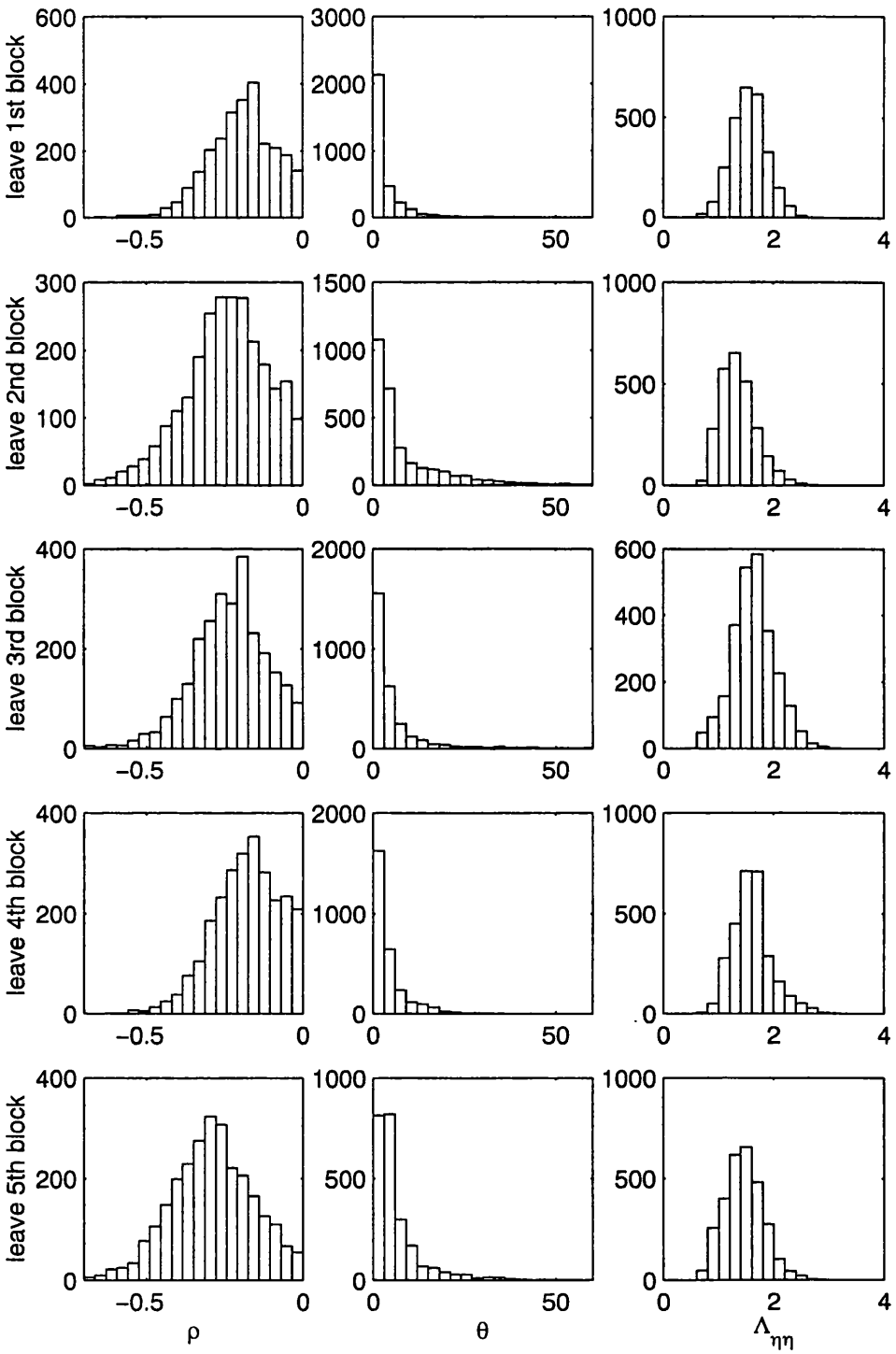

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
ρ
θ
$\Lambda_{\eta\eta}$

Figure B.8 (b)

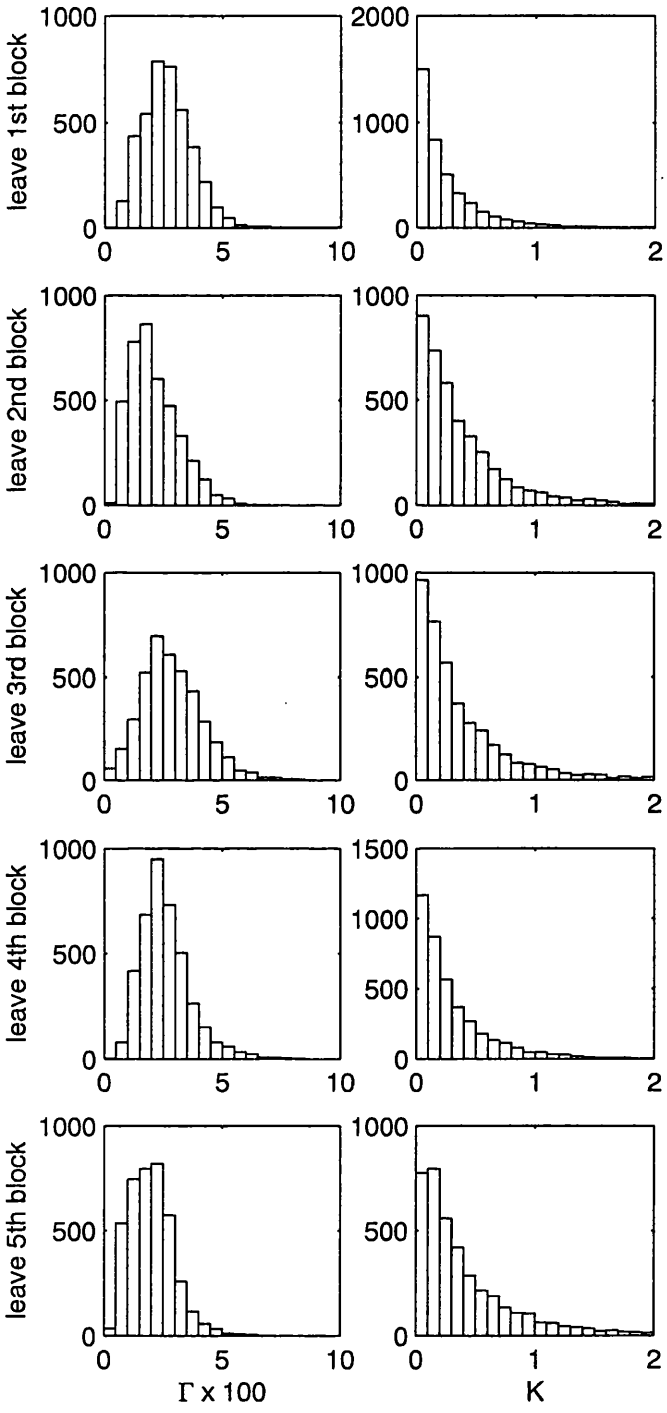

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
1000
500
0
2000
1000
1500
0
5
10
1
2
Γ x 100
K

Figure B.8 (c)

Figure B.9: Histograms of MCMC samples for the parameters in regression model M.d

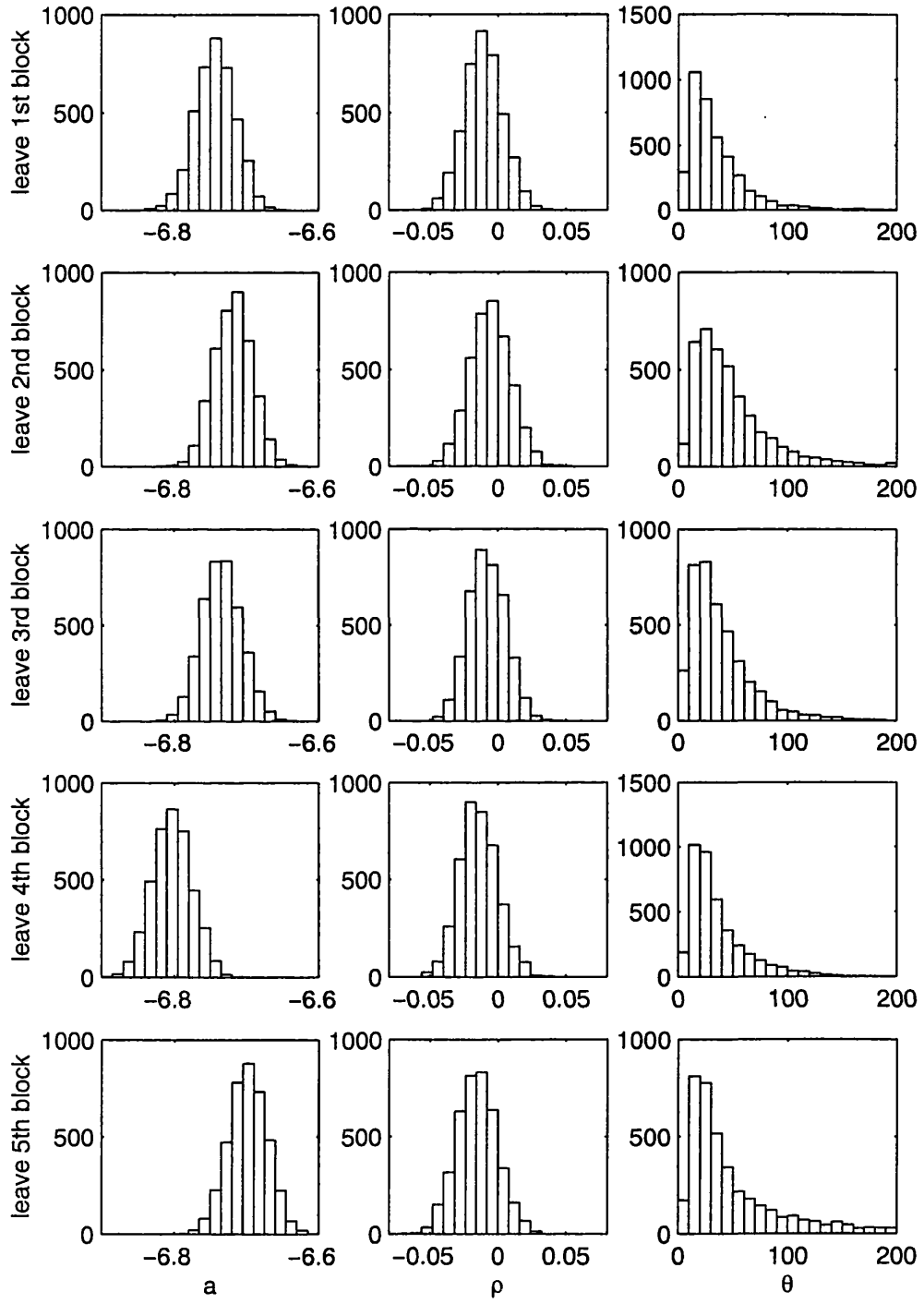


Figure B.9 (a)

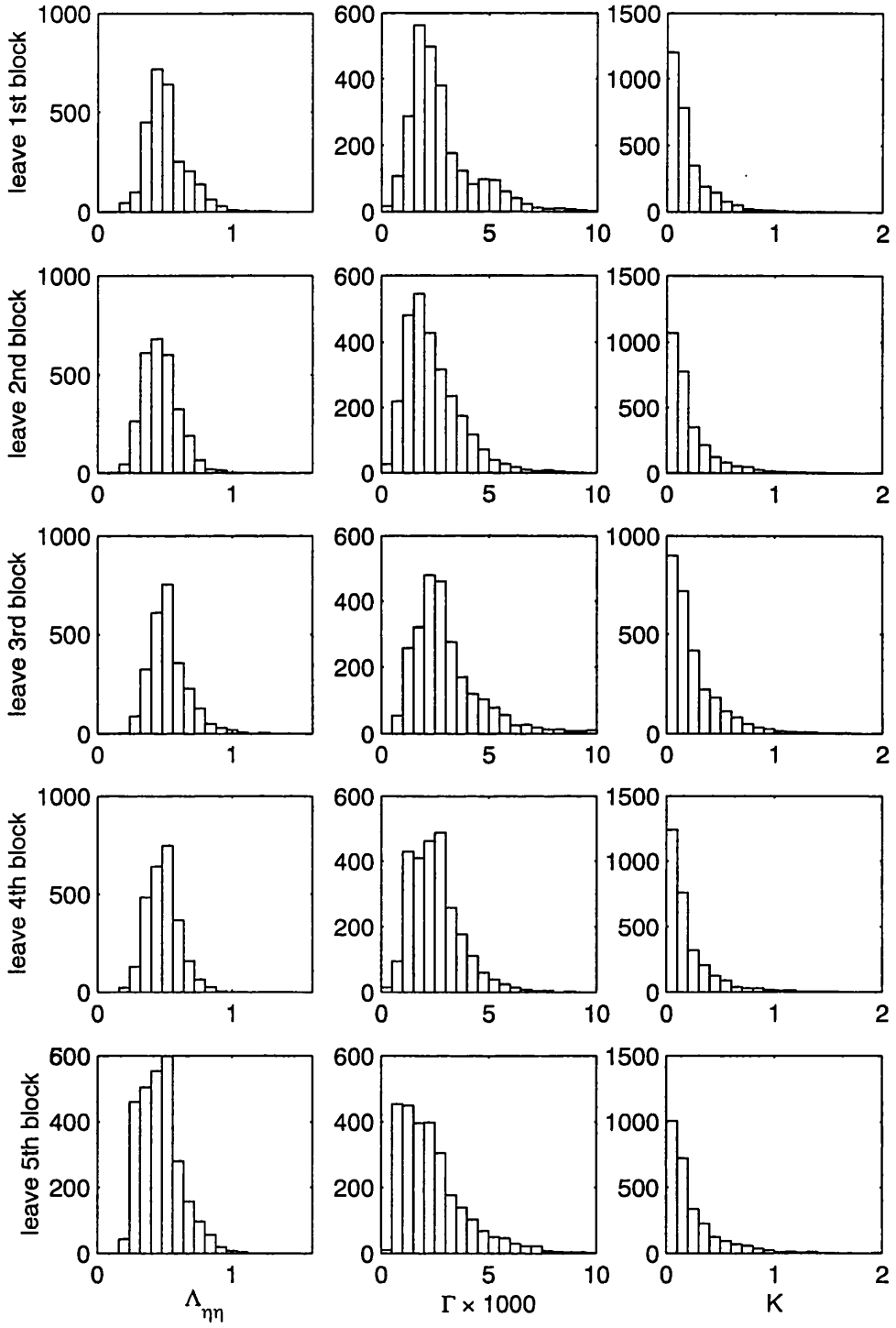

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
$\Lambda_{\eta\eta}$
$\Gamma \times 1000$
K

Figure B.9 (b)

Figure B.10: Histograms of MCMC samples for the parameters in regression model M.e

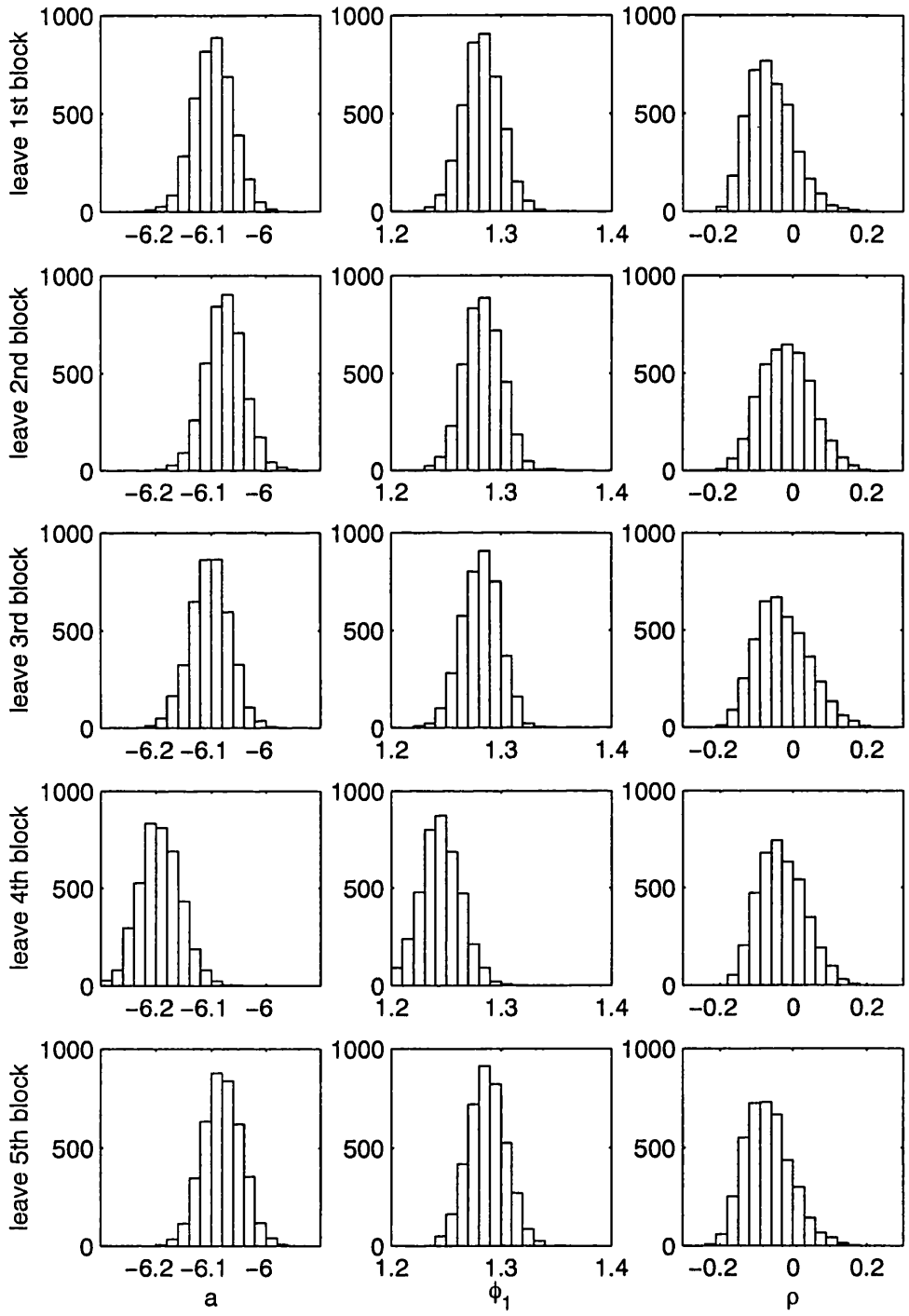


Figure B.10 (a)

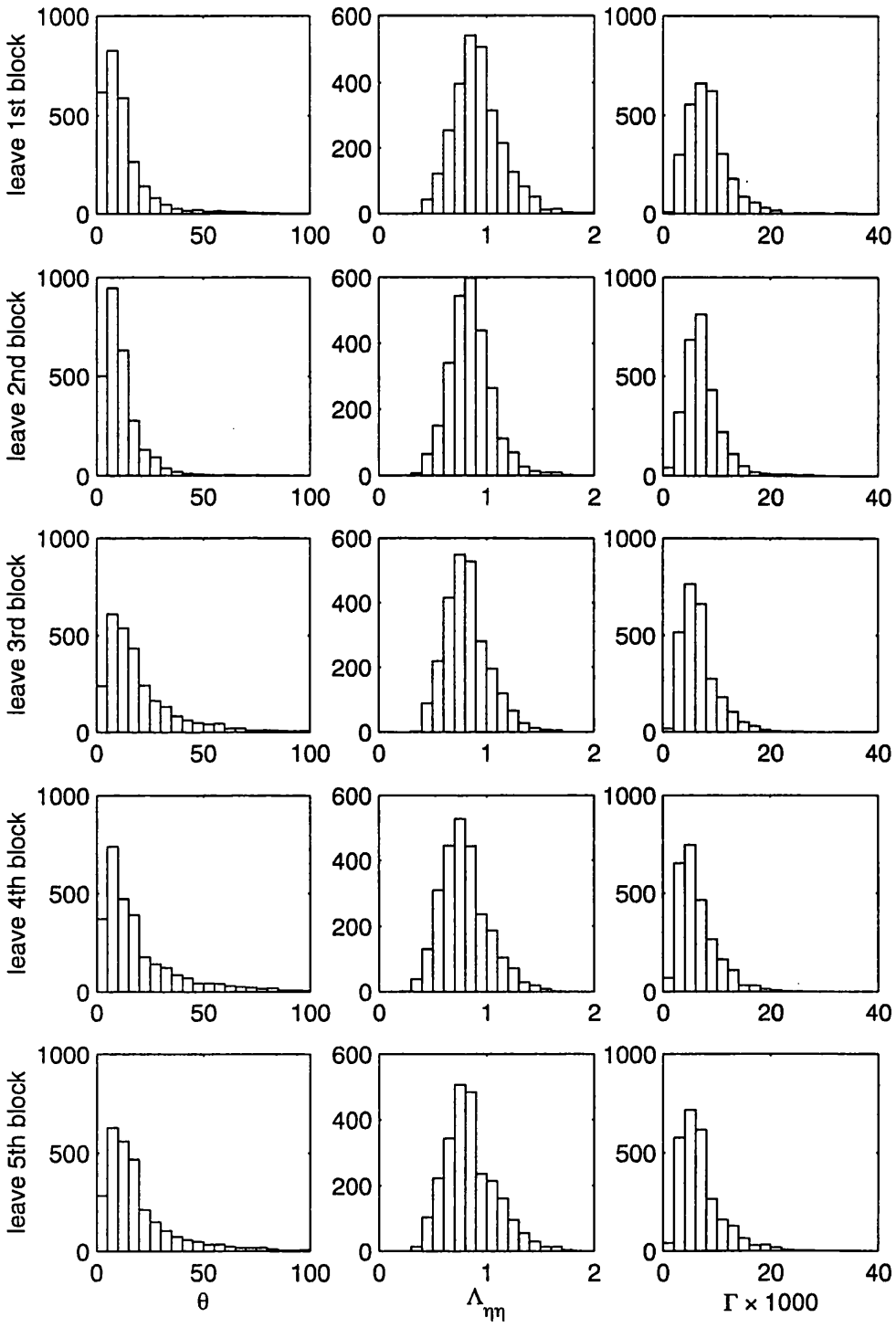

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
θ
$\Lambda_{\eta\eta}$
Γ × 1000

Figure B.10 (b)

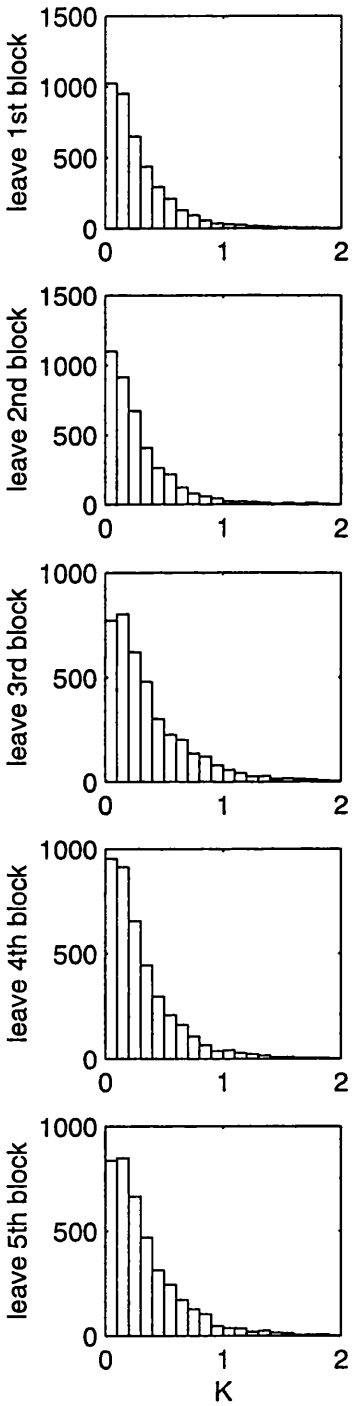

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
1500
1000
500
0
1
2
K

Figure B.10 (c)

Appendix C

Results for Regression Models for the Artificial Data

Figure C.1: Histograms of MCMC samples for the parameters in regression model M.a

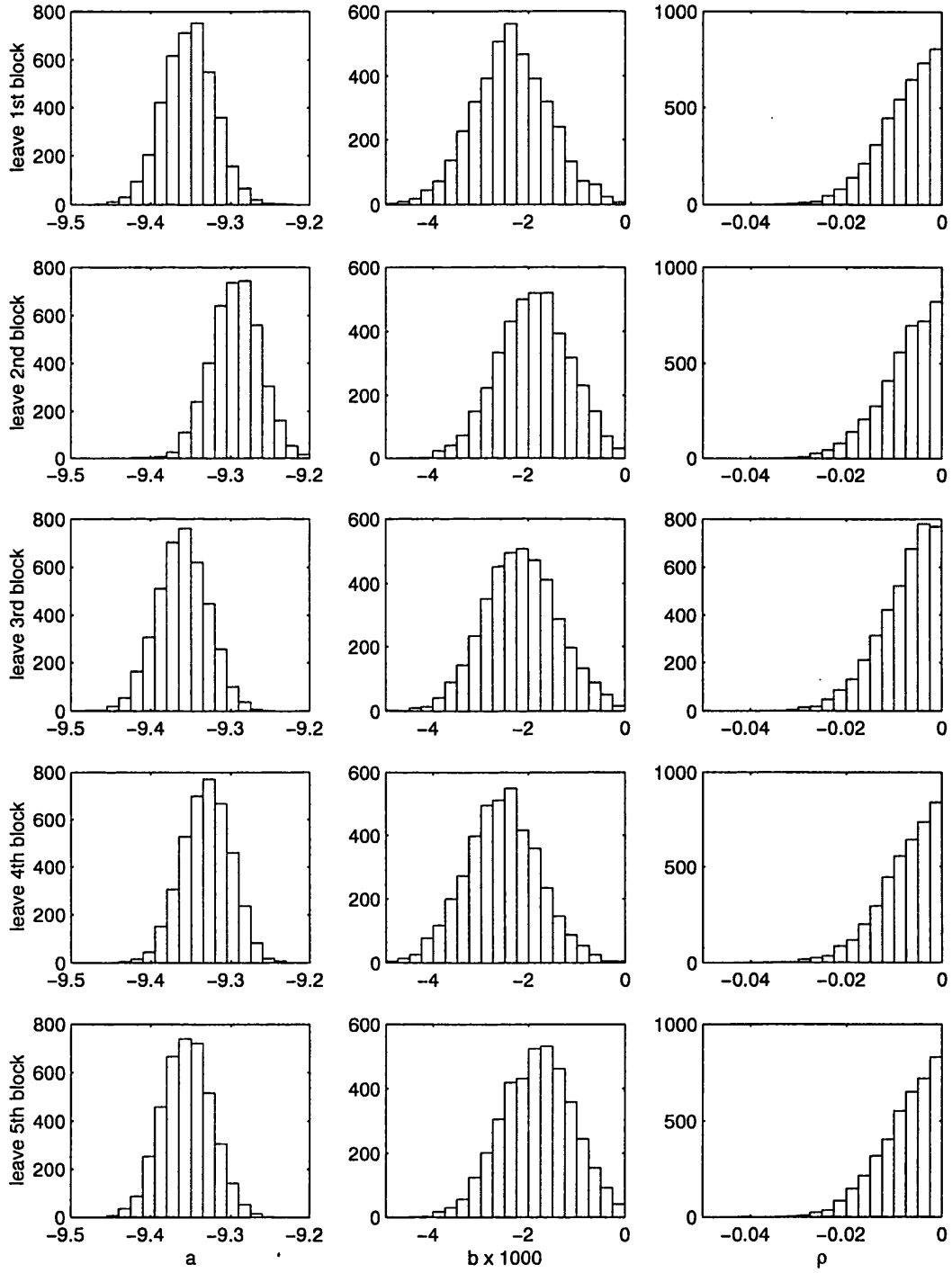


Figure C.1 (a)

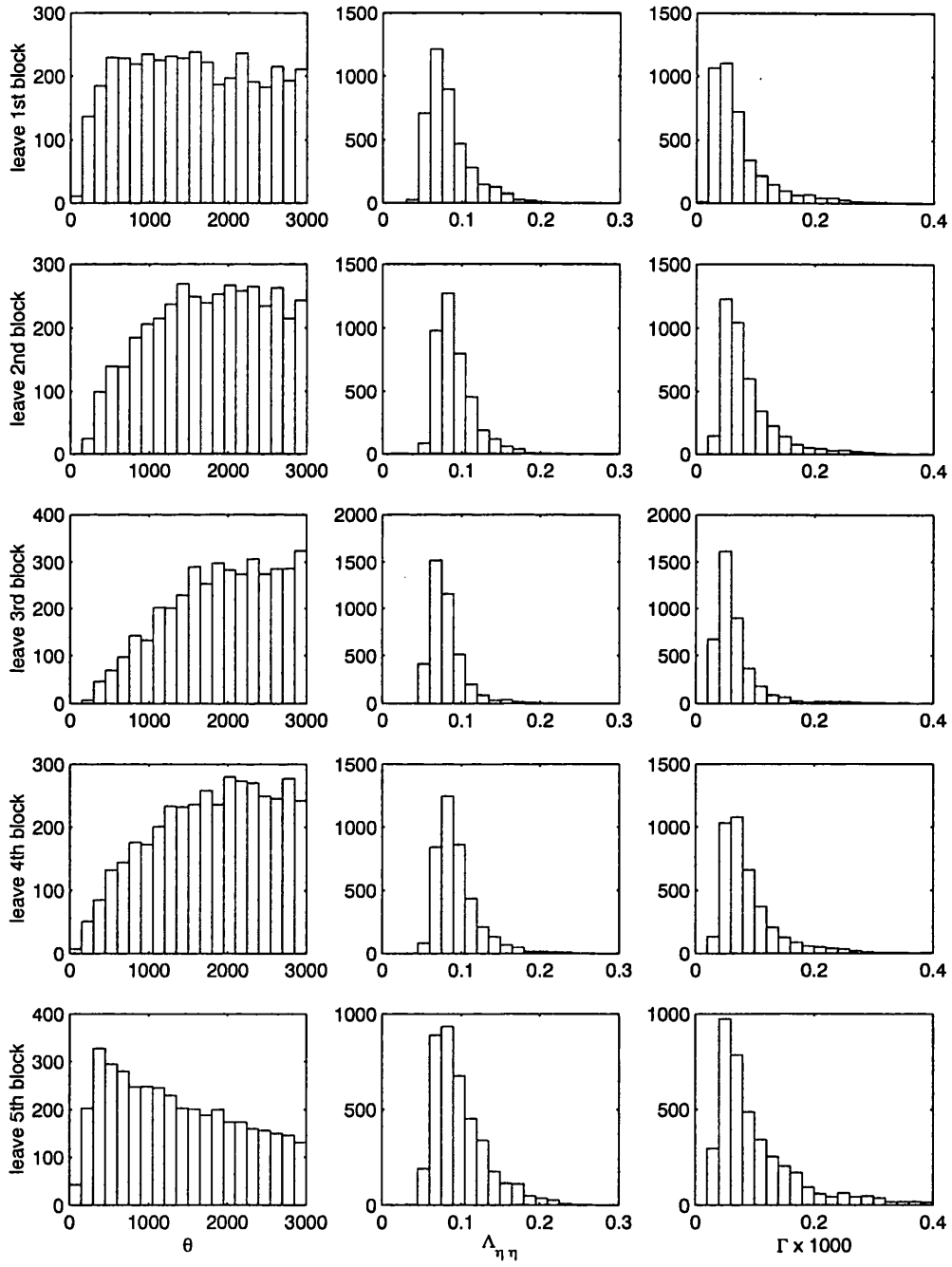

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
θ
$\Lambda_{\eta\eta}$
Γ x 1000

Figure C.1 (b)

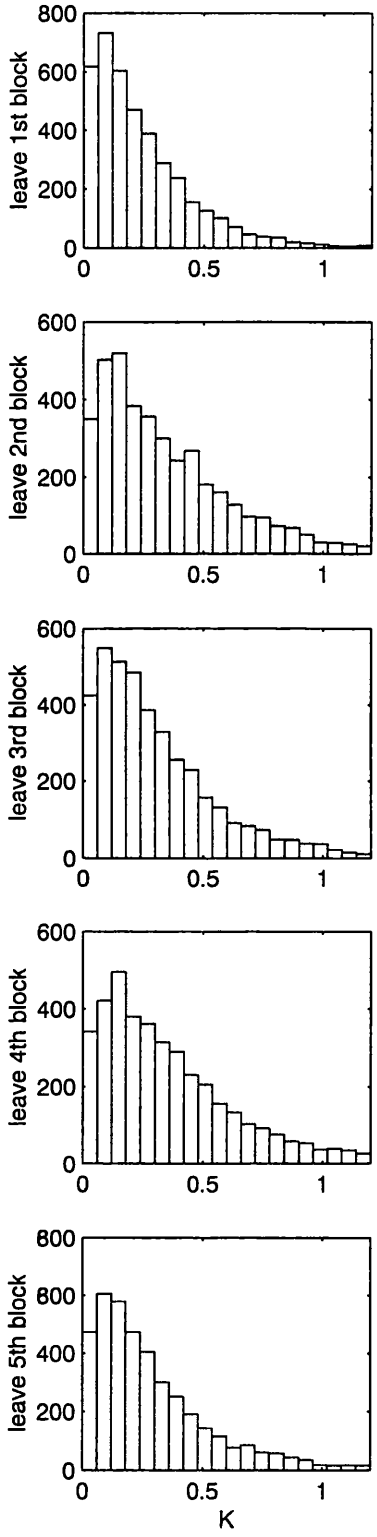

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
800
600
400
200
0
0.5
1
K

Figure C.1 (c)

Figure C.2: Histograms of MCMC samples for the parameters in regression model M.b

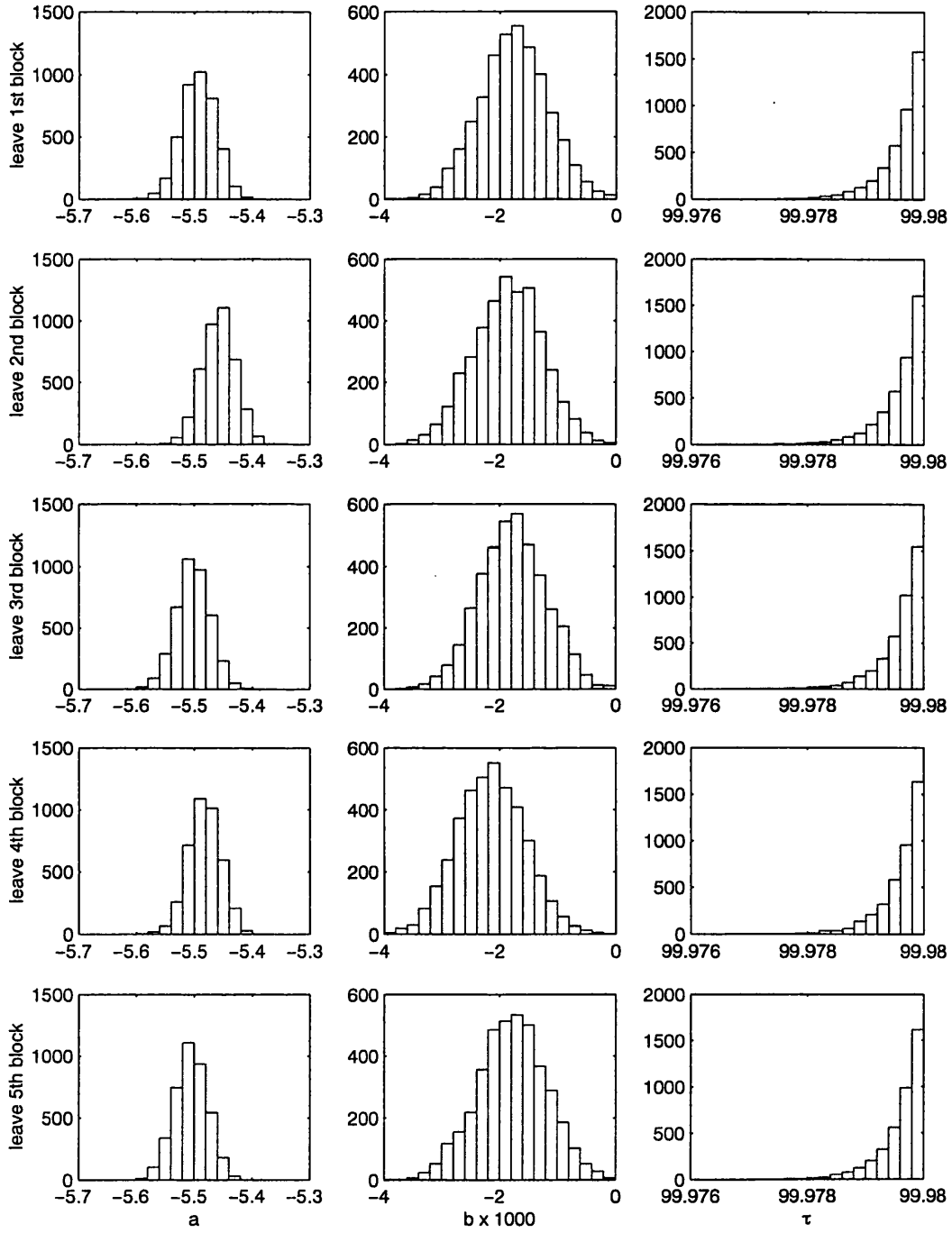


Figure C.2 (a)

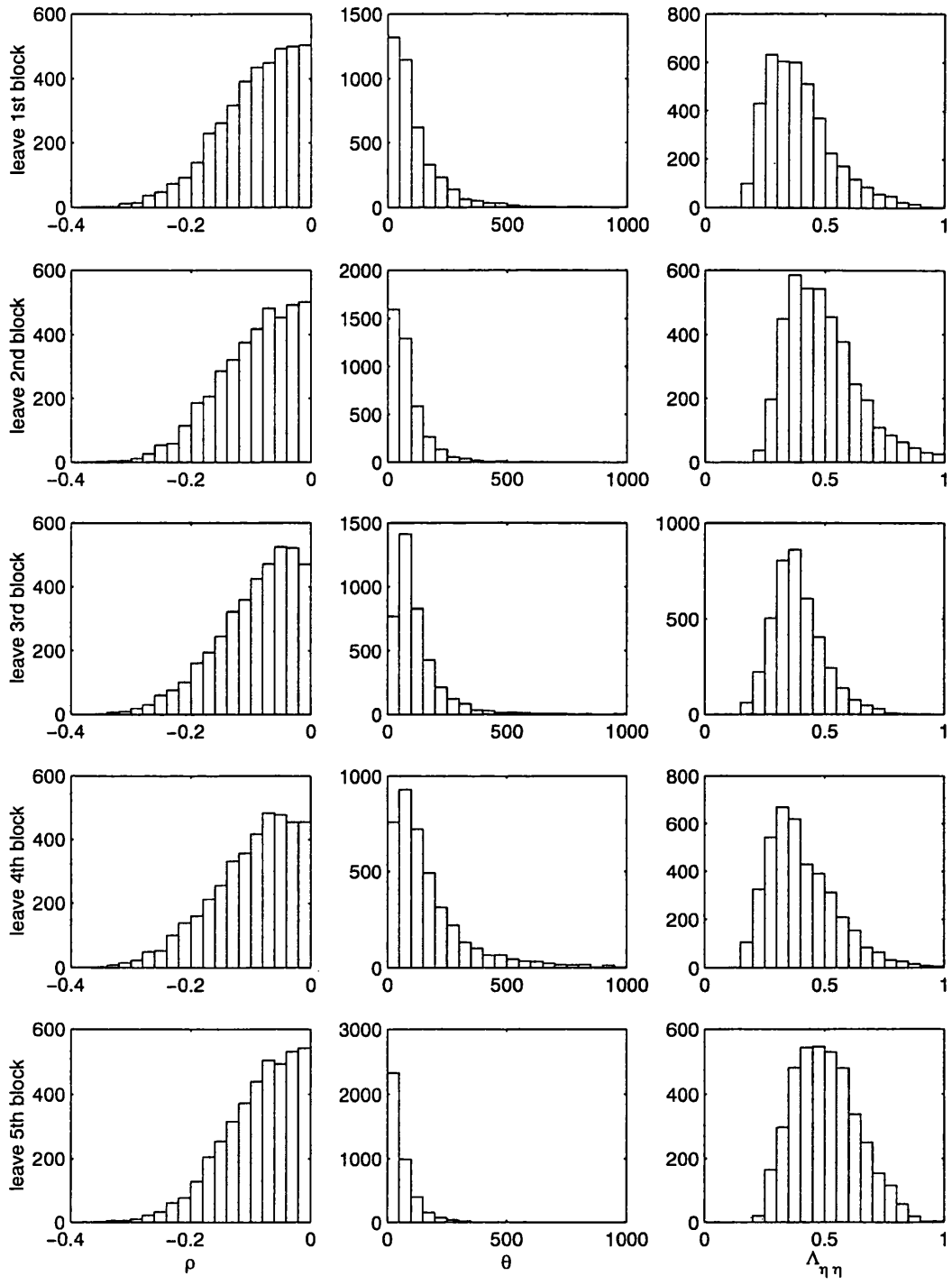

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
ρ
θ
$\Lambda_{\eta\eta}$

Figure C.2 (b)

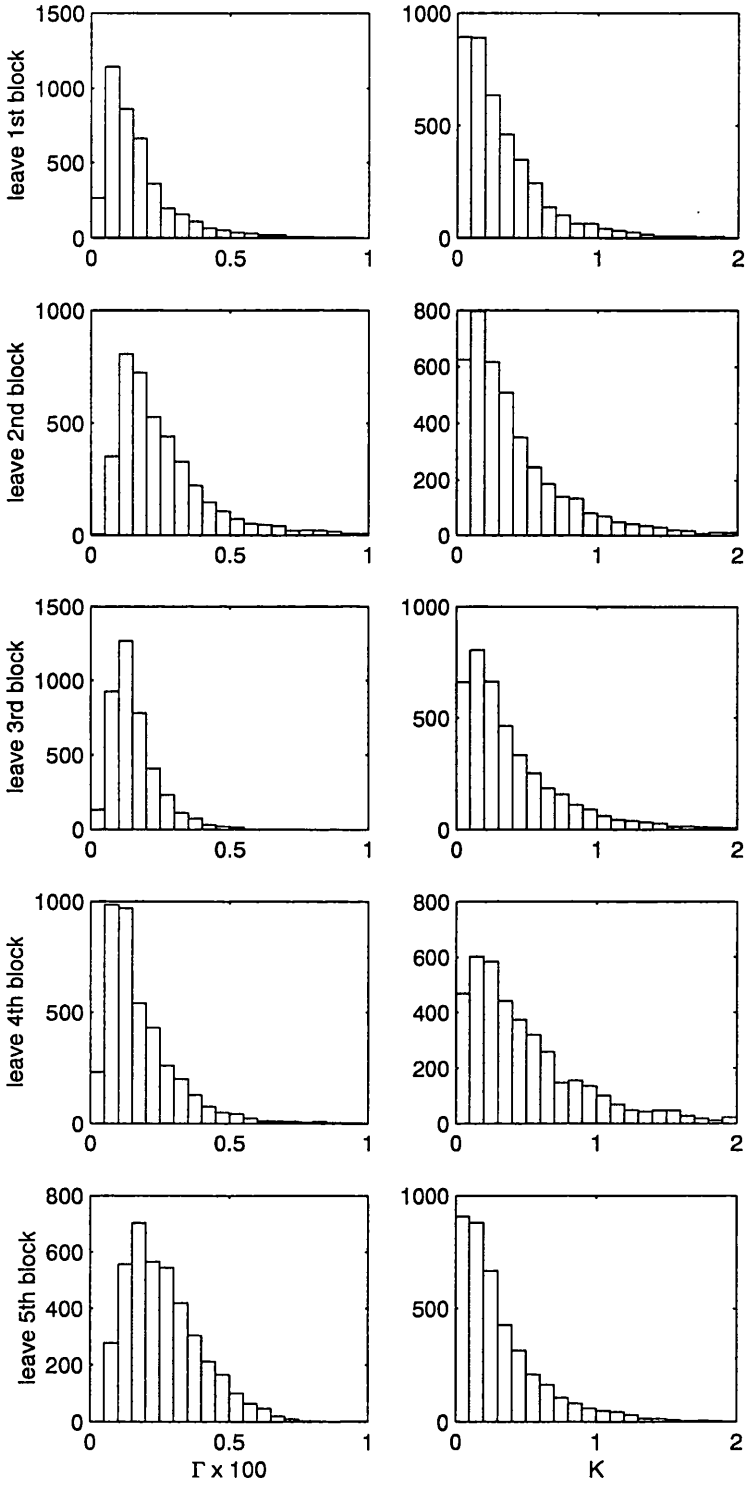

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
Γ x 100
K

Figure C.2 (c)

Figure C.3: Histograms of MCMC samples for the parameters in regression model M.c

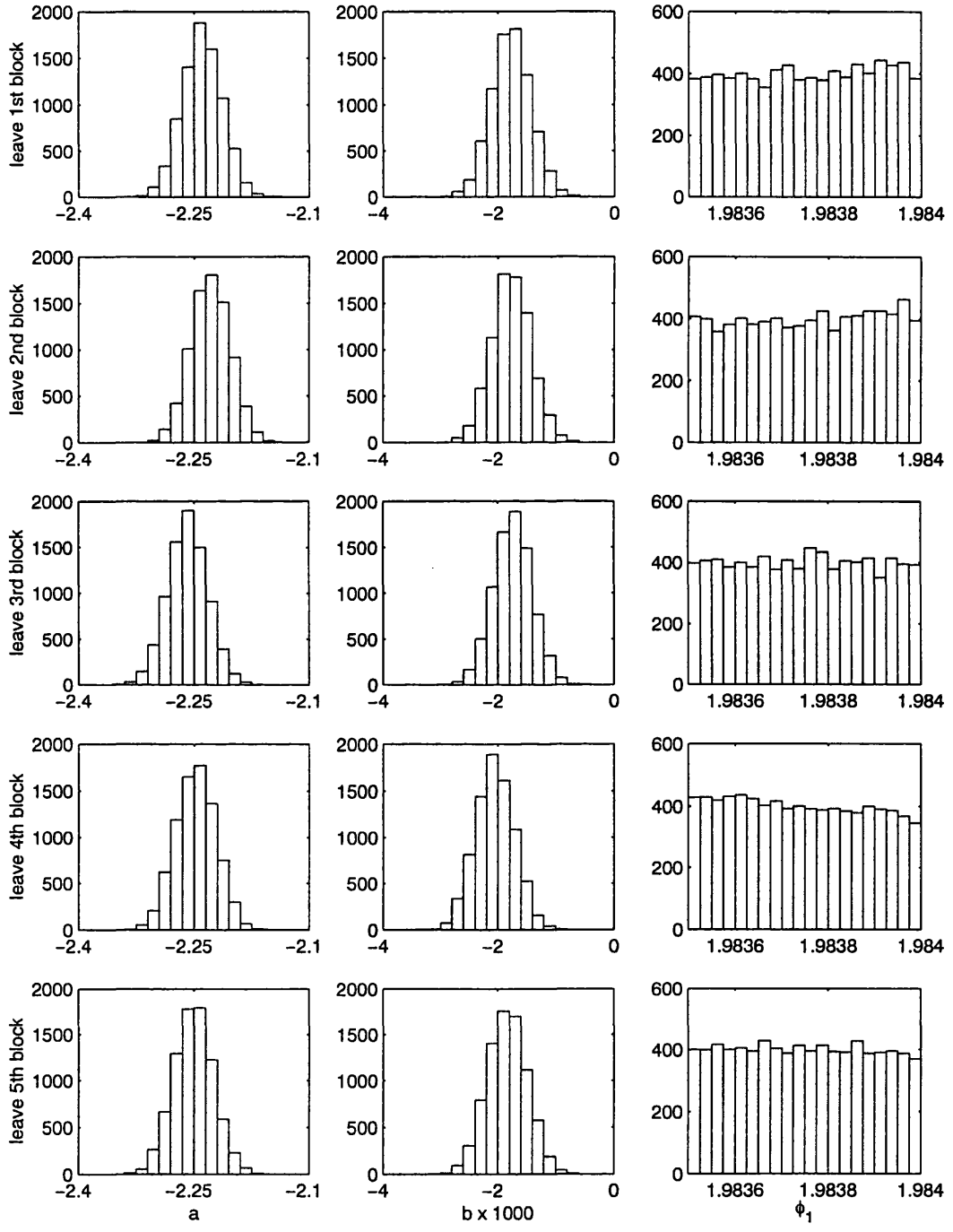


Figure C.3 (a)

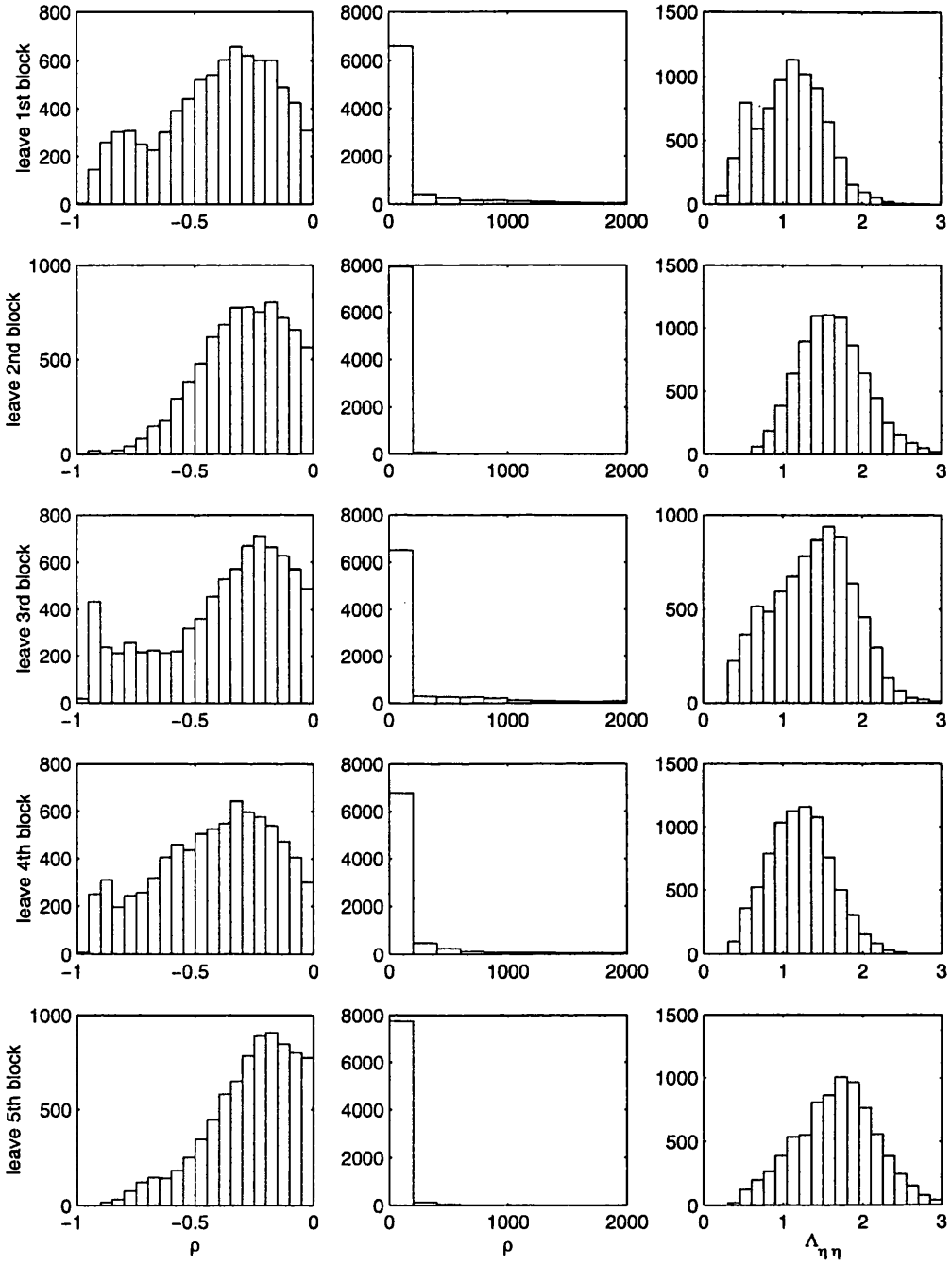

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
ρ
ρ
$\Lambda_{\eta\eta}$

Figure C.3 (b)

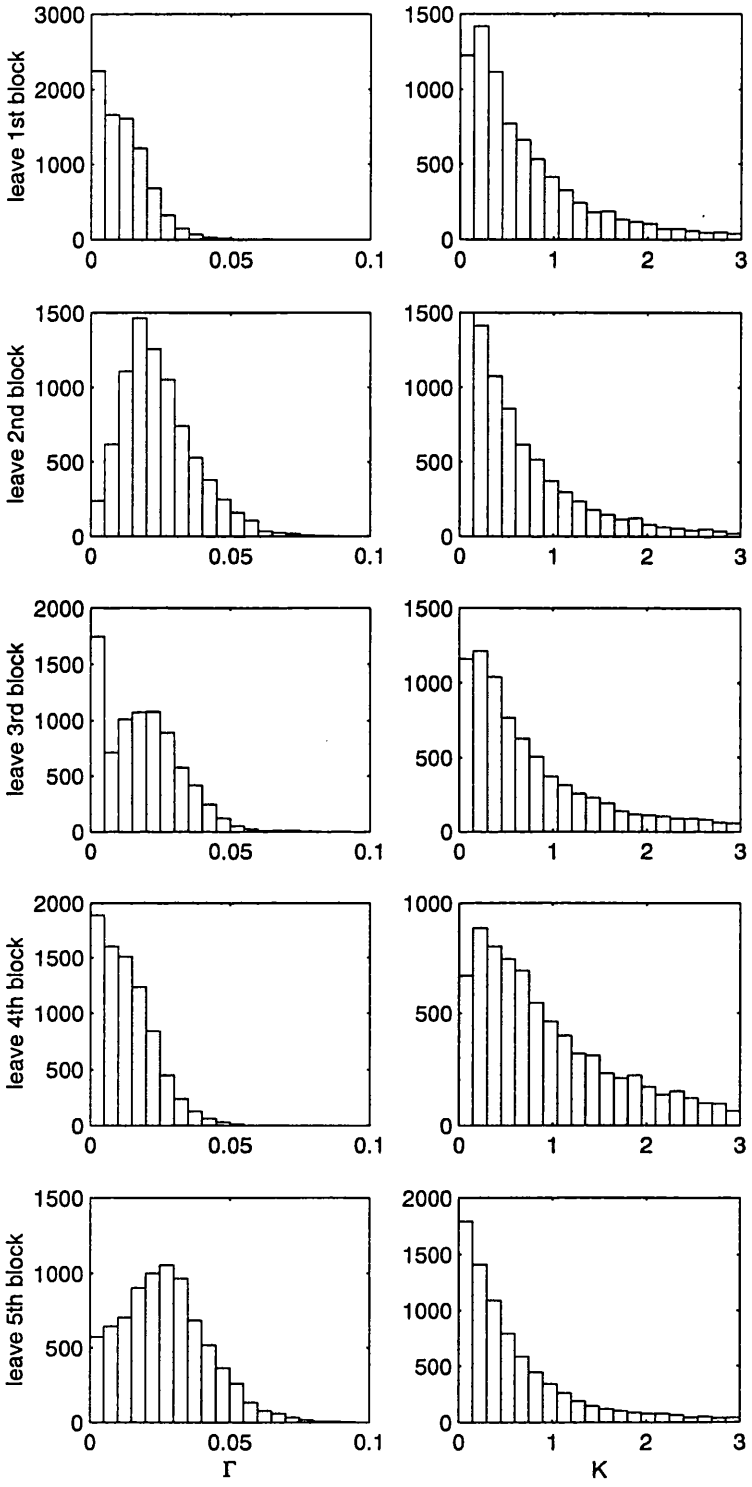

leave 1st block
leave 2nd block
leave 3rd block
leave 4th block
leave 5th block
Γ
K

Figure C.3 (c)

Appendix D

Results for Discriminant Analysis

Table D.1: Variance ratios for M.a, M.b and M.c

The figures in the row of Variance ratio (VR) method 1 are the $\sqrt{\widehat{R}_G}$; in the row of VR method 2 are the $\sqrt{\widehat{R}_C}$; in the row of VR method 3 are the $\widehat{R}_{interval}$ with $\alpha = 0.05$.

		M.a		M.b			M.c		
δ	VR method	a	b	a	b	τ	a	b	ϕ_1
3	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	0.99	1.01	1.01	0.99	1.00	0.99	1.00
250	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	1.00	1.01	1.00	1.00	1.00	1.00	1.00
500	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.01	1.00	0.99	1.00	0.99	1.00	1.00	1.00

Figure D.1: MCMC output of discriminant analysis M.a

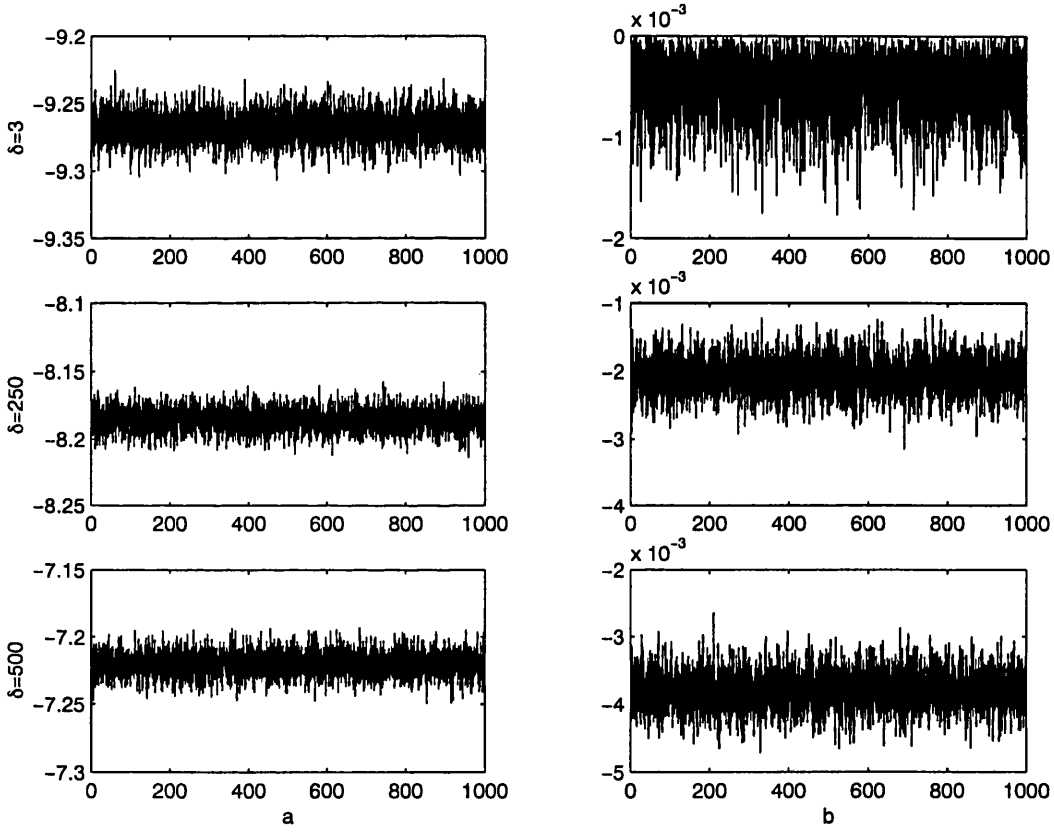


Figure D.2: MCMC output of discriminant analysis M.b

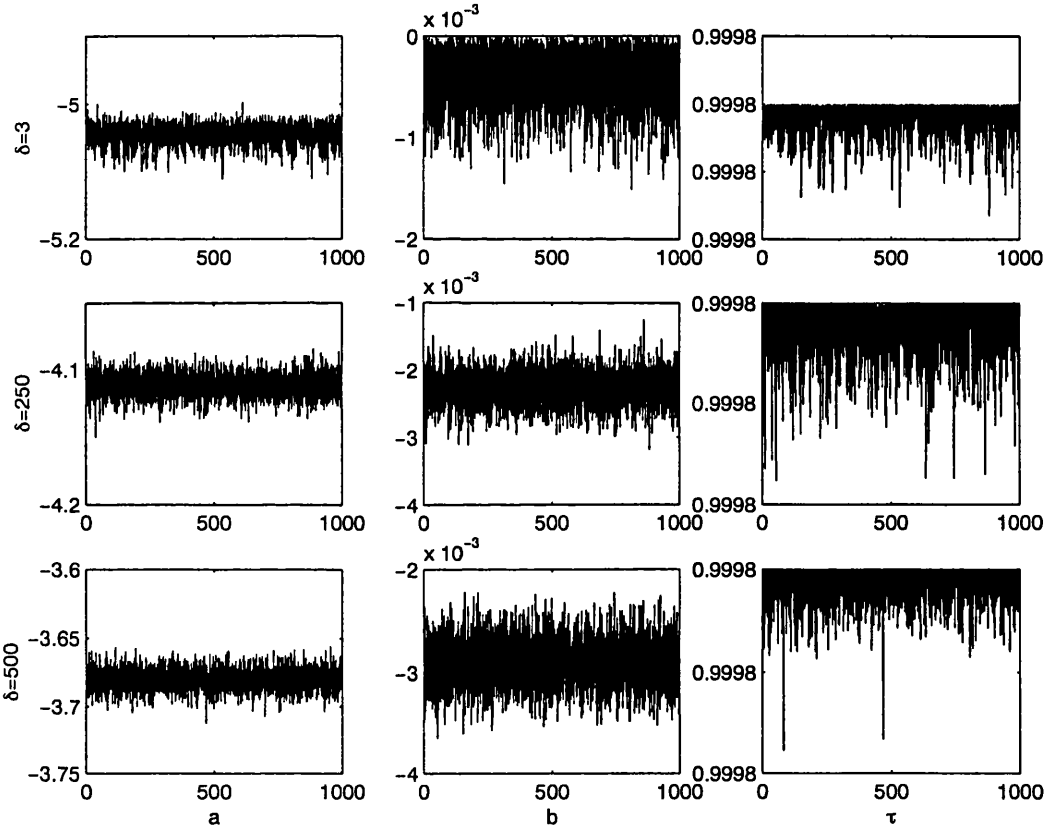


Figure D.3: MCMC output of discriminant analysis M.c

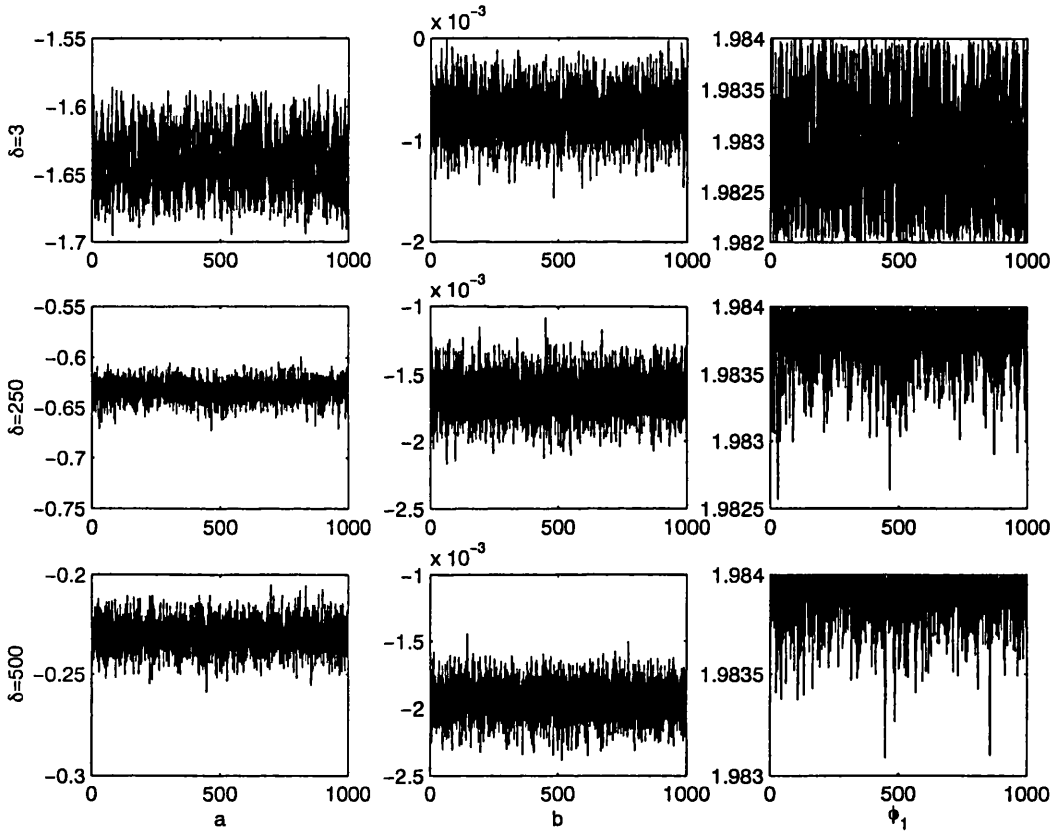


Figure D.4: Histograms of samples for the parameters in discriminant analysis M.a

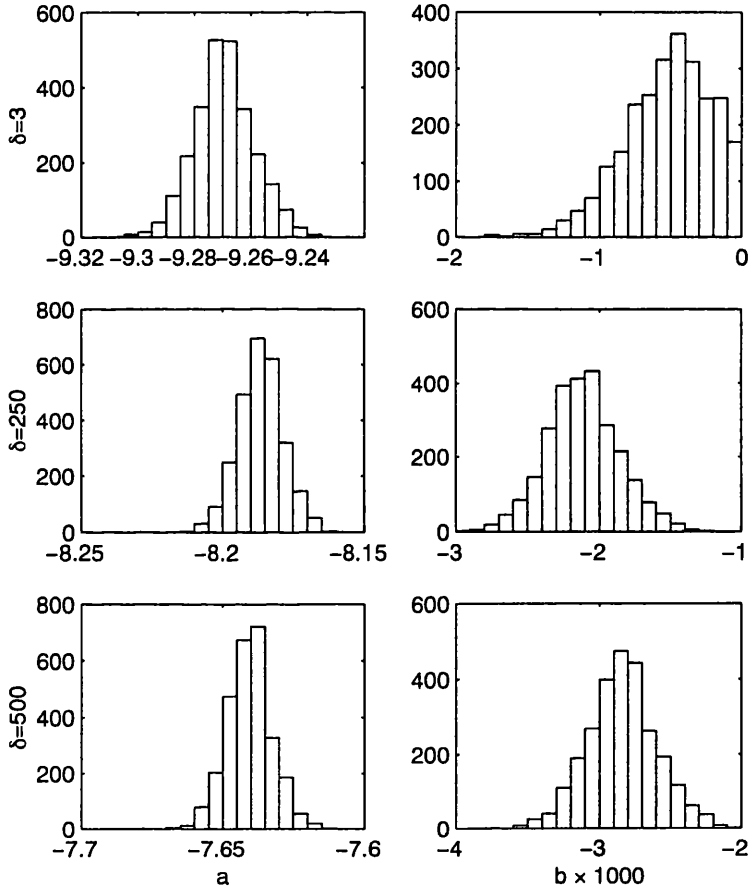


Figure D.5: Histograms of samples for the parameters in discriminant analysis M.b

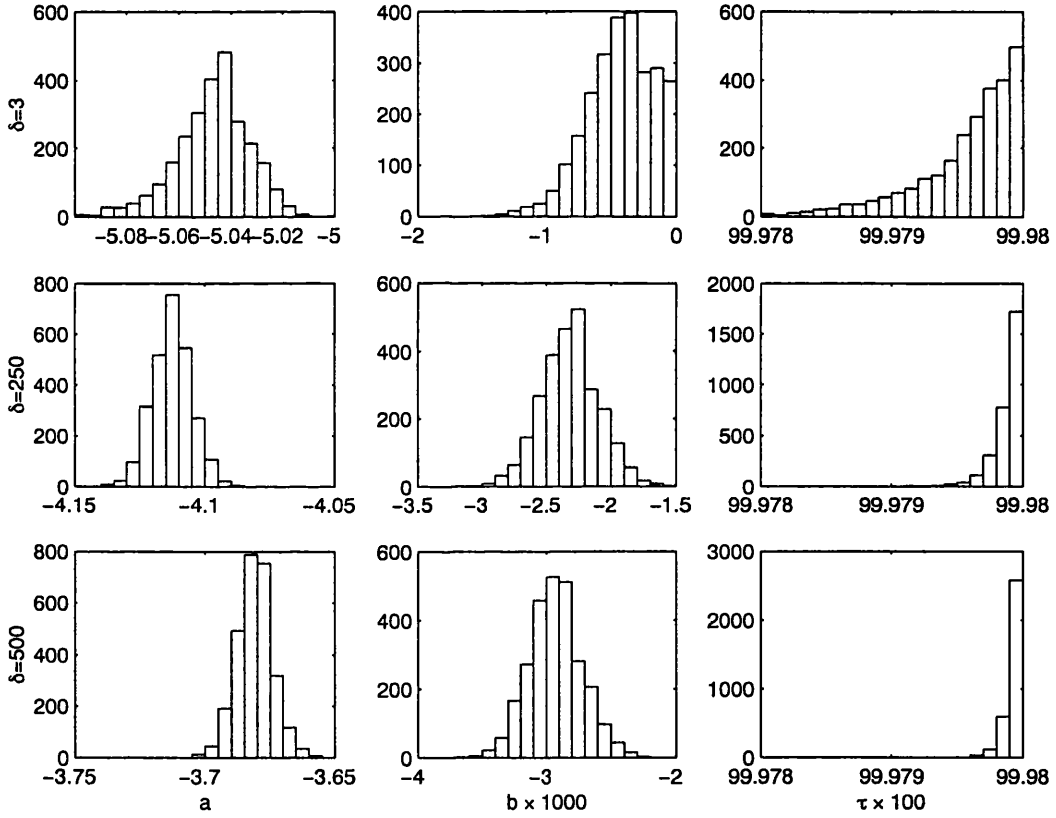
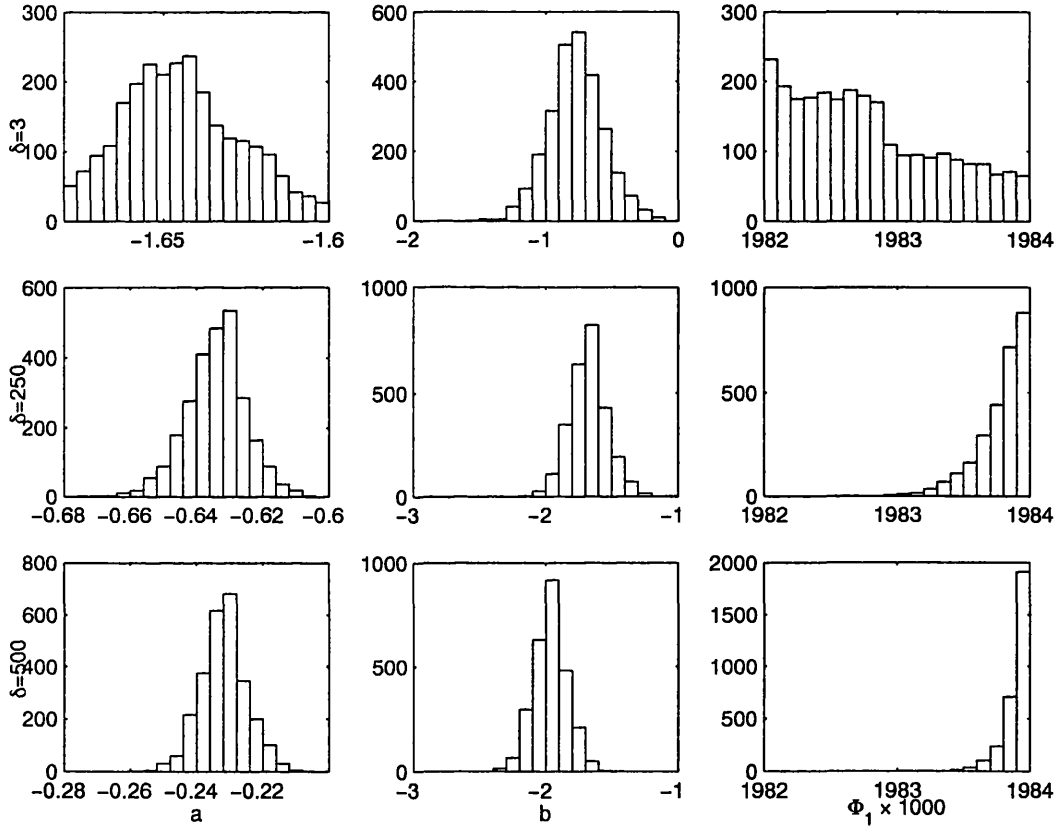


Figure D.6: Histograms of samples for the parameters in discriminant analysis M.c



Note: The two tails of the histogram for a at $\delta = 3$ are not shown, but the posterior density of a for $\delta = 3$ is not a truncated density function.

Appendix E

$(1+\theta)Q_{xx}$, ${X_t}^t X_t$, $(1+\theta)Q_{x\eta}$, and ${X_t}^t Y_t$

Figure E.1: $X_t{}^t X_t$ for the original spectra

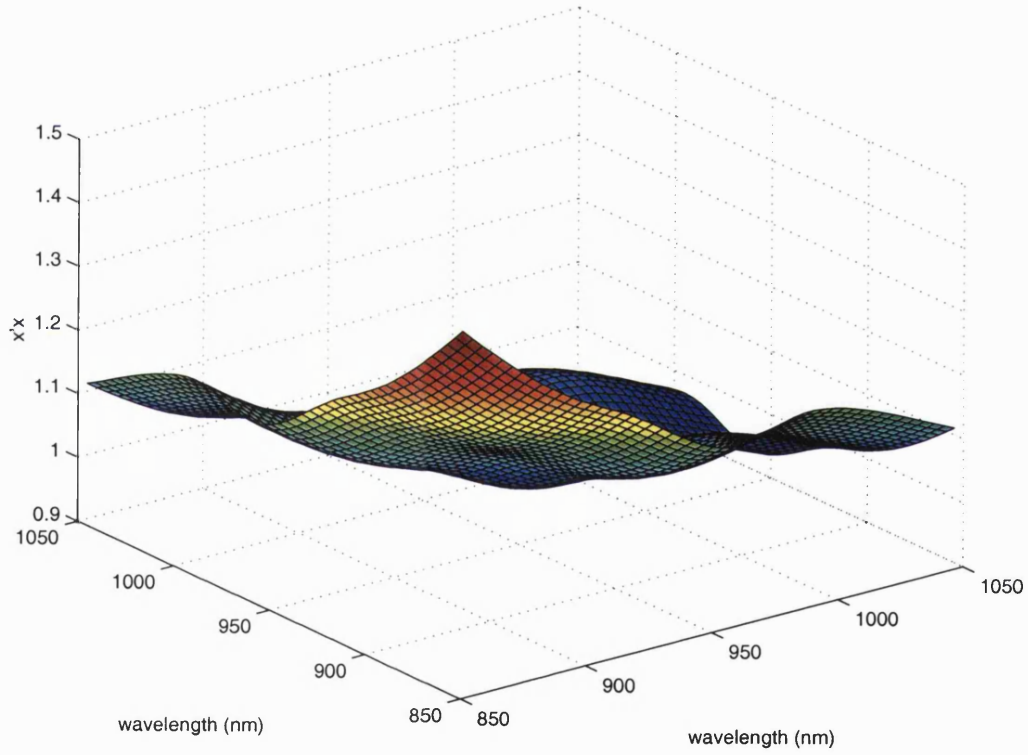


Figure E.2: MCMC estimate of the mean marginal $(1+\theta)Q_{xx}$ for M.a

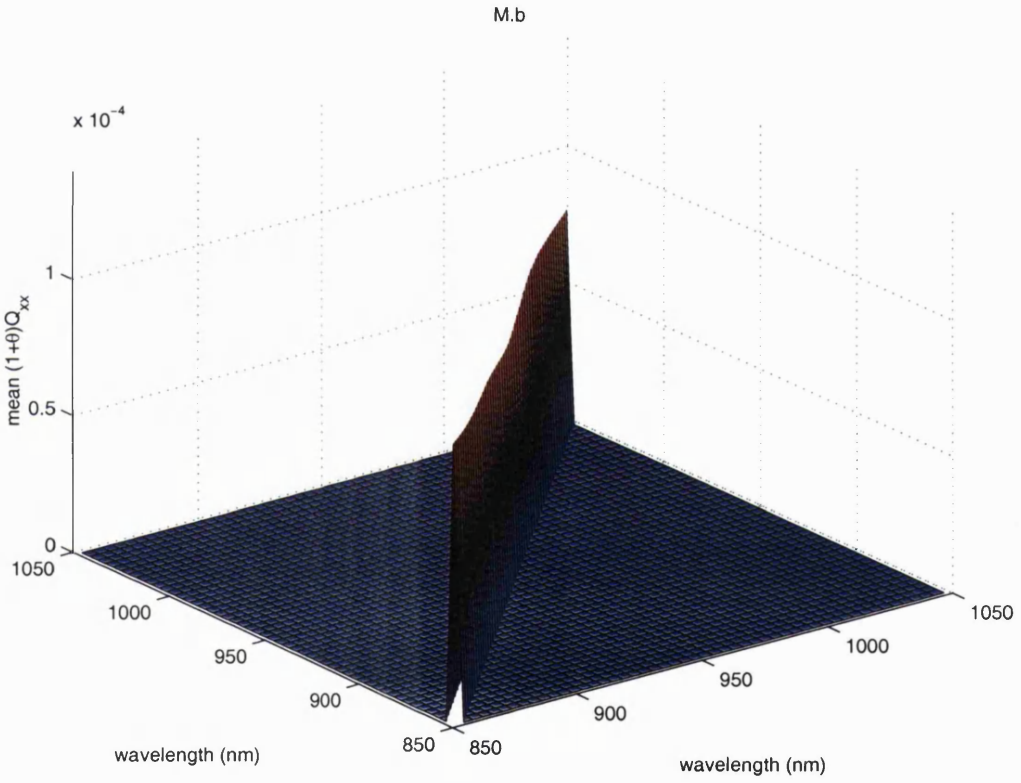


Figure E.3: MCMC estimate of the mean marginal $(1+\theta)Q_{xx}$ for M.b

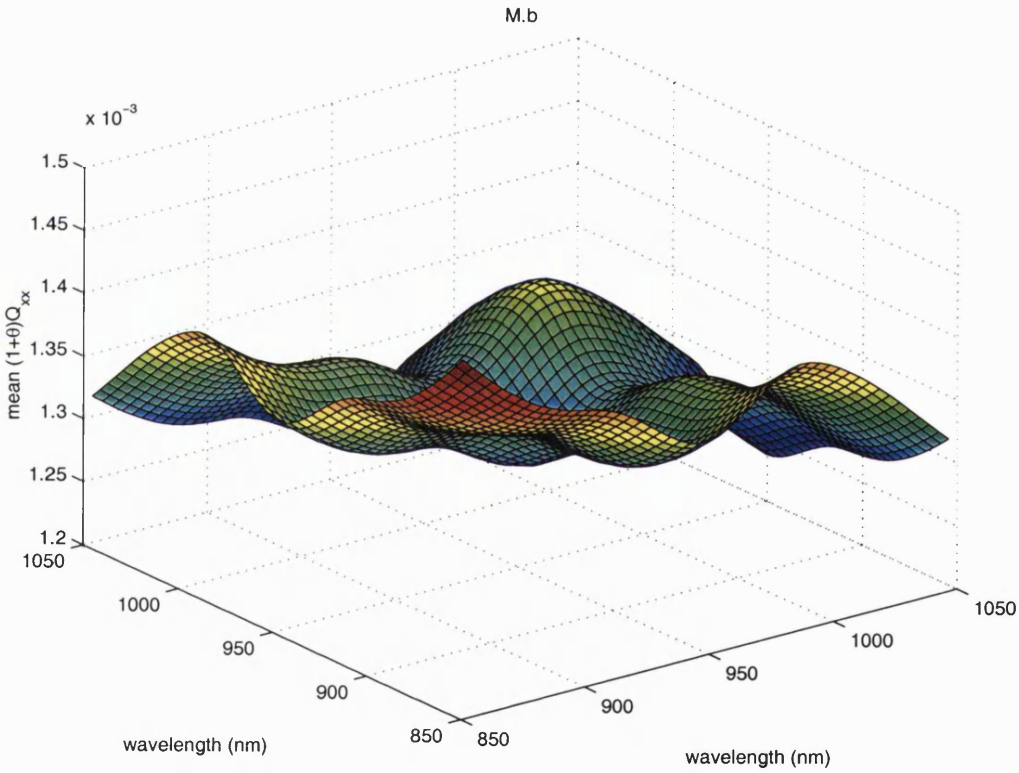


Figure E.4: MCMC estimate of the mean marginal $(1+\theta)Q_{xx}$ for M.c

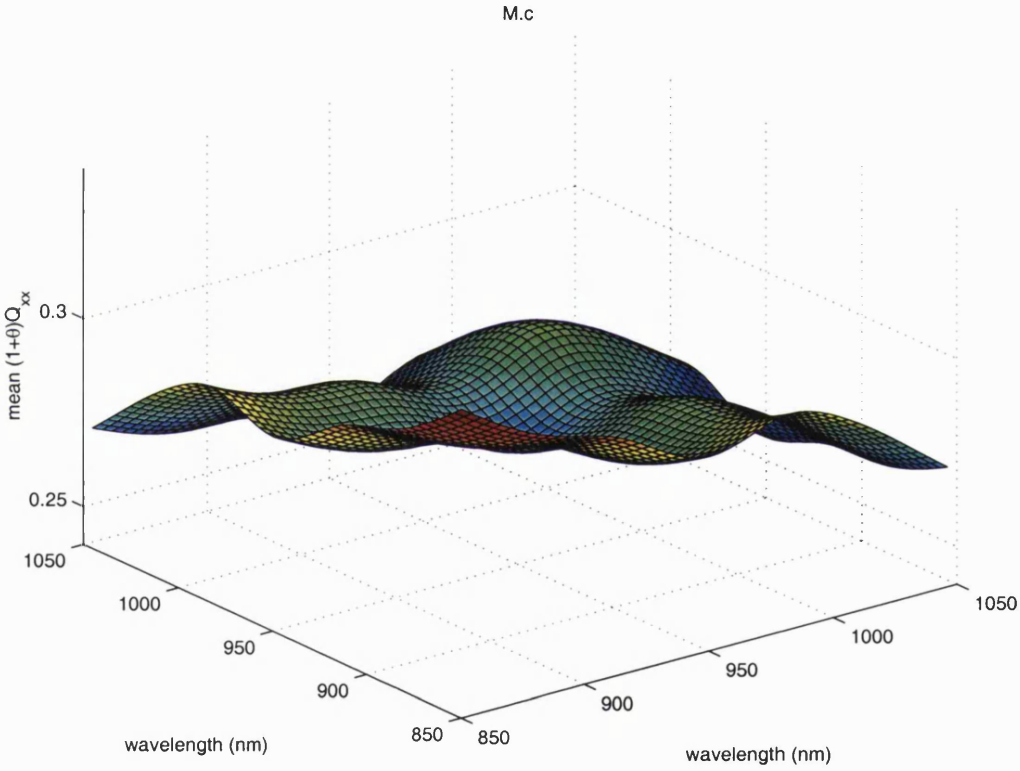


Figure E.5: $X_t{}^t X_t$ for the 2nd derivative spectra

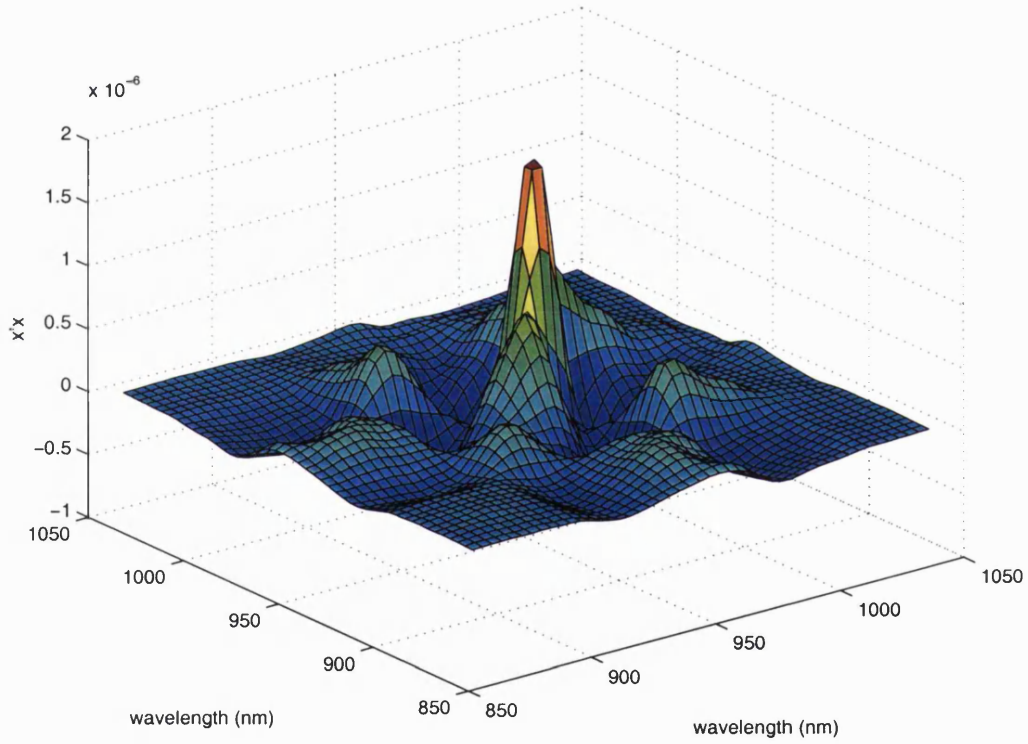


Figure E.6: MCMC estimate of the mean marginal $(1+\theta)Q_{xx}$ for M.d

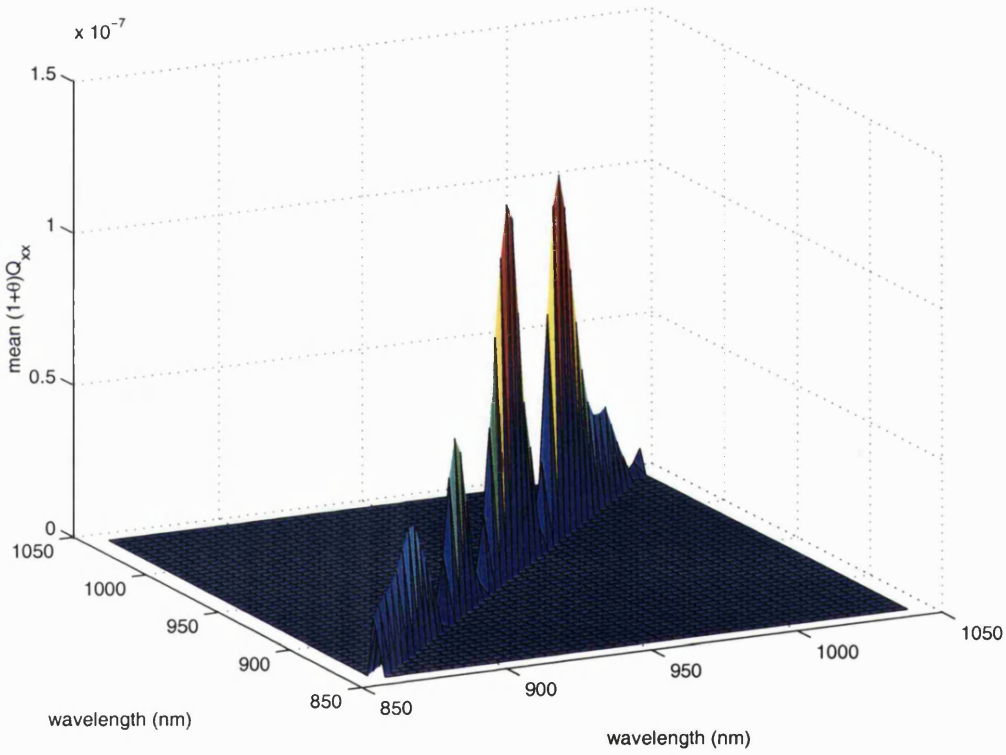


Figure E.7: MCMC estimate of the mean marginal $(1+\theta)Q_{xx}$ for M.e

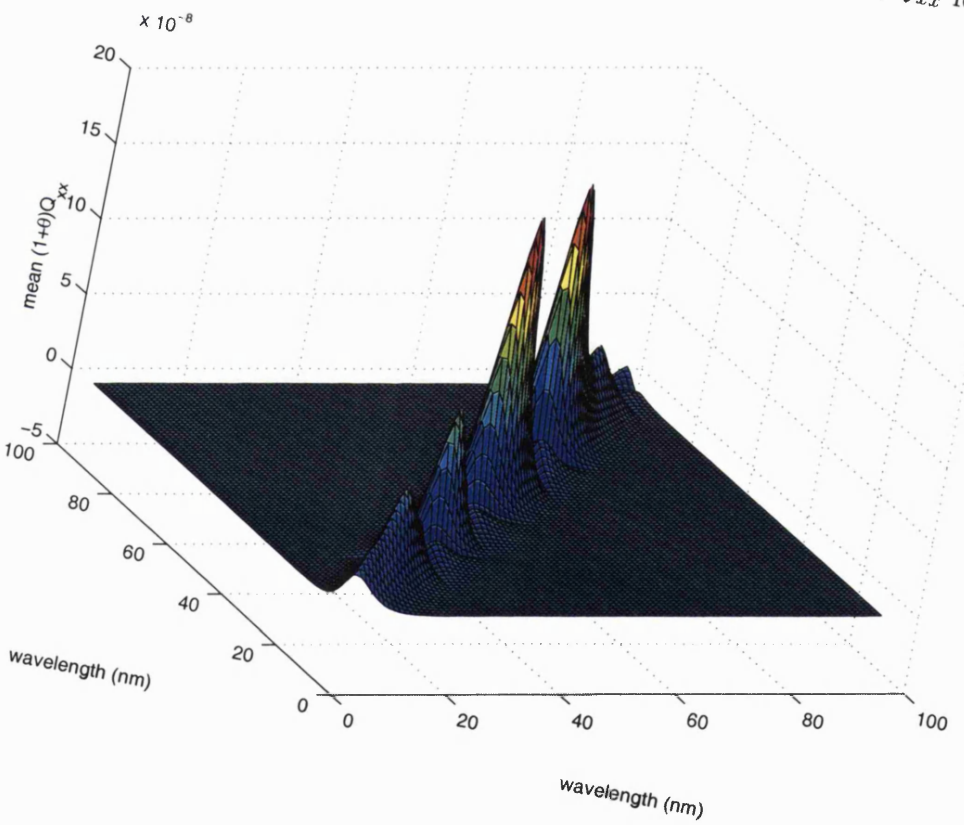


Figure E.9: ${X_t}^t Y_t$ of the 2nd derivative spectra and MCMC estimates of the mean marginal $(1+\theta)Q_{x\eta}$ of M.d, and M.e

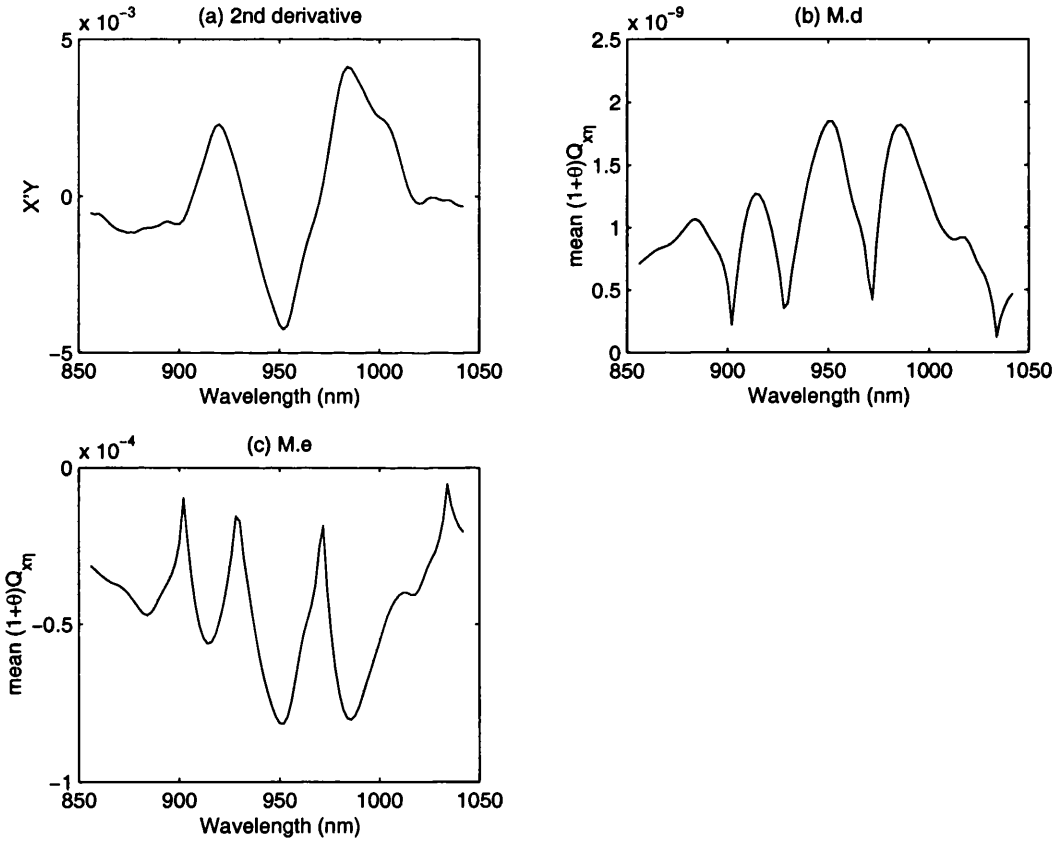


Figure E.9: ${X_t}^t Y_t$ of the 2nd derivative spectra and MCMC estimates of the mean marginal $(1+\theta)Q_{x\eta}$ of M.d, and M.e