Copyright © 2006 by Joseph Edward Lucas
All rights reserved

SPARSITY MODELING FOR HIGH DIMENSIONAL SYSTEMS: APPLICATIONS IN GENOMICS AND STRUCTURAL BIOLOGY

by

Joseph Edward Lucas

Institute of Statistics and Decision Sciences
Duke University

Date: ______________________________
Approved:

____
Dr. Mike West, Co-chair

____
Dr. Scott Schmidler, Co-chair

____
Dr. David Dunson

____
Dr. Terrance G. Oas

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2006

ABSTRACT

(Statistics)

SPARSITY MODELING FOR HIGH DIMENSIONAL SYSTEMS: APPLICATIONS IN GENOMICS AND STRUCTURAL BIOLOGY

by

Joseph Edward Lucas

Institute of Statistics and Decision Sciences
Duke University

Date: ______________________________
Approved:

____
Dr. Mike West, Co-chair

____
Dr. Scott Schmidler, Co-chair

____
Dr. David Dunson

____
Dr. Terrance G. Oas

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2006

Abstract

The availability of very high dimensional data has brought sparsity modeling to the forefront of statistical research in recent years. From complex physical models with hundreds of parameters to DNA microarrays which offer observations in tens to hundreds of thousands of dimensions, separating relevant and irrelevant parameters is becoming more and more important. This dissertation will focus on innovations in the area of variable and model selection as they pertain to these high dimensional systems.

Chapter 1 will discuss work from the literature on the areas of variable and model selection.

Chapter 2 will describe an innovation to hierarchical variable selection modeling that corrects errors that stem from assuming incorrectly that multiple thousands of observations are informing about the same distribution.

In Chapter 3, we introduce a novel technique for applying variable selection priors to induce sparsity in variance modeling.

One of the weaknesses of DNA microarrays is their sensitivity to the conditions under which they were prepared. Chapter 4 describes a technique for correcting the systematic bias that is introduced by these extreme sensitivities.

Chapters 5 and 6 are both case studies. They focus on implementing the techniques described in chapters 2-4 in real world situations in order to ferret out pathway signatures and to apply those to clinical situations.

Chapter 7 will introduce a new technique for sampling from a point mass mixture prior when calculation of the conditional probability is impossible.

In Chapter 8, we apply this technique to a challenging problem in structural

biology.

For Chapter 9, we switch gears somewhat and apply some of the techniques of decision theory the protein folding problem introduced in chapter 8. We are able to use the results of our model fitting to inform future decisions for studying polypeptide helicity.

Finally, we close, in Chapter 10, with some areas for future work that have opened up as a result of studying these variable selection techniques.

Acknowledgements

My heartfelt thanks go to my two advisors, Mike West and Scott Schmidler, for the immense amounts of time they have spent trying to turn me into a statistician. Their joy for the work they are doing is infectious, and has inspired my own.

Thanks goes to David Dunson for conversations that spark many ideas for statistical innovation.

Without collaboration, advancement of statistics becomes directionless and uninteresting. I owe much to the careful and inspired laboratory work of Andrea Bild, Jen-Tsan Chi, and Terry Oas.

I want to thank Carlos Carvalho, Eric Vance, Jarad Niemi, and all of my fellow students at ISDS. There have been a lot of fun chats about statistics and many other topics. It is a pleasure to come into an office where one respects ones co-workers and looks forward to their company.

Without the support of my wife, Miha, finishing this dissertation would have been entirely impossible. Through all of the writing, editing and revising, she has encouraged and helped me in many ways.

Though they don't know it, my children also deserve thanks, as it is with their interests in mind that I make most of my choices.

Contents

List of Figures

Chapter 1

Bayesian Variable Selection

The problem of model selection is pervasive in modern statistics. Determining which of the (potentially many) predictors are relevant to a response variable of interest is a surprisingly difficult prospect because of the exponential number of potential predictor interactions. The volume of literature on model selection is vast, but focuses mostly on the issue as it pertains to generalized linear models. An excellent survey is available in Clyde and George (2004).

In this situation we have a variable Y that we want to explain in terms of a subset of measured variables $\{X_i\}_{i=1}^{k}$. Let $\varphi = 1, \cdots, 2^k$ index the subsets of our predictor variables. Then we want to choose from among the 2^k models of the form

$$Y = X_\varphi \beta_\varphi + \epsilon.$$

With Markov Chain Monte Carlo and the ability to do fast calculation, statisticians are now able to attack this problem directly with hierarchical models. One puts prior distributions on the different models $\pi(M_\varphi)$ and priors on the parameters $\beta_\varphi | M_\varphi$ of each model. The most studied of the hierarchical models is of the form

$$\begin{aligned}\beta_\varphi|\varphi,\Sigma &\sim N(0,\Sigma),\\ \varphi|p &\sim Bin(r_\varphi, k-r_\varphi),\end{aligned}$$

where r_φ is the number of predictors included in model φ and p is the probability of including any particular variable in the model. When Σ is diagonal, this formulation is equivalent to using a mixture of a normal distribution and a point mass at zero for each of the k β's as follows:

$$\beta_i \sim (1-p)\delta_0(\beta_i) + pN(\beta_i|0,\phi)$$

The term "Spike and Slab" has been used to describe this model for obvious reasons. Its popularity is due, in part, to its tractability to calculation. In the linear model situation outlined above this prior is conditionally conjugate, which allows for complete posterior inference via Gibbs sampling in MCMC.

1.1 Some History and Related Ideas

There is a large body of work built up around a different approach to the problem of model selection. This involves minimizing a penalized sum of squared errors function

$$\frac{SSE_\varphi}{\hat{\sigma}^2} + \lambda(k)\cdot r_\varphi$$

where $\hat{\sigma}^2$ is an estimate of variance under the "true" model, $\lambda(k)$ is a function of the size of the largest model under consideration, and r_φ is the size of model φ. Techniques for choosing penalty functions are myriad. The Akaike Information

Criterion (AIC) seeks to minimize the expected Kullback-Leibler distance (Akaike, 1973). The Bayesian Information Criterion (BIC), introduced in Schwarz (1978), chooses a penalty that is asymptotically equivalent to model selection based on Bayes' factors (Kass and Raftery, 1995). The advantage of penalized least squares models is that they are "automatic" in the sense that they allow one to avoid having to specify a potentially large collection of hyper-parameters. More recently, C_{MML} and C_{CML}, introduced in George and Foster (2000), use Empirical Bayes to automatically define hyper-parameters in the context of hierarchical models. They are able to show the equivalence of this technique to a penalty function that varies with the structure and amount of data available. A concise survey of the history of these models and others can be found in George (2000).

These techniques are prevalent in applications due to their ease of use, though there has been some statistical work in producing new penalty functions from the perspective of Empirical Bayes. Software tools for model selection based on AIC and BIC are readily available and there are many users who are comfortable with them. However George and Foster (2000) show that, given a particular (fixed) penalty function, there is a choice of (constant) Σ and p so that the ordering of the posterior probabilities of the models, $L(M_\varphi)$, is exactly the same for the hierarchical model selection and the approach that seeks to minimize a penalized least squares (George and Foster, 2000). Because it offers significant increases in flexibility over older approaches (we need not leave p and Σ constant), the main body of work in variable selection since the introduction of variable selection priors in Box and Meyer (1986); Mitchell and Beauchamp (1988) has focused on hierarchical modeling.

There are other approaches to variable selection that are closely related to or derivatives of the hierarchical model described briefly above. One such, introduced by George and McCulloch (1993), uses the following prior for the coefficients of regressors:

$$\beta_i \sim (1-p)N(\beta_i|0,\rho) + pN(\beta_i|0,\sigma\rho), \tag{1.1}$$

$$p \sim (1-r)\delta_0(p) + r\delta_1(p). \tag{1.2}$$

The differences here are the replacement of the point mass part of the mixture with a spike normal distribution and the replacement of a continuous distribution for p with point masses at zero and 1. We are estimating β_i with a mixture of a very low variance (ρ) normal distribution and a higher variance, $\sigma\rho$ with $\sigma >> 1$, normal distribution. In order to avoid problems of identifiability and to aid in interpretability, we are forced to fix ρ at a value that would keep β_i small enough to be estimable by zero and to keep σ large so that if β_i is drawn from $N(0,\sigma\rho)$ then it is unlikely to be estimable by zero. The strongest reason to use this model over the variable selection model defined above is when the measurement error in the data is low enough, or the number of observations is high enough to identify non-zero coefficients that would, nevertheless, be accurately approximated by zero (George and McCulloch, 1997).

The LASSO method for variable selection, introduced by Tibshirani (1996), uses a standard linear model setup with double exponential priors on the coefficients rather than normal priors. This is equivalent to a normal prior distribution with a certain gamma prior for the variance (Andrews and Mallows, 1974). While it is evident that this does not lead to posterior mass at zero for any of the coefficients, it can lead to maximum a posteriori (MAP) estimates of zero for many of them. It is equivalent to (and indeed it was introduced as) a classical least squares regression model in which one constrains the sum of the absolute values of the coefficients, $\sum |\beta_i| \leq k$. The Lasso produces a single "best" answer for any given value of k, and so any ability to rank the models, or use model averaging to improve predictive accuracy is lost. For the same reason, there is no information about posterior density or uncertainty. However, Tibshirani has discovered an algorithm (Efron *et al.*, 2004) which allows the computation of the LASSO solution to a linear regression problem for all values of the constraint, k, and which has the same computational complexity as basic least squares regression. Thus, for variable selection problems, LASSO offers the possibility of solving problems that would be computationally infeasible for MCMC algorithms such as those required by hierarchical models with mixture priors.

1.2 Specification of Prior Distributions

The "Spike and Slab" models allow full posterior inference via MCMC algorithms. We may compute posterior probability of inclusion as a variable selection criterion, we automatically generate full posterior distributions for all model parameters, and we have all the information required to use model averaging for the purposes of prediction. In this sense, Bayesian hierarchical models provide a complete solution to the variable selection problem in generalized linear models. However, this does not mean that there is no more work to be done.

In this case, the selection of informative prior distributions and hyper-parameters is a challenge due to the potentially large number of parameters to choose. Additionally, the consequences of choosing specific priors on posterior bias have not yet been fully elucidated. Indeed, the effect of a particular prior structure is different for problems with high dimension and low observation number data versus low dimensional problems with many potential predictors.

1.2.1 Variable Selection with g-Priors

There are two different prior specifications to be made in the "Spike and Slab" model. The first is the prior probability of variable inclusion, p, and the second is the prior variance structure, Σ. Often the prior variance is expressed in terms of Zellner's g-priors as follows:

$$\beta_i \sim (1-p)\delta_0(\beta_i) + pN(\beta_i|0, g\sigma^2(X'X)^{-1}).$$

This leaves the choice of only p and g. Some suggested choices are given in Kass and Wasserman (1995), Foster and George (1994), and Fernandez *et al.* (2001). This type of model suffers from two well known issues. First, if one endeavors to be as uninformative as possible by choosing g large, one biases the results toward the null model, $\beta_i = 0$ for all i (This is a well known problem and is called either Lindley's or Bartlett's paradox). Second, as the information supporting a specific model, M_φ, accumulates, the Bayes' factor for that model tends to a constant that is strictly less than 1 (Information Paradox).

There are two alternatives to the approach of assigning constant values to p and g. Empirical Bayes (EB) seeks to choose p and g to maximize the model marginal likelihood. This is worked out in George and Foster (2000) under the assumption of a single g for all models. One might follow the same procedure assuming a different g for each model. Liang *et al.* (2005) refer to these as Global and Local Empirical Bayes, respectively. These are conceptually very similar to the approach of fixing g and p, but now these parameters will depend on the data. It has also been shown that they are equivalent to least squares regression with a particular, data dependent penalty function.

The other alternative is hierarchical modeling. Liang *et al.* (2005) describe Hyper-g priors, which are priors with polynomial tails for the parameter g. Use of such a prior admits the calculation of a closed form for the posterior of g and is, therefore, fast and easy. Zellner and Siow (1980) propose Cauchy priors on the regression coefficients, arguably more difficult to calculate with, but offering the best consistency results. See Liang *et al.* (2005) for the details of the relevant theorem and its proof.

All of these approaches resolve the consistency issue and outperform models with constant g and p. Additionally, it is unclear whether EB or hierarchical modelling performs better as choice of hyper-parameters seems to make a difference, for example Liang *et al.* (2005) and Cui and George (2004). A standard complaint about Empirical Bayes is that the use of the data to "compute" the prior leads to underestimation of variation in the posteriors. It should be pointed out that g-priors in general are subject to this complaint because of the use of the data covariance matrix in the construction of the prior variance, Σ, though it has not been shown that they maximize marginal likelihood in the empirical Bayes sense.

1.2.2 Independent Priors

Approaches to independent prior specification for the variable selection model

$$\beta_i \quad \sim \quad (1-p)\delta_0(\beta_i) + pf(\beta_i|\Phi)$$

are numerous and, for the most part, lack a coherent focus. (Here, f is a known distribution with parameters Φ.) This should not be surprising as the problems to which they can be applied vary widely. In social data, one may have many potential predictors of some outcome variable. On the other hand, in the case of gene expression data, we may have just one experimental group, and we want to determine in which of multiple thousands of genes there is an effect. In the case of wavelet density estimation, we may have just one "observation" at each basis function, and we need to determine which wavelet coefficients can be best estimated by zero.

The simplest possible solution to the problem is to assign a fixed value to p and let f be a mean zero normal distribution (with or without a known variance). Unfortunately, this approach has problems with consistency (Bartlett's Paradox, Information Paradox) and suffers from slow convergence as data accumulates. Additionally, fixing p can lead to significant bias in the posterior (see chapter 2). Finally, the lack of a prescribed technique for assigning p leads to significant variability in the results from application to application. This kind of dependence on the prior is excellent cause for rejection of the results of any analysis in which it is used. There are generally two types of solutions to these problems. The first, Empirical Bayes, seeks to use the data to produce automatic estimates for the parameters of the model. The chief complaint regarding this technique is that, by using the data to construct the prior distributions, one underestimates the true variance in the posterior. The second, hierarchical modeling, seeks to add levels to the hierarchy in order to allow the choice of diffuse hyper-parameters which have minimal effect on the outcome. This approach, however, leaves open the question of what constitutes minimal effect. Additionally, one is still left with the question of how to choose hyper-parameters, even though the negative consequences of a poor choice of prior are lessened.

1.3 Empirical Bayes

Empirical Bayes is the approach to the problem that has seen the most statistical research. This is not surprising, as determining the maximum likelihood estimate of a particular hyper-parameter is a statistical problem. Additionally, EB provides a technique for selecting hyper-parameters automatically, a feature that is important to users of the procedures who do not have the knowledge to make informed choices of hyper-parameters, or do not understand the procedures. The use of hierarchical modeling is more often used as a tool for specific data analysis problems in fields where expert opinion on the prior distributions is present.

In a pair of papers Johnstone and Silverman (2004) and Johnstone and Silverman (2005) outline a variable selection method specifically for a thresholding problem and apply it to wavelet density estimation. Suppose that we have many observations $\{x_i\}_{i=1}^n$, each of which are either zero (with some noise) or drawn from some symmetric distribution (with some noise), and that we want to estimate the mean of each observation. We must either choose zero or some non-zero value for each mean. The best one can do in this situation is threshold the data at some value, t, and estimate $\mu_i = 0$ when $x_i < t$ and $\mu_i = x_i$ otherwise. For varying levels of sparsity, the optimal value of the threshold parameter (in terms of mean squared error of estimation), varies widely (small p leads to high thresholds). Johnstone and Silverman proceed by assuming the model

$$x_i \sim (1-p)\delta_0(x_i) + pf(x_i|\Phi)$$

where $f(x_i|\Phi)$ is known, and choosing p to maximize the data likelihood. This leads to an adaptive threshold that is close to optimal in a rigorous sense, described by Johnstone and Silverman (2004), across all levels of sparsity. However, assuming that the method can be generalized to multiple observations from each μ_i, this leaves open the question of choosing $f(x_i|\Phi)$ and provides no information on the uncertainty of each particular estimate of the mean. (Though the latter would be of dubious use in the case of just one observation from each distribution.)

Another example of the use of Empirical Bayes is given in Yuan and Lin (2005). They formulate the model as follows. First, let γ be binary vector indicating which variables are non-zero, and define $|\gamma|$ to be the number of non-zero variables included in the model described by γ. Then β_γ is a vector of coefficients, $|\gamma|$ of which are non-zero. This model is of the form

$$\begin{aligned} Y|\gamma, \beta_\gamma &\sim N(X_\gamma \beta_\gamma, \sigma^2 I_n), \\ \beta_i|\gamma_i &\sim (1-\gamma_i)\delta_0 + \gamma_i DE(0, \tau), \\ f(\gamma) &\propto q^{|\gamma|}(1-q)^{p-|\gamma|}\sqrt{det(X_\gamma' X_\gamma)}. \end{aligned}$$

There are two significant changes to the basic formulation (equation 1.2) described above. The first is the use of a double exponential (DE) rather than a normal distribution in the mixture prior for β_i. The double exponential distribution has heavier tails, and, therefore, should be able to accommodate large ranges in the posteriors for the set of β_i. The second innovation is the addition of $\sqrt{det(X_\gamma' X_\gamma)}$ in the prior for γ. This has the desirable effect of lowering the prior probability for models that include highly collinear predictors. This model formulation has three hyper-parameters, σ, τ, and q, which need to be set. Yuan et al. propose using Empirical Bayesian techniques to automatically select these values, though the possibility of adding levels to the hierarchical model is clear. Yuan et al. were able to show in simulation studies that this model outperforms the EB model introduced above by George and Foster (2000) in many situations, and is tractable to calculation even in cases of high dimensional data.

1.3.1 Hierarchical approaches

Approaches to hierarchical modeling are somewhat varied. West *et al.* (2001) and West (2003) use a beta prior for p with a conjugate inverse gamma prior for σ^2, and indeed this seems to have become the default choice in hierarchical variable selection models (Clyde and George, 2004). Wolfe *et al.* (2004) describes the use of an inverse gamma (IG) prior for σ^2 with a one parameter family of gamma priors for the shape parameter. Additionally, they use a prior that allows the modeling of dependence structure in the p_i's in cases where there is some idea of a metric on the β_i's. Geweke (1996) prefers a non-conjugate χ^2 prior for σ^2 for reasons of interpretability. Finally, there is some evidence (Yuan and Lin, 2005) that mixing a point mass with a reflected exponential, rather than a normal, distribution leads to more robust results. With the notable exception of Wolfe *et al.* (2004), little attention has been paid to prior distributions for p. Additionally, hierarchical modeling is something like the Wild West; because there is no consensus as to the best choice for any of the prior distributions or hyper-parameters, researchers must fend for themselves.

1.4 Data and Examples

The first of three Affymetrix data sets we use in the paper is from a Human Mammary Epithelial cell line (HMEC). The arrays were generated by upregulating specific known oncogenes (Bild *et al.*, 2006). The data set consists of 15 samples from a control group, and 7-10 samples from each of the 9 experimental groups. The 9 experimental groups correspond to the upregulation of the following genes: myc, src, b-cat, E2F3, H-Ras, Np63α, AKT1, E2F1, and Pik3CA. All of the samples were generated from the same cloned cell line, thereby avoiding significant gene expression changes due to normal biological variability. Thus, observed variation in gene expression should be due to oncogenic intervention exclusively. However, the samples from this experiment come in three groups which were collected months apart. The first consists of 10 of the controls along with the first 5 experimental groups, the second consists of the last 4 experimental groups, and the remaining 5 controls were collected last. Array data is notoriously difficult to collect in a standard way, and there are clear temporal effects observable in the data. Additionally, there is evidence in some genes of large changes in expression variance across the experimental groups.

The second of the Affymetrix data sets, also obtained from HMEC's, involved a study of the effects of hypoxia and lactic acidosis. Three samples were taken from the cell line in each of four conditions: control, in the presence of lactic acid, in the absence of oxygen, and with lactic acid in addition to hypoxia. These cell lines do not show the temporal variability that can be seen in the up-regulation experiment, but with only three observations in each group there is no way to observe any differences in the variance of expression levels across the experimental conditions.

The final Affymetrix data set was generated on RNA extracted from the breast tumors of hospital patients (Miller *et al.*, 2005). In addition to biological variability across patients, the tumors themselves are heterogeneous, so we expect to find distributions of expression levels in this case that look much more unimodal. We have survival data from these patients, and this will allow us the opportunity to find collections of genes from the previous data sets, and then test these collections for their ability to predict patient survival.

In general, RMA data is noisy, so we will drop from consideration any gene which remains low across all samples. Also, from prior experience with microarrays, we have some idea of likely levels of experimental variance of the observations. This will allow us to inform some of our prior distributions where appropriate.

For the second part of the text, we will focus on a different type of data set. We have collected from publications a set of polypeptides with experimentally measured alpha helicity. For each such polypeptide, we have information on the temperature and pH at which the measurement was taken. There are 360 distinct polypeptides with such measurements and there are included a number of temperature and pH curves, bringing the total number of observations to 1187. Of the 360, 142 are designed polypeptides and the remainder are naturally occurring. The designed sequences contain multiple instances of alanine and lysine. Because of this and other biases in peptide design, we can not assume that the polypeptides were sampled randomly from polypeptide space.

1.5 Summary

While there has been significant work on the variable selection problem as it pertains to linear models, there is a conspicuous absence of work on hierarchical models. In chapter 2, we will examine the choice of prior probability of inclusion of a model parameter, p, and offer an expanded model to correct for bias in the posterior probability of inclusion. In chapter 3, we will examine the potential application of variable selection priors in models of heterogeneous variance in expression. Chapters 4 and 5 are case studies in which we apply our multilevel hierarchical model to the gene expression data sets described above.

Another area of importance for variable selection is in non-linear models. In chapter 7 we introduce a particular graphical model and discuss the possibility of variable selection in this context. Chapters 8 and 9 are devoted to the application of this model to the prediction of helicity in short polypeptides. Thus we will retain our biomedical/genomics theme throughout the thesis, with our theoretical focus centering on model and variable selection.

Chapter 2

Shrinkage in “Large p Small n” Analysis

Since the introduction of variable selection priors, there has been little focus on specification of the prior probability that a variable is included in a model. This prior probability is often treated as a nuisance parameter. Empirical Bayes solutions set the parameter to a constant and hierarchical Bayes solutions make the obvious choice of giving it a weak beta prior.

It is almost always the case, however, that an experimenter is specifically interested in the posterior probability of inclusion as a study variable. In these cases, it is difficult to justify a constant prior even if it is derived from a maximum likelihood estimate. The loss of some of the posterior uncertainty can be shrugged off when the posterior of the given variable is of secondary interest, but it is not clear that this is the case if it is of primary interest.

The hierarchical Bayes solution of using a beta hyper-prior has a different problem. By definition, it assumes that all of the parameters under consideration are being drawn from the same distribution. That is, the variable inclusion/exclusion indicators are viewed as exchangeable. This is often a dubious assumption. Consider the oncogene experiment described in chapter 1. We have 10 observations from a control group, and 10 from and experimental group in which a particular gene has been artificially up-regulated in a human cell line. We may have results from 50,000 or more genes. If we make the prior assumption that the probability of seeing a shift in mean expression level follows a single beta distribution, then we are in effect saying that each of the 50,000 experiments gives us information about the same distribution. Now consider what happens when we eliminate 40,000 of the genes because the associated expression levels are all below the detection threshold of the experiment. If there were about 1000 genes that show a change in mean due to our experimental up-regulation, then the posterior probability of inclusion has changed from 1/50 to 1/10 (because a beta distribution updated with either 10,000 or 50,000 observations is essentially a constant). Eliminating the “noise” genes, which are providing no information about the variables of interest, has significantly changed our posterior.

The main problem here is the assumption that all of the 50,000 expression levels are providing information about the same distribution. In fact, some of the microarray probes have nothing at all to do with our artificially up-regulated gene. Indeed, some of them may have nothing to do with human mRNA at all!

What we truly want is a hyper-prior that implies that some of the expression levels are within the pathway of our up-regulated gene and some are not, but we do not know, a priorí, which are which. In this chapter we will outline new statistical thinking and resulting modelling methodology that aims to better differentiate between noise and useful information in such a situation. There is some similar work in Ishwaran and Rao (2005) for use with spike and slab type variable selection models.

2.1 Framework and Notation

There are three main data sets that we will be using throughout the first part of the text for illustrative purposes. These are Affymetrix DNA microarray data sets. For these, we will use the standard RMA estimates of the intensity of the flourescence at each point on the chip, (Irizarry *et al.*, 2003a,b).

We will write $y_{g,i}$ to refer to the expression level of gene g from sample i. We will reserve p for the number of genes and n for the number of samples. Let y_i denote the column vector of expression levels for sample i and $Y = [y_1, \cdots y_n]$ be the $p \times n$ dimensional matrix of expression levels.

We will use linear regression in dealing with these data sets:

$$y_i = \mu + Bx_i + \epsilon_i$$

where μ is a p dimensional vector of means (constant across samples), x_i is a d-dimensional design vector of known covariates, and B is a $p \times d$ dimensional vector of unknown coefficients. Our error in the model, ϵ, is a p-dimensional vector of zero mean normal random variables, $\epsilon \sim N(0, \psi)$, where ψ is a $p \times p$ diagonal covariance matrix. We write μ_g for the g^{th} element of μ and β_g' for the g^{th} row of B. If we let $j \in 1 \cdots d$, then we write $\beta_{g,j}$ as the element of B in the g^{th} row and j^{th} column. Our model above can then be written:

$$y_{g,i} = \mu_g + \sum_{j=1}^{d} \beta_{g,j} x_{j,i} + \epsilon_{g,i}$$

where the $\epsilon_{g,i}$ are conditionally independent $N(\epsilon_{g,i}|0, \psi_g)$.

2.2 Prior Assumptions about Probability of Inclusion

The default choice of distribution for prior probability of inclusion of a model parameter is a beta distribution. A simple form of the model, with a single regressor, x_i, on sample i is:

$$\begin{aligned} y_{g,i} &= \beta_{g,0} + \beta_{g,1} x_i + \epsilon_{g,i}, && (2.1) \\ \beta_{g,0} &\sim N(m, v), && (2.2) \\ \beta_{g,1} &\sim (1-\pi)\delta_0(\beta_{g,1}) + \pi N(\beta_{g,1}|0, \tau), && (2.3) \\ \pi &\sim Be(\alpha, \gamma), && (2.4) \\ \epsilon_{g,i} &\sim N(0, \psi_g). && (2.5) \end{aligned}$$

where $1 \leq i \leq n$ is an index on the observations and $1 \leq g \leq p$ where p is the (large) number of variables, (genes in a microarray experiment).

This is a shrinkage prior on the variable π which tends to shrink the posterior in surprising ways. Specifically, consider a simulated example containing 1,000 "genes", each with 10 control group and 10 experimental group observations. Suppose that all controls are drawn from a standard normal distribution. Additionally, suppose that in 700 of the trials the experimental group is drawn from a standard normal distribution while in the remaining 300 trials the result is from a $N(1.5, 1)$ distribution. Take a uniform Be(1,1) prior for π and $Ga(1, 100)$ priors for both $1/\tau$ and each of the $1/\psi_g$.

As one would expect (see Figures 2.1 and 2.2), the posterior means for $\beta_{1,i}$ and ψ_g fit what we know of the data. However, consider the posterior probability that $\beta_{g,1} \neq 0$ shown in Figure 2.3. We see that the bulk of the mean posterior

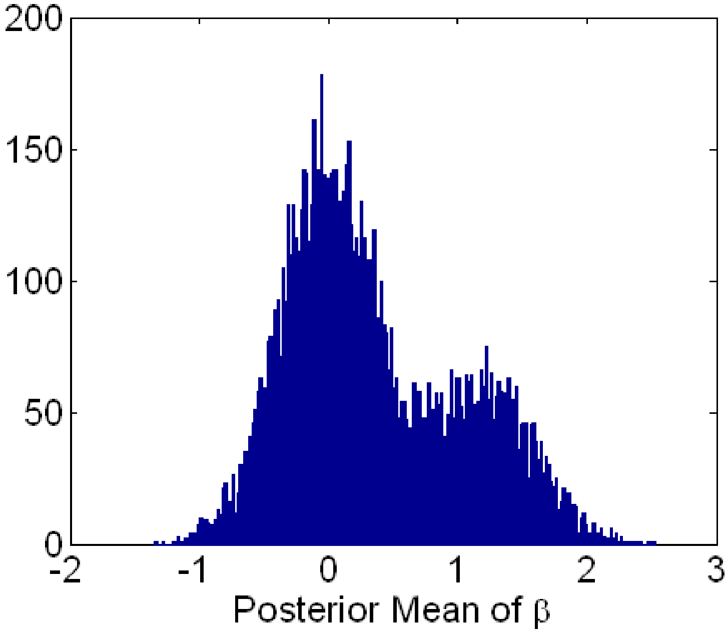


Figure 2.1: A histogram of the posterior means of the regression coefficients, $\beta_{g,1}$, from the model outlined in equations 2.1 through 2.5 on simulated data. The data was generated such that the "truth" is 0 for 700 of the trials and 1 for 300 of the trials. The standard Bayesian variable selection model correctly obtains a bimodal posterior distribution for the coefficient, which matches the simulated data set.

probability of $\beta_{1,i} \neq 0$ is above .4 and that all of the 1000 posteriors are larger than .2. This is surprising considering that there are 123 of the 1000 trials in which the absolute difference in mean from control to experimental is less than .01 and the minimum difference is 10^{-4}. The model in this situation predicts a 30% probability of an experimental effect when there is in fact no evidence whatsoever in the data.

In fact, it is possible to observe the same mean from a $N(0,1)$ process and a $N(1.5,1)$ process; the probability, given 10 observations from each, that the second process will produce a smaller mean than the first is only about .04%. Clearly, then, a posterior of 30% is undesirable. Let us consider what else is going

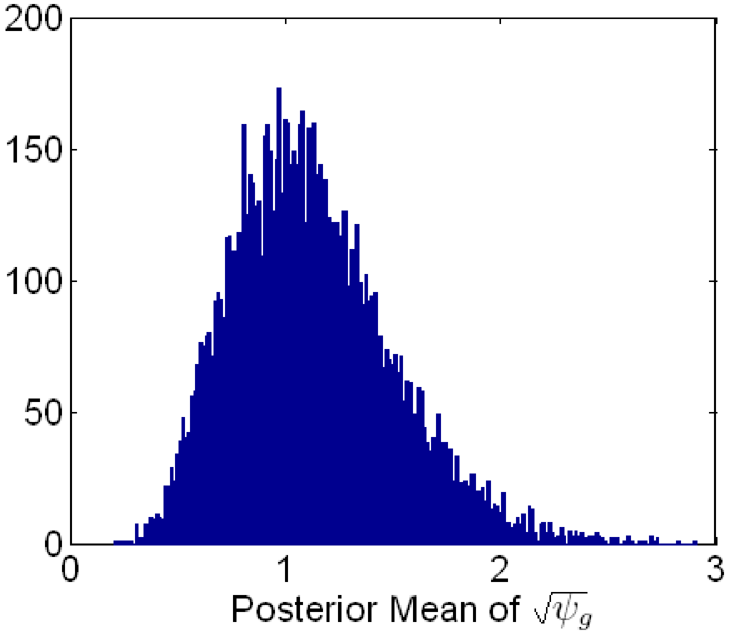


Figure 2.2: A histogram of the posterior means of the standard deviations, $\sqrt{\psi_g}$, from the model outlined in equations 2.1 through 2.5 on simulated data. The data was all generated from normal distributions with variance 1. The actual variance for the data was 1.

on.

Let $p(\beta_{g,1}|-)$ be the conditional density of $\beta_{g,1}$. The prior for $\beta_{g,1}$ is conditionally conjugate, so that

$$p(\beta_{g,1}|-) = (1 - \hat{\pi}_g)\delta_0(\beta_{g,1}) + \hat{\pi}_g N(\beta_{g,1}|m, v),$$

where the values of the parameters of this posterior are shown in equation 2.9 of section 2.3.1.

The strongest evidence for $\beta_{g,1} = 0$ occurs when $\sum_i y_{g,i} - \beta_{g,0}$ is minimal. In this case,

$$\hat{\pi} \;=\; \frac{\pi/(1-\pi)}{\sqrt{\tau/v} + \pi/(1-\pi)}.$$

We know from the construction of our data set that π is drawn from a beta distribution with mean approximately .3 after being updated with 1,000 observations. This implies a $>$ 99% probability that $\pi/(1-\pi)$ will be less than .46. Notice that v is the variance of the estimate of $\beta_{g,0}$ derived from 10 observations which is approximately $1/\sqrt{10}$. τ is the variance of the normal distribution that approximates the distribution of $\beta_{g,1}$, which in this case should be about .5 (the variance of a vector of 700 zeros and 300 values of 1.5). Under these assumptions, we find that $\hat{\pi} \approx .33$. Hence we expect to see a non-zero $\beta_{g,1}$ more than 33% of the time even when there is no evidence at all in the data. This phenomenon is what we see in Figure 2.3. (Note: we have used a $Be(.5, .5)$ prior for π and diffuse inverse gamma priors for τ and ψ.)

Consider the experiment we have just outlined. We have 1000 different trials, but implicit in this assumption is that there is something different across the trials, otherwise we would treat the data as coming from just one large experiment with 10000 control observations and 10000 experimental observations. On the other hand, the standard model assumes that the process by which an experimental effect occurs is the same across all of the trials. That is to say, from a modeling standpoint, our original experiment would have been no different if we had just repeated trial i 1000 times. Based on the standard Bayesian variable selection model, we are simultaneously assuming that the trials have some feature that differentiates them and that they are exchangeable.

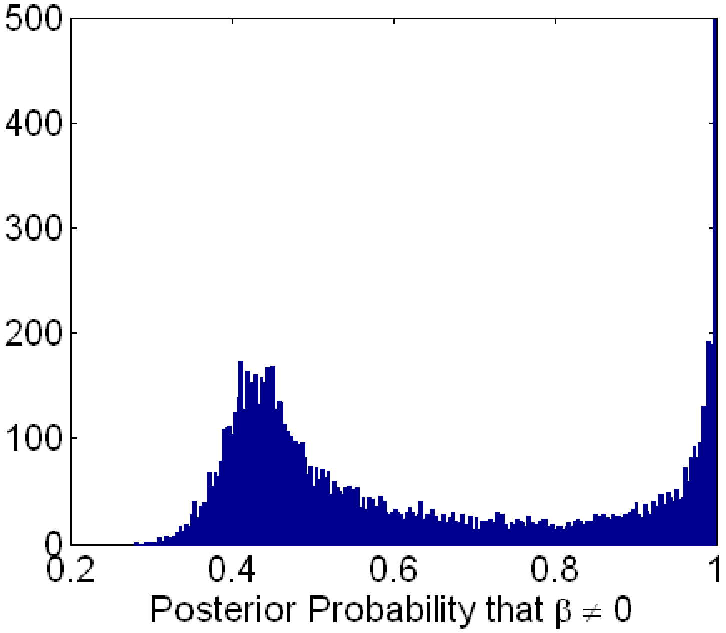


Figure 2.3: The posterior probabilities from the model outlined in equations 2.1 through 2.5 of $P(\beta_{g,1} \neq 0)$ for the 1,000 trials in the simulated data. The data contain only 300 trails in which $\beta \neq 0$. Shrinkage has caused $\beta_{1,i}$ to shift much higher than the data should suggest.

In truth, we believe that the results from some of the trials are best explained by assuming that $\beta_{g,1} = 0$, others are best explained by assuming that $\beta_{g,1} \neq 0$, and a priorí we do not know which are which. However, we do believe that the results from trial i should be consistent if we were to repeat the experiment. One model for $\beta_{g,1}$ that better reflects this situation is:

$$\begin{aligned} \beta_{g,1} &\sim (1-\pi_g)\delta_0(\beta_{g,1}) + \pi_g N(\beta_{g,1}|0,\tau), & (2.6) \\ \pi_g &\sim (1-r)\delta_0(\pi_g) + rBe(\pi_g|\alpha,\gamma), & (2.7) \\ r &\sim Be(\rho,\varphi). & (2.8) \end{aligned}$$

We now have a notion of π_g as a variable-specific (or in the case of microarrays, gene specific) prior probability that $\beta_{g,1} \neq 0$. By assuming a mixture of a point mass at zero and a beta distribution, we assume that some of the variables are best explained by assuming $\beta_{g,1} = 0$ with probability 1. When we repeat our experiment with the same simulated data, we obtain a much clearer separation of signal from noise, and lower our false discovery rate significantly (see figure 2.4). Full details of this prior and the resulting Bayesian analysis are described in section 2.3.2.

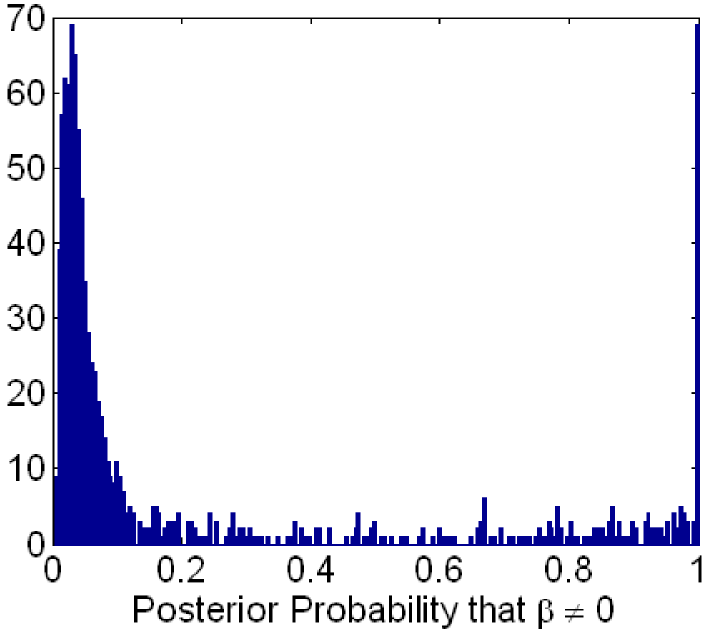


Figure 2.4: A histogram of the posterior probability of $\beta_{g,1} \neq 0$ for the model outlined in equations 2.6 through 2.8 on the simulated data set. The new model for π_g does a much better job of separating the noise from the clear signal. As before, 300 of 1000 experimental group results are drawn from N(1.5,1) while the rest, along with the controls, are drawn from N(0,1).

Choice of Hyper-Parameters

First consider the mixture prior for π_g. Notice that a low mean beta distribution can be almost indistinguishable from a point mass at zero. In what follows, we are considering the conditional densities of $\beta_{g,1}$ and π_g. With this in mind, we drop $\beta_{g,0}$ from consideration, as it can be incorporated into the data as a residual. Additionally, because $\beta_{g,1}$ and $\beta_{g',1}$ are conditionally independent, we will drop the subscript notation and simply write $\beta = \beta_{g,1}$ and $\pi = \pi_g$ here. Consider the density for π conditional on $\beta = 0$:

$$\begin{aligned}
p(\pi|\beta = 0, r) &\propto (1-\pi)\delta_0(\beta)[(1-r)\delta_0(\pi) + r \cdot Be(\pi|\alpha,\gamma)] \\
&\propto (1-r)(1-\pi)\delta_0(\pi) + r(1-\pi)\frac{\Gamma(\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\gamma)}\pi^{\alpha-1}(1-\pi)^{\gamma-1} \\
&\propto (1-r)\delta_0(\pi) + r\frac{\gamma}{\alpha+\gamma}Be(\pi|\alpha,\gamma+1) \\
&\propto \frac{(1-r)(\alpha+\gamma)}{\alpha(1-r)+\gamma}\delta_0(\pi) + \frac{\gamma_\pi r}{\alpha(1-r)+\gamma}Be(\pi|\alpha,\gamma+1) \\
&= (1-\hat{r})\delta_0(\pi) + \hat{r}Be(\pi|\alpha,\gamma+1)
\end{aligned}$$

with $\hat{r}$ implicitly defined.

Thus, when $\beta = 0$ an MCMC analysis will sample π from $Be(\alpha, \gamma + 1)$ with probability $\hat{r}$. The ideal choice of α and γ will depend on the data. If $\hat{r}$ is near zero, then π (and therefore β) will stick to zero leading to convergence issues in the MCMC chain. On the other hand, if $\hat{r}/r$ is near 1, then there is clearly an identifiability issue, as $\beta = 0$ is the strongest evidence possible for $\pi = 0$. $\hat{r}/r = \gamma/(\alpha + \gamma - \alpha r)$, so the ratio approaches 1 when γ is large relative to α. This is not surprising since a $Be(1, 100)$ distribution (for example) is difficult to visually distinguish from a point mass at zero.

The prior values for ρ and φ will determine the mean and variance of the prior for r. For a very low variance, r will be essentially constant, and our model treats each variable independently. That is to say, the status of $\beta_{g,1}$ provides no information on $\beta_{g',1}$ if $g \neq g'$. If this is the situation we want to achieve, we may need to choose a prior for r that is much more concentrated.

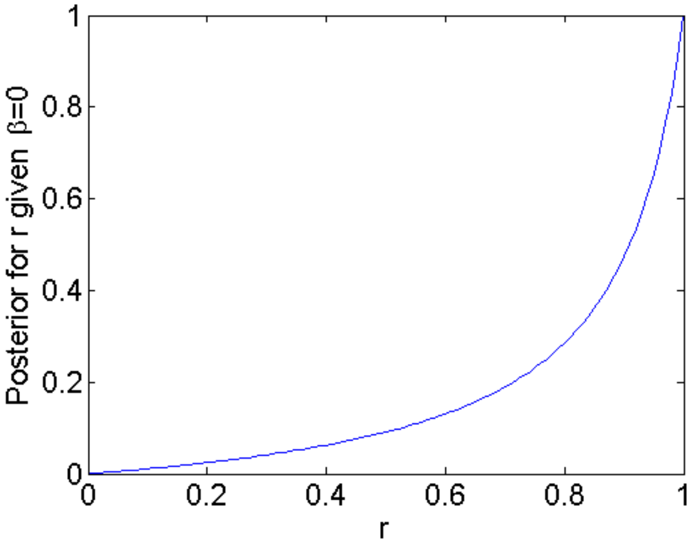


Figure 2.5: When we set $\alpha = .9$ and $\gamma = .1$ we obtain this relationship between $\hat{r}$ and r. This shows the relationship between the prior value for r and its posterior after we update when β is observed to be zero (the strongest evidence that $r = 0$).

Experimentally, we have determined that a choice of $\alpha = .9$ and $\gamma = .1$ work well for the point mass mixture prior for π (see Figure 2.5) in a number of analyses of simulated and real data sets.

As in many complex hierarchical models, there is some lack of clarity on the criteria for choosing hyperparameters. There is the possibility of making such decisions based on maximizing predictive accuracy, or by minimizing misclassification rates. This is an area for future work.

2.3 Markov Chain Monte Carlo

Thus far, we have focused our attention on a very simple model for illustrative purposes. In more practical situations, we must expect to have multiple covariates. In order to make the following sections more directly applicable, we will expand our model as

$$Y \sim N(\beta H', \psi I_p)$$

where we have a $n \times d$ design matrix, H, with columns corresponding to the d design vectors, a $p \times d$ matrix of regression coefficients, and a p-dimensional vector of variances, $\psi = (\psi_g)$. We will index the elements of β as $\beta_{g,j}$.

We have developed a relatively complex model which includes some point mass mixtures. Without careful design of a sampling algorithm, we will have difficulty with convergence. However, because of the structure we have chosen, we are able to use Gibbs sampling for all of the variables. Techniques for updating regression coefficients that have normal priors (such as the mean expression parameter, $\beta_{g,0}$) are standard, and will not be covered here. Instead we will focus on the updating schemes for the coefficients with variable selection priors.

2.3.1 Updating β

Note that, given all of the other variables, $\beta_{g,j}$ and $\beta_{g',j}$ are independent. Thus, in the interest of clarity, we will leave g out of the notation in this section. Thus we will write $y_{g,i} = y_i$, $\beta_{g,j} = \beta_j$, etc. We update each of the β_j's ($j = \{1, 2, \cdots, d\}$) sequentially and independently of the others. If H_j is the design matrix with the jth column set to zero, then let $t_i = y_i - \beta H_j'$. Our model for t_i is then $t_i \sim N(\beta_j x_i, \psi_i)$ where x_i is the ith element of the zeroed out column from the design matrix. Notice that our design matrix may contain some cases where x_i is neither 1 nor 0. For the purposes of updating β_j, we throw out observations corresponding $x_i = 0$. Define $\phi_i = \psi_i / x_i^2$ and $z_i = t_i / x_i$. Then, we can then write $z_i \sim N(\beta_j, \phi_i)$. Notice that we have allowed the variance parameter, ψ_i to be different across the observations. Our reasons for this will follow in Chapter 3. Now our complete conditional posterior density for β_j is

$$\begin{aligned}
P(\beta_j|-) &\propto \left(\prod_i N(z_i|\beta_j, \phi_i)\right)[(1-\pi_j)\delta_0(\beta_j) + \pi_j N(\beta_j|0, \tau)], \\
&\propto (1-\pi)\delta_0(\beta_j) e^{-\frac{1}{2}\sum_i z_i^2/\phi_i} + \pi e^{-\frac{1}{2}\sum_i (z_i - \beta_j)^2/\phi_i} \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{1}{2\tau}\beta_j^2}, \\
&\propto (1-\pi)\delta_0(\beta_j) + \pi e^{-\frac{1}{2}\left[\left(\frac{1}{\tau}+\sum_i \frac{1}{\phi_i}\right)\beta_j^2 - \left(2\sum_i \frac{z_i}{\phi_i}\right)\beta_j\right]} \frac{1}{\sqrt{2\pi\tau}}.
\end{aligned}$$

If we let $1/v = 1/\tau + \sum_i 1/\phi_i$ and $m = v \cdot \sum_i z_i/\phi_i$, then

$$\begin{aligned}
P(\beta_j|-) &\propto (1-\pi)\delta_0(\beta_j) + \pi N(\beta|m, v)\frac{\sqrt{v}}{\sqrt{\tau}} e^{m^2/(2v)}, \\
&= (1-\hat{\pi})\delta_0(\beta_j) + \hat{\pi} N(\beta_j|m, v)
\end{aligned}$$

where $\hat{\pi} = c/(1+c)$ and

$$c = \frac{\pi}{1-\pi}\frac{\sqrt{v}}{\sqrt{\tau}} e^{m^2/(2v)}.$$

Thus, to update β_j we let $\beta_j = 0$ with probability $1 - \hat{\pi}$ and otherwise sample $\beta_j \sim N(m, v)$.

2.3.2 Updating π

The other step in the MCMC of note is the updating of π. The conditional posterior is

$$\begin{aligned}
p(\pi|\beta=0,r) &\propto (1-\pi)\delta_0(\beta)[(1-r)\delta_0(\pi)+r\cdot Be(\pi|\alpha_\pi,\gamma_\pi)] \\
&\propto (1-r)(1-\pi)\delta_0(\pi)+r(1-\pi)\frac{\Gamma(\alpha_\pi+\gamma_\pi)}{\Gamma(\alpha_\pi)\Gamma(\gamma_\pi)}\pi^{\alpha_\pi-1}(1-\pi)^{\gamma_\pi-1} \\
&\propto (1-r)\delta_0(\pi)+r\frac{\gamma_\pi}{\alpha_\pi+\gamma_\pi}Be(\pi|\alpha_\pi,\gamma_\pi+1) \\
&\propto \frac{(1-r)(\alpha_\pi+\gamma_\pi)}{\alpha_\pi(1-r)+\gamma_\pi}\delta_0(\pi)+\frac{\gamma_\pi r}{\alpha_\pi(1-r)+\gamma_\pi}Be(\pi|\alpha_\pi,\gamma_\pi+1) \\
&= (1-\hat{r})\delta_0(\pi)+\hat{r}Be(\pi|\alpha_\pi,\gamma_\pi+1).
\end{aligned}$$

Thus, when $\beta=0$ we sample π from $Be(\alpha_\pi,\gamma_\pi+1)$ with probability $\hat{r}$ and otherwise we sample $\pi=0$.

Similarly, if we suppose that $\beta\neq 0$, then

$$P(\pi|\beta\neq 0,r) \propto \pi N(\beta|0,\tau)[(1-r)\delta_0(\pi)+r\cdot Be(\pi|\alpha_\pi,\gamma_\pi)].$$

In this case, however, we know that $\pi\neq 0$ since this would force $\beta=0$, thus

$$\begin{aligned}
P(\pi|\beta\neq 0,r) &\propto \pi N(\beta|0,\tau)r\cdot Be(\pi|\alpha_\pi,\gamma_\pi) \\
&\propto \pi\cdot Be(\pi|\alpha_\pi,\gamma_\pi) \\
&\propto Be(\pi|\alpha_\pi+1,\gamma_\pi).
\end{aligned}$$

So when $\beta\neq 0$, we draw π from a $Be(\pi|\alpha_\pi+1,\gamma_\pi)$ distribution.

2.3.3 Updating r_j and s_j

The parameter r_j is a prior on $I_{[\pi_{g,j}=0]}$. As a beta prior on a binary variable, it is conjugate. If we observe $\pi_{g,j} = 0$ for f of the genes, then we update r_g from the distribution:

$$r_j \sim Be(\alpha_{r_j} + p - f, \gamma_{r_j} + f).$$

Because of the conjugacy of these parameters, we are able to update and converge to our posterior distribution quickly.

2.4 Results

2.4.1 Gene Up-regulation Experiment

The modification of our hierarchical priors has a profound effect on the posterior results in our gene expression experiment. Figures 2.6 and 2.7 show the analysis of the MYC experimental group using the old beta prior and the new multi-level prior respectively.

As can be seen in figures 2.8 and 2.9, this trend is seen in all of the different experimental groups. The much lower discovery rate is indicative of a lower false discovery rate, and is more consistent with the expectation of the size of these gene pathways.

2.4.2 Lactic Acidosis Signature

We may apply the same analysis to another data set. In this experiment, we have a single cell line used for 12 gene microarray assays. There are three observations in each of four groups: control, lactic acid, hypoxia, and both. We will focus on the lactic acidosis vs control group for this section.

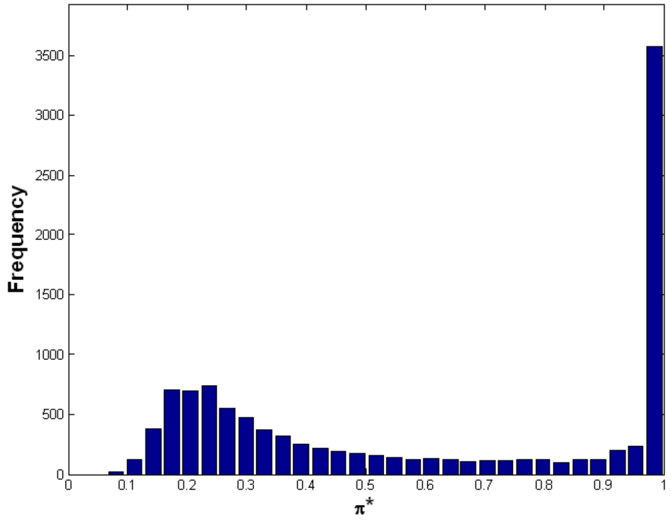


Figure 2.6: A histogram of the posterior probabilities of change in mean between control and the MYC experimental group as computed from the standard hierarchical Bayesian variable selection priors. Notice that the model finds that > 3500 of the 12,000 genes under study are significantly affected by the up-regulation of MYC. This is difficult to believe, as we are studying the pathway of just one gene.

With the standard hierarchical model, we find 3,350 genes that are relevant (posterior probability that $\beta_1 \neq 0 > .95$) in the lactic acidosis signature. Figure 2.10 shows the expression levels of all 12 of the experimental cells on the first two principal components of the expression levels of the 3,350 genes found to be relevant to the lactic acidosis intervention. We call these principal components the lactic acidosis metagene or the lactic acidosis signature. Notice that the cells grown in the presence of lactic acid are relatively poorly separated from the others.

Using the new hierarchical structure, we find that there are only 808 relevant genes in the lactic acidosis signature. Without the irrelevant genes that are “discovered” by the older hierarchical structure, we find a much cleaner separation of the lactic acidosis cells from the others in the experiment (see Figure 2.11).

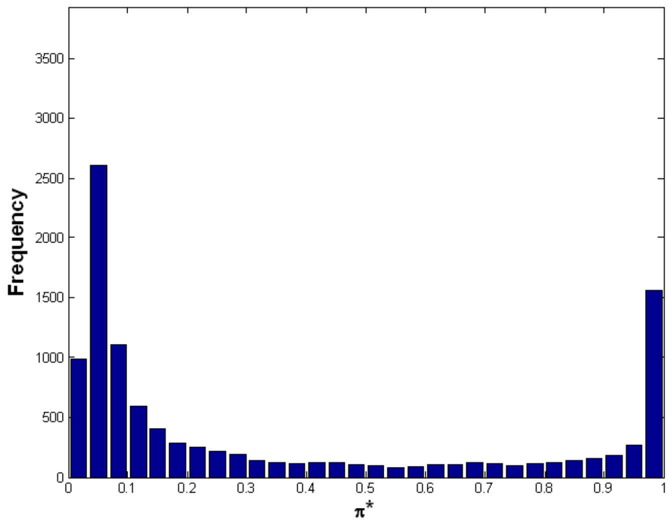


Figure 2.7: When we use the new, multi-level hierarchical prior, there is a large drop, from 3500 to around 1500, in the number of genes that are determined to be significantly affected by the up-regulation of MYC. Also notice that, compared to the standard hierarchical prior, there are many genes that are determined to have a very low probability of being in the MYC pathway. The genes with the least such probability using the standard model are still assigned a probability near 10%.

2.5 Summary

We have presented a new sparsity prior for multiple regression and anova,

$$\begin{aligned} y_{g,i} &\sim N(x_i'\beta_g, \psi_g), \\ \beta_{g,j} &\sim (1-\pi_{g,j})\delta_0 + \pi_{g,j}N(0,\tau_j), \end{aligned}$$

which assumes a prior probability of inclusion that is different for every gene and every group. These prior probabilities are independent, conditional on knowledge of a group specific overall inclusion probability, r_j:

$$\begin{aligned} \pi_{g,j} &\sim (1-r_j)\delta_0 + r_j Be(\alpha_\pi, \gamma_\pi), \\ r_j &\sim Be(\alpha_r, \gamma_r). \end{aligned}$$

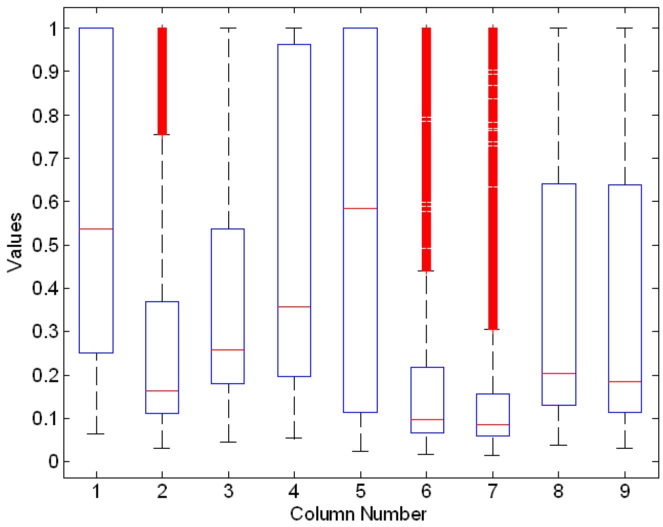


Figure 2.8: Boxplots of the posterior probabilities of a change in mean for each of the nine experimental groups across the 12,000 genes under study. These plots are of posteriors generated from the standard beta prior probability of a difference in mean. Notice that the median level of posterior probability of a difference in mean is much higher than that obtained using the new multi-level hierarchical prior.

This hierarchical structure leads to a conditional probability of a change in mean

$$\hat{\pi}_{g,j} = \frac{c_{g,j}}{1 + c_{g,j}}$$

$$c_{g,j} = \frac{\pi_{g,j}}{1 - \pi_{g,j}} \frac{\sqrt{v_{g,j}}}{\sqrt{\tau_j}} e^{m_{g,j}^2/(2v_{g,j})}$$

where $v_{g,j}$ and $m_{g,j}$ are defined as in section 2.3.1. This model leads to improved estimation of the probability of a true difference in mean between control and experimental groups that is both more true to our prior beliefs, and which produces posterior results that are nearer to expectation. Additionally, there are few additional complexities introduced in the Markov Chain Monte Carlo, with all steps updated by Gibbs sampling. While there is further work to be done in understanding the implications of choosing particular hyper-parameters, the benefits,

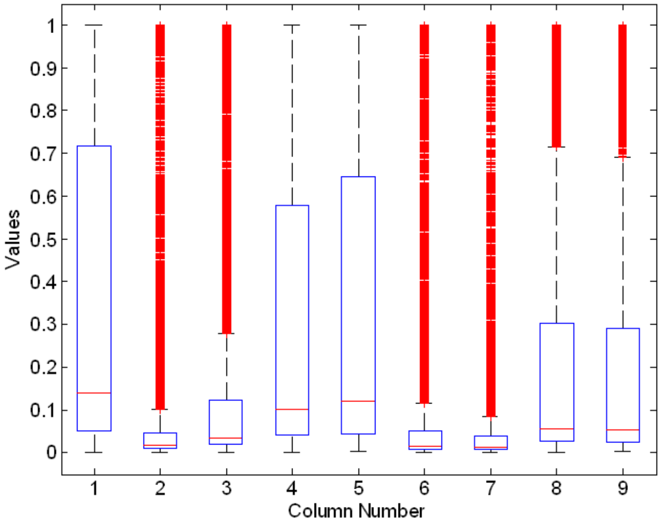


Figure 2.9: Boxplots of the posterior probabilities of a change in mean for each of the nine experimental groups across the 12,000 genes under study. These plots are of posteriors generated from the new multi-level hierarchical prior probability of a difference in mean. Notice that the median level of posterior probability of a difference in mean is much lower than that obtained using the standard beta hierarchical prior.

in terms of improved posterior inclusion probabilities, are significant.

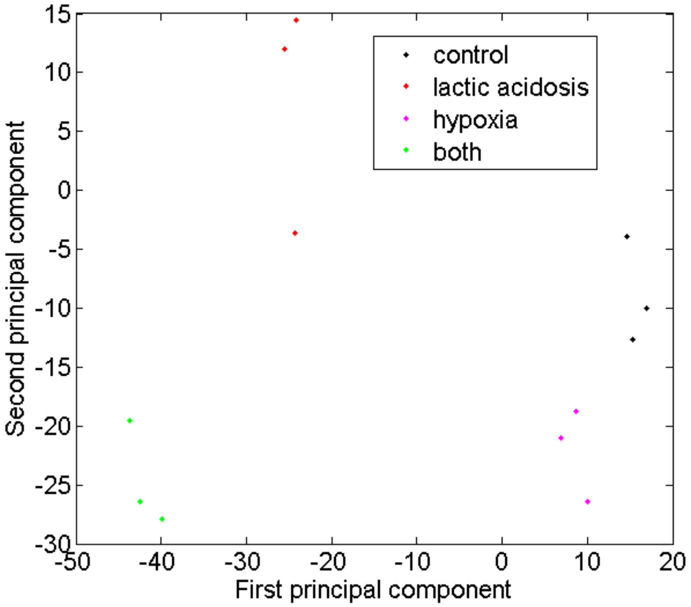


Figure 2.10: Expression levels for the first and second principal components of the lactic acidosis meta-gene. The meta-gene is derived with the standard model for hierarchical Bayesian variable selection, and encompasses 3,350 genes.

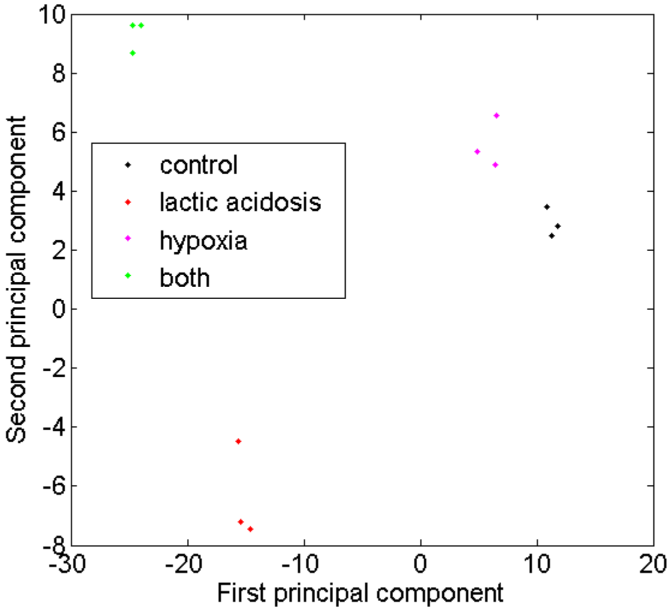


Figure 2.11: Expression levels for the first and second principal components of the lactic acidosis meta-gene. The meta-gene is derived with the use of the new hierarchical prior. This meta-gene contains only 808 genes and leads to a much cleaner separation of the cells grown in the presence of lactic acid from the rest of the cells in the experiment.

Chapter 3

Sparsity in Variance Modeling

Revolutionary advances in microbiology have led to the ability to assay mRNA expression levels for thousands of genes concurrently. This ability has led to an explosion of gene expression data in many different organisms and many different cell lines. These arrays have already been used in research institutions everywhere in the study of aortic aneurysms (Tung *et al.*, 2001), esophageal cancer (Hu *et al.*, 2001), breast cancer (West *et al.*, 2001), atherosclerosis (Karra *et al.*, 2005), and many other studies.

In many of these experiments, the researcher will obtain multiple observations from control and experimental groups. In this situation, the researcher may want to know which of the genes show a response to the treatment. As discussed in chapter 2, an excellent way to approach this is via the use of a variable selection prior. One of the weaknesses of this approach is the assumption of a constant variance across both control and treatment groups. This assumption causes loss of power, and may lead to both Type I and Type II error.

Implicit in the standard Bayesian selection model is the assumption that the experimental intervention will have no effect on the variance of the observations, ie.,

$$\begin{aligned} y_{g,i} &\sim N(\beta_{g,0} + x_i\beta_{g,1}, \psi_g), \\ 1/\psi_g &\sim Ga(a, b), \end{aligned}$$

implicitly assuming that ψ_g is constant as x_i varies. (Note: for simplicity, we have reverted back to a model that has just one variable with a selection prior.) While this can be a useful approximation, it is often inappropriate. Consider a typical dose response curve shown in Figure 3.1. At the upper and lower ends of the curve, the response will have much lower variance. With this in mind, it is much

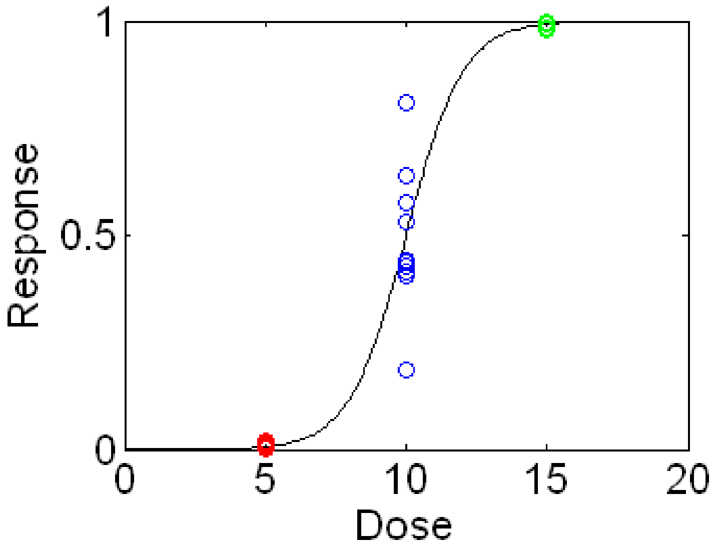


Figure 3.1: A hypothetical dose response curve showing the response from dosing at 5, 10, and 15 when the dose is normally distributed with a variance of 1. This illustrates the obvious result that there is much more variance in the response when the dose level varies around 10.

easier to understand results such as that shown in Figure 3.2.

There have been some ad hoc attempts to deal with differences in variance between genes (Delmar *et al.*, 2005), but, to date, there has been no attempt to

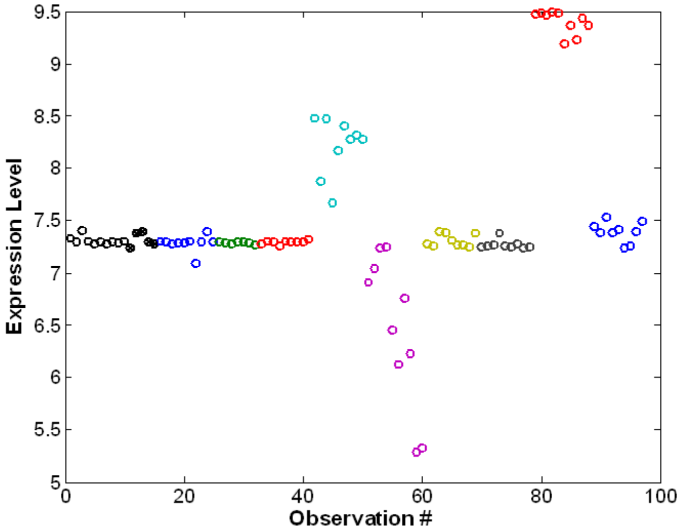


Figure 3.2: A scatter plot of the expression levels of the gene CDKn2a across all 97 samples from the oncogene experiment described in section 1.4. There are clear changes in the variance of expression levels for E2F3, Ras, and E2F1.

model differences in variance across treatment groups within a gene in general models. We will present a model for doing just that in a selective manner which allows the assumption of constant variance to be kept when appropriate.

3.1 Sparsity Priors and Variance Modeling

In the ideal situation, one would like to detect and adapt to differences in variance only when they are present. This would allow us to keep the lower false discovery rate associated with the assumption of constant variance when appropriate. In fact the tools for accomplishing this are already available to us in the form of variable selection. Rather than the more typical inverse gamma distribution for the variance of the zero mean error, one might assume a log-normal distribution. This variable can then be modeled in exactly the same way as any other normal random variable including regression with variable selection priors. For example,

$$\begin{aligned}
y_{g,i} &\sim N(\beta_{g,0} + x_i\beta_{g,1}, \psi_{g,i}),\\
\frac{\log(\psi_{g,i})}{2} &\sim N(\Delta_{g,0} + x_i\Delta_{g,1}, \rho),\\
\Delta_{g,1} &\sim (1-q)\delta_0 + qN(0,\tau).
\end{aligned}$$

However, we are now apparently free to place any type of hierarchy structure on the coefficients Δ of our model for variance. In particular, we may use the hierarchical variable selection priors such as those described in chapter 2. This will accomplish our goal of modeling changes in variance only where appropriate. That is,

$$\begin{aligned}
\Delta_{g,1} &\sim (1-q_g)\delta_0(\Delta_{g,1}) + q_g N(\Delta_{g,1}|0,\tau),\\
q_g &\sim (1-s)\delta_0(q_g) + sBe(q_g|\alpha,\gamma),\\
s &\sim Be(\alpha_s,\gamma_s).
\end{aligned}$$

Given $\lambda_{g,i} = \log(\psi_{g,i})/2$, model fitting proceeds exactly as in standard variable selection. In fact, computer code may be reused exactly with no changes, thus implementation complexity is not increased at all for these elements of the model.

There are, however, two important hyper-parameters to consider when modeling variance in this way, ρ and τ, and some additional complexity crops up in the study of these parameters. The default technique in each case would be to insert an inverse gamma prior for each of these parameters, and let the computer do its thing. However, it turns out that this is the wrong decision in both cases.

3.1.1 The Conditional Density for ψ

The complete conditional posterior for each $\psi_{g,i}$ is the product of an inverse gamma distribution and a log-normal distribution. Based on the model,

$$\begin{aligned} y_{g,i} &\sim N(x_i'\beta_g, \psi_{g,i}), \\ \frac{\log(\psi_{g,i})}{2} &\sim N(x_i'\Delta_g, \rho). \end{aligned}$$

First consider the contribution from the inverse gamma distribution:

$$\begin{aligned} y_{g,i} &\sim N(x_i'\beta_g, \psi_{g,i}), \\ \frac{y_{g,i} - x_i'\beta_g}{\sqrt{\psi_{g,i}}} &\sim N(0,1), \\ \frac{(y_{g,i} - x_i'\beta_g)^2}{\psi_{g,i}} &\sim \chi_1^2, \\ \log|y_{g,i} - x_i'\beta_g| = \lambda_{g,i} &= \frac{\log(\psi_{g,i})}{2} + \kappa_{g,i}, \\ \lambda_{g,i} &= z_{g,i} - \kappa_{g,i}. \end{aligned}$$

where the $\kappa_{g,i}$ are 1/2 of the log of a χ_1^2 random variable, $\lambda_{g,i} = \log(\psi_{g,i})/2$ and $z_{g,i} = \log|y_{g,i} - x_i'\beta_g|$.

Thus, the full conditional density for $\lambda_{g,i}$ is:

$$\begin{aligned} p(\lambda_{g,i}|-) &\propto f_\kappa(z_{g,i} - \lambda_{g,i}) \cdot N(\lambda_{g,i}|x_i'\Delta_g, \rho), && (3.1) \\ &\propto e^{z_{g,i}-\lambda_{g,i}} e^{-\frac{1}{2}\exp(2(z_{g,i}-\lambda_{g,i}))} \cdot e^{-\frac{1}{2\rho}(x_i'\Delta_g - \lambda_{g,i})^2}. && (3.2) \end{aligned}$$

Equation 3.2 gives us the density we need to perform updates of $\psi_{g,i}$ during Markov Chain Monte Carlo. Specifically, in order to update $\psi_{g,i}$ we first draw $\lambda_{g,i}'$ from a random walk distribution, $N(\lambda_{g,i}, .2)$. (It was determined that a standard deviation of .2 lead to an acceptance rate of around 40%.) We then accept a move to $\lambda'_{g,i}$ with probability

$$min\Big(1, \frac{p(\lambda_{g,i}|-)'}{p(\lambda_{g,i}|-)}\Big).$$

The variance, $\psi_{g,i}$, is then set to $exp(2\lambda_{g,i})$.

3.1.2 Prior Distributions on Variance Hyper-parameters within the Variance Model

Notice that ρ is a parameter that will allow for changes in variance across the samples in the same experimental group. Since one usually assumes, for good reason, that observations in the same experimental group come from the same distribution, it is clear that we want this parameter to be quite small. Additionally, consider that ρ is truly a nuisance parameter which will provide no information about our data. Finally, $\psi_{g,i}$ are variables in the model, unlike $y_{g,i}$. Thus, seemingly reasonable priors on ρ will tend to cause the $\psi_{g,i}$ to shrink to a common mean. This in turn will lead to a decrease in ρ, thereby strengthening the tendency of the $\psi_{g,i}$ to shrink to a common mean, as shown in the trace of ρ (Figure 3.3).

This tendency of ρ to shrink to zero can have disastrous consequences for the convergence of other parameters. The traces for the $\Delta_{1,1}$ parameters from two different MCMC runs of the exact same data are shown in Figure 3.4. Note that one goes to zero just after step 2000 while the other converges to a non-zero distribution.

There is a separate issue at work here as well. Notice that we are updating each of our $\psi_{g,i}$ around a mean $\Delta_{g,0} + x_i\Delta_{g,1}$ and subsequently updating $\Delta_{g,0}$ and $\Delta_{g,1}$

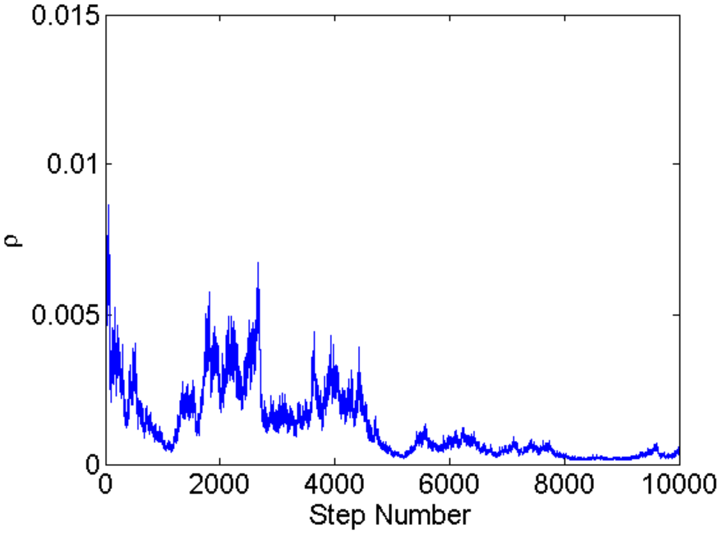


Figure 3.3: The trace of the parameter ρ (from equation 3.1) when it is given an inverse gamma prior. This trace was generated on a generated data set of size $p = 20$ and $n = 20$. Depending on the initial data, there is a tendency for ρ to limit to zero.

given $\{\psi_{g,i}\}$. If, for example, we have 20 observations in our experimental group, we will find that, even if convergence is possible, it will be very slow because each of the 20 separate $\psi_{g,i}$ parameters must move toward its correct distribution with a step size that is partially determined by ρ while at the same time remaining near the other $\psi_{g,j}$. When ρ decreases to zero at the same time, we find that the model never converges (as in figures 3.3 and 3.4).

3.1.3 Prior on ρ

The solution to these problems comes in two parts. First note that the size of ρ will have a direct bearing on the level of variation in the estimates of ψ_i within a trial. Due to the nature of variable selection, higher variability will lead to lower posterior probability that $\Delta_1 \neq 0$. We generated a data set consisting of two

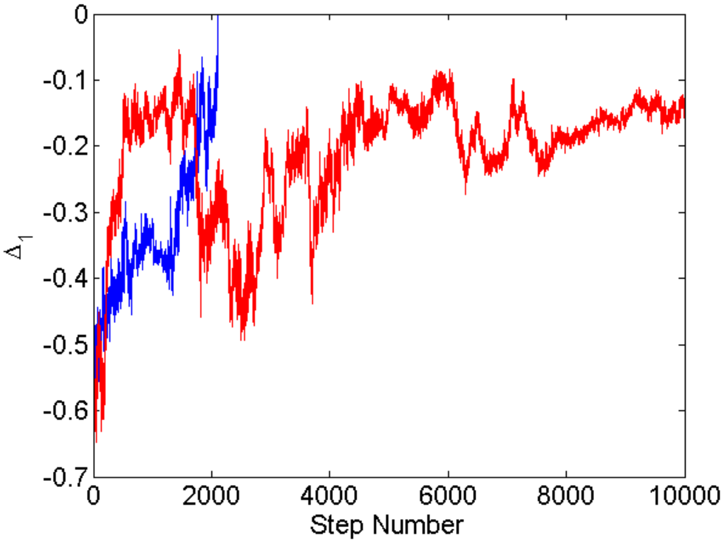


Figure 3.4: When ρ does limit to zero, as in Figure 3.3, this can have disastrous consequences for the convergence of the variance parameters. Here we see the traces for the parameter $\Delta_{1,1}$ from two separate runs on exactly the same data used to generate Figure 3.3.

groups of 10 each and 500 "genes", of which all 10×500 from the control group were generated from a standard normal, and all 10×500 from the experimental group were generated from $N(0, .5)$. Figures 3.5 and 3.6 show the posterior probabilities on this data set for $\Delta_1 \neq 0$ when ρ is set to constant value of .1 and .5 respectively. We find empirically that: 1) when there are 10 observations in both control and experimental groups, 2) the prior probability of variance difference is set to .5, and 3) ρ is set to .2, that there is about a 50% chance of detecting a doubling or halfing of the variance from control to experimental observations.

3.1.4 Independent Movement of ψ_i

We have an exact formula for the conditional probability of $\lambda_{g,i}|z_{g,i}, \Delta$ (see equation 3.2), so we can use a random walk Metropolis step to update it (and hence

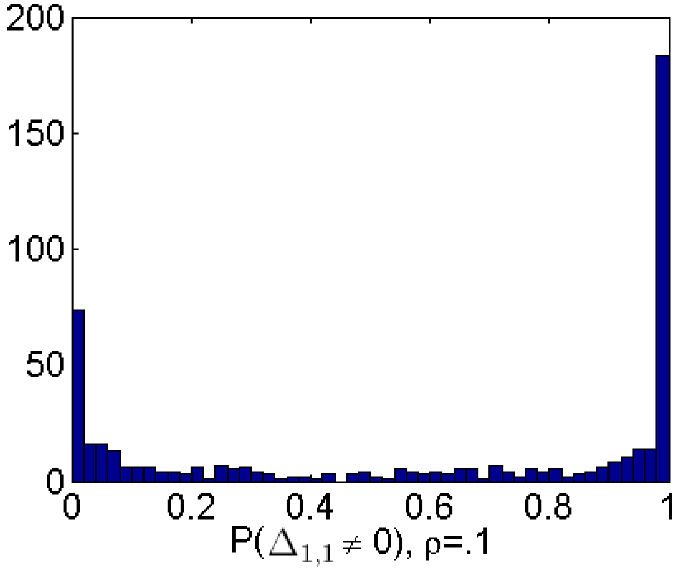


Figure 3.5: When ρ is small, there is a high chance of finding variance differences in the data set.

$\psi_{g,i}$). First we set $\lambda_{g,i} = \log(\psi_{g,i})/2$. We then draw $\hat{\lambda}_{g,i} \sim N(\lambda_{g,i}, .2)$. The standard deviation of .2 was chosen to induce approximately a 40% acceptance rate. The proposed new value, $\lambda'_{g,i}$ is then accepted with probability $\min\left(1, \frac{p(\hat{\lambda}_{g,i})}{p(\lambda_{g,i})}\right)$. If it is accepted, then we set $\psi_{g,i} = e^{2\lambda'_{g,i}}$.

With the model and updating scheme as we have described it so far, we have eliminated the issue (shown in Figure 3.3) of convergence to zero. This, however, does not entirely address all of the convergence issues.

3.1.5 Convergence

While fixing ρ will control the problem of non-convergence to some extent, there is still an issue when there are multiple observations moving individually, but also constrained to be near each other. Additionally, notice that we are poten-

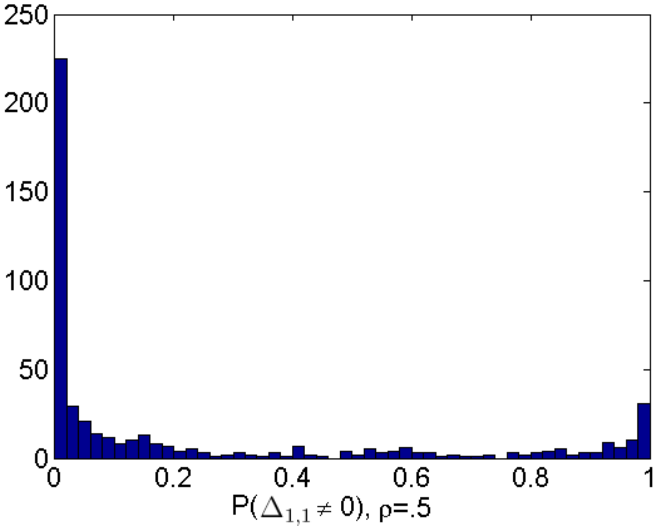


Figure 3.6: When ρ is large, there is less chance of finding variance differences, as compared to Figure 3.5.

tially inducing multimodality in $\Delta_{g,1}$ with our point mass mixture prior. This will translate directly into multimodality in the posterior of $\psi_{g,i}$ with the two modes corresponding to $\Delta_{g,1} = 0$ and $\Delta_{g,1} \neq 0$. If we can move all of the $\psi_{g,i}$ simultaneously, the problem with the convergence rate can be substantially improved, and if we can do this with a Gibbs step, we can solve the problem of movement between modes. Consider the conditional posterior (from above) for $\lambda_{g,i} = \log(\psi_{g,i})/2$,

$$p(\lambda_{g,i}|-) \quad \propto \quad e^{z_{g,i}-\lambda_{g,i}} e^{-\frac{1}{2}\exp(2(z_{g,i}-\lambda_{g,i}))} \cdot e^{-\frac{1}{2\rho}(x_i'\Delta_g-\lambda_{g,i})^2},$$

and suppose that we use an improper uniform prior distribution for $\Delta_{g,1}|(\Delta_{g,1} \neq 0)$. We will introduce a new latent "offshift" variable, δ_g, by which amount we will simultaneously move $\psi_{g,i}$ and $\Delta_{g,1}$. It is important to note that this step can only be taken when $\Delta_{g,1} \neq 0$. Then the conditional density for δ_g is:

$$\begin{aligned}
p(\delta_g|-) &\propto \prod_i e^{z_{g,i}-\lambda_{g,i}} e^{-\frac{1}{2}\exp(2(z_{g,i}-\lambda_{g,i}))} \cdot e^{-\frac{1}{2\rho}(x_i'(\Delta_{g,1}+\delta_g)-(\lambda_{g,i}+\delta-g))^2}, \\
&= \prod_i e^{z_{g,i}-\lambda_{g,i}} e^{-\frac{1}{2}\exp(2(z_{g,i}-\lambda_{g,i}))} \cdot e^{-\frac{1}{2\rho}(x_i'(\Delta_{g,1}+\delta_{g,i})-(\lambda_{g,i}+\delta_g))^2}, \\
&\propto \prod_i e^{z_{g,i}-\lambda_{g,i}} e^{-\frac{1}{2}\exp(2(z_{g,i}-\lambda_{g,i}))}, \\
&= e^{n\delta_g} e^{-1/2\sum_i \exp(2(z_i-\lambda_{g,i}))\exp(2\delta_g)}, \\
&\propto e^{n\delta_g} e^{-1/2K\exp(2\delta_g)},
\end{aligned}$$

where $K = \sum_i \exp(2(z_{g,i} - \lambda_{g,i}))$. Now, if we let $\nu_g = \exp(\delta_g)$ we find

$$\begin{aligned}
p(\nu_g|-) &\propto \nu_g^n e^{-1/2K\nu_g^2}(1/\nu_g), \\
&= \nu_g^{n-1} e^{-1/2K\nu_g^2}.
\end{aligned}$$

Now we may draw $\nu_g^2 \sim Ga((n+1)/2, K/2)$ and set $\delta_g = \log(\nu_g)$. A similar assumption of an improper uniform prior on $\Delta_{g,0}$ allows us to move in the opposite direction. Thus, we can move between the two modes in the posterior distribution of $\psi_{g,i}$ that occur due to the multimodality of $\Delta_{g,1}$ and, at the same time, improve convergence significantly.

We ran 1,000 different simulated trials, each of them on two different MCMC chains. Figure 3.7 demonstrates convergence of the separate chains after 10,000 steps, assuming the use of the Gibbs step allowing joint movement of $\psi_{g,i}$ and $\Delta_{g,1}$. Figures 3.8, 3.9 and 3.10 are the results from running the test without the step allowing joint movement of $\psi_{g,i}$ and $\Delta_{g,1}$. The three figures show non-convergence after 10,000 steps, 20,000 steps, and 40,000 steps respectively. Points where there is disagreement by more than 20% are marked with red x's. Note that the three experiments showing non-convergence were performed independently of each other. The 10,000 steps from the shortest of the three are not a subset of the steps from either of the other two.

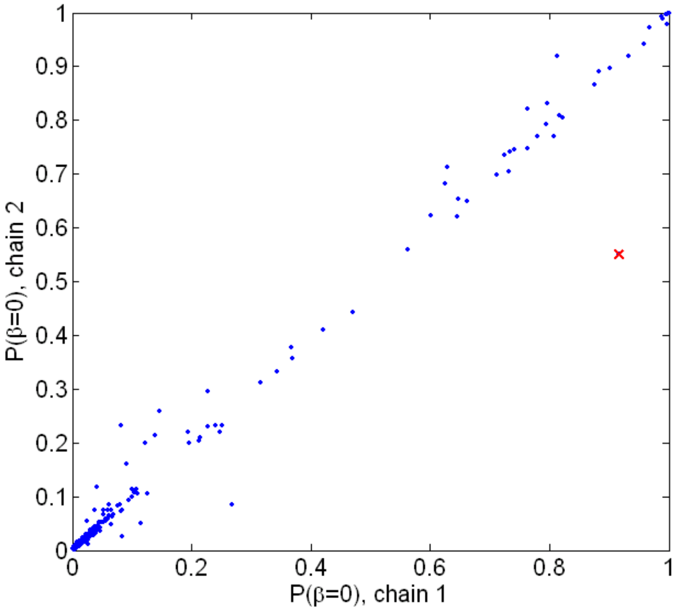


Figure 3.7: Convergence of two separate chains after 10,000 steps, and tested on 1,000 different trials. This experiment uses the joint move of $\psi_{g,i}$ and $\Delta_{g,1}$. The red 'x' makes the single point for which the posteriors disagree by more than 20%.

3.2 Experimentation

We have introduced a model for variance in a linear model which will allow for non-constant variance only where it is appropriate. We have demonstrated calculation for this model, and shown how to avoid certain convergence issues. We will now discuss some of the results that come from using this model. In section 3.2.1, we will discuss some simulated data sets, and demonstrate differences in posteriors between standard hierarchical models and ours. In section 3.3 we will show the

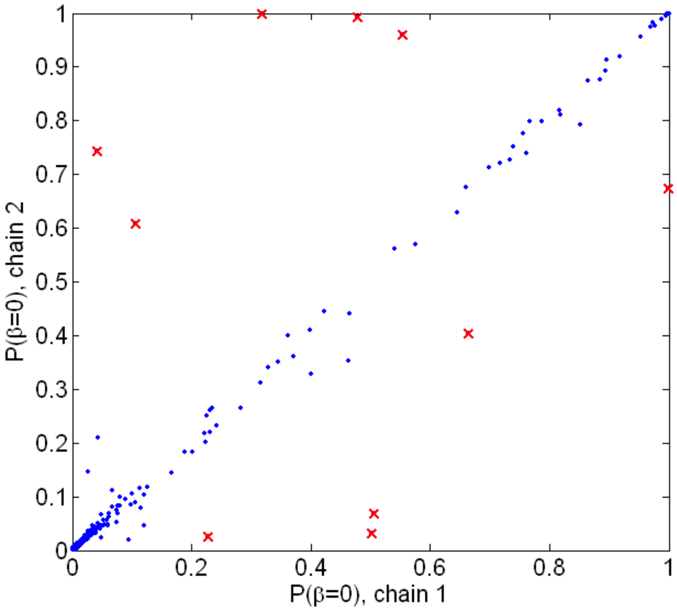


Figure 3.8: If the joint move of $\psi_{g,i}$ and $\Delta_{g,1}$ is excluded, there is significant non-convergence after 10,000 steps. Red x's mark points at which the posteriors disagree by more than 20%.

consequences of using the model in a real gene microarray experiment.

3.2.1 Simulated data

Consider the following experiment with a simulated data set. We will use 1000 "trials", each consisting of 10 "control group" observations and 10 "experimental group" observations. The control group observations will be sampled from a $N(0,1)$ distribution in all trials. For 700 of the trials, the experimental group will be sampled from $N(0,1)$ and the remainder will be sampled from $N(.75,.5)$.

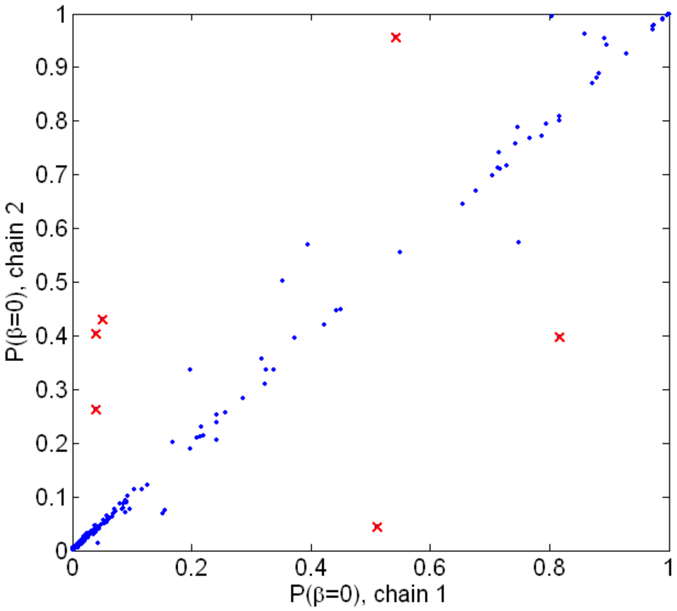


Figure 3.9: The same experiment as shown in Figure 3.8, but this shows non–convergence after 20,000 steps.

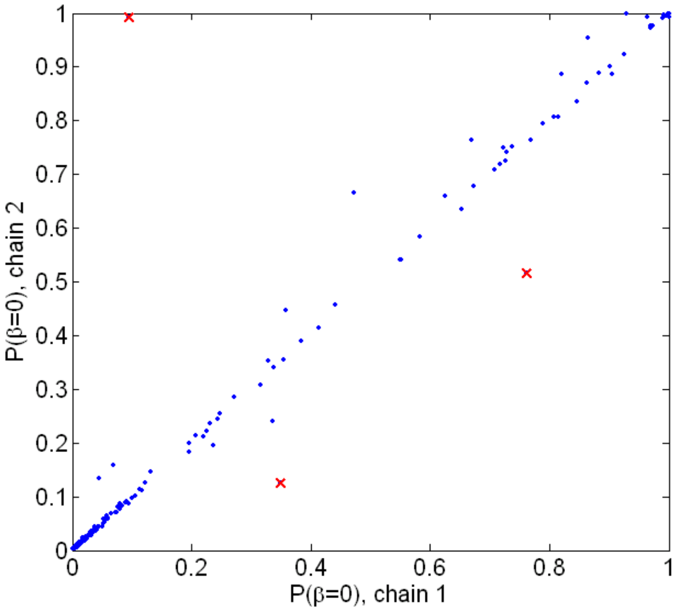
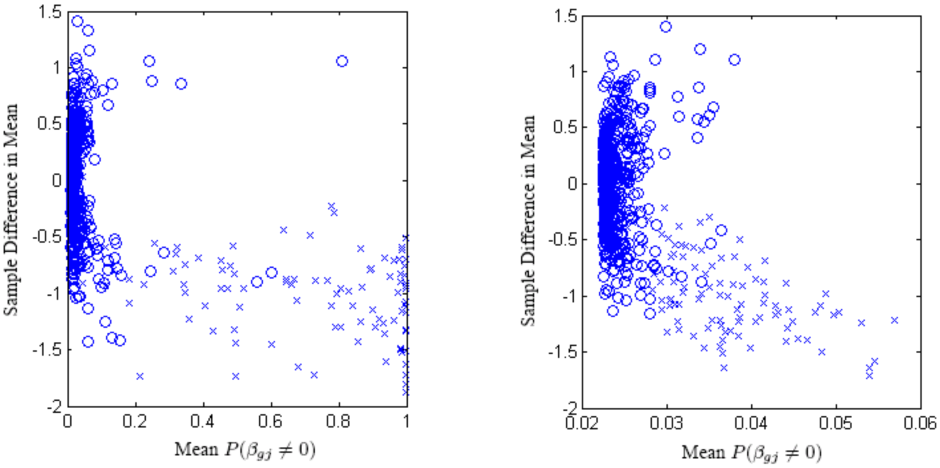


Figure 3.10: The same experiment as shown in figures 3.8 and 3.9, but this shows non-convergence after 40,000 steps. Note: the three non-convergence experiments were independent of each other.



(a)

(b)

Figure 3.11: Figures (a) and (b) show the posteriors from the full variance model and the constant variance model. The 'x' represents data that was generated with a different mean. The data was generated as described in Section 3.2.1. In (b) we see that for all g, $P(\beta_{g,1} \neq 0|-) < .06$. Modification of the prior to raise these probabilities leads to significant type I error.

We find experimentally that we pick up much more information about the data with our inclusion of a model for variance (Figure 3.11a). Without our variance model (Figure 3.11b), there is very little separation between the trials with an effect and those without. Experimentally, we find that increasing the discovery rate artificially through strong priors induces a high false discovery rate.

3.3 A DNA Microarray Experiment

One of the most prevalent things one finds microarrays being used for is to determine what genes are in a particular pathway. For example, to study which genes are affected (up or down regulated) by a particular promoter one might grow two cell lines, one control and one with an artificially up-regulated promoter gene. Other examples include comparing cells grown in different media, in the presence or absence of toxic (or other) substances, and in high or low oxygen concentrations.

In one such study, introduced by Bild *et al.* (2006) and described in section 1.4, we have 15 control group observations with between 7 and 10 observations from each of a set of nine experimental groups. The non-control groups have up-regulated Myc, Src, β-Catenin, E2F3, Ras, P53, AKTA, E2F1, and P110.

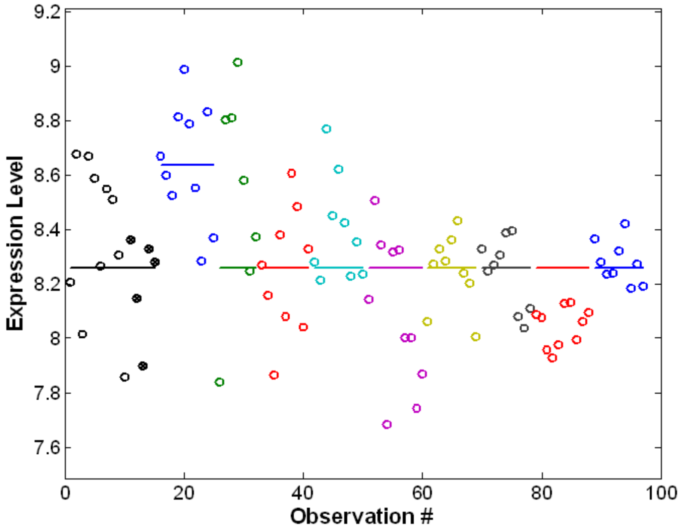


Figure 3.12: Without the possibility of a difference in variance, the standard model is unable to find a difference in means between control and the E2F1 experimental group. The colors represent the different experimental groups.

Among the genes on the microarray is the gene for Calcium-Transporting ATPase, ATP2A2. This gene is known to be a coregulator, along with some members of the E2F family, of the expression of many genes. Additionally, it is known that retinoblastoma-related protein, p107, regulates the expression of some E2F genes (Zhu *et al.*, 1993) as well as the expression of ATP2A2. Thus, in an experiment in which E2F1 is artificially up-regulated, it is stands to reason that some feedback loop would raise the levels of p107 in an attempt to lower expression levels of E2F1.

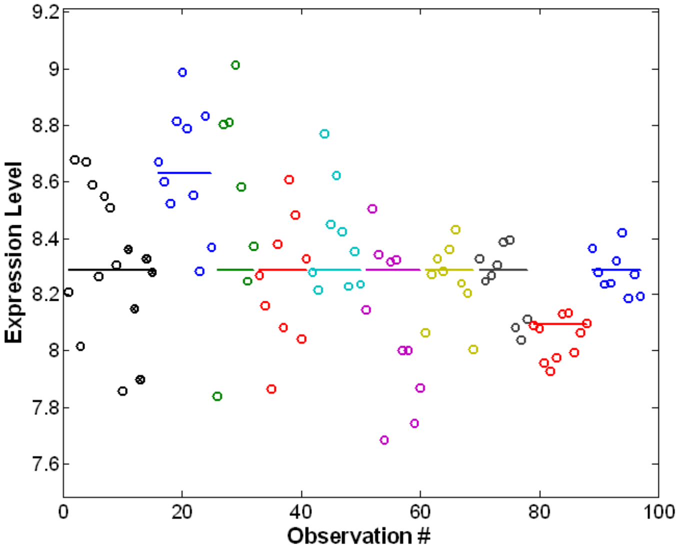


Figure 3.13: A model that includes variable selection on the variance terms will allow the detection of small shifts in mean if there is a simultaneous decrease in variance. With a variable selection prior for the variance model, there is a $> 95\%$ posterior probability of a difference in mean between control and the E2F1 experimental group.

A side effect of this happening would be the down-regulation of ATP2A2. Figures 3.12 and 3.13 show the results from a standard model with constant variance and from a variable selection model for variance, respectively. If π is the posterior probability that $\beta_1 \neq 0$, then the lines are the posterior mean for $\beta_0 + I_{\pi > .95}\beta_1$, where I is an indicator function for $\pi > .95$. In the case of the standard model, the posterior probability that there is a difference in means between control and the E2F1 experimental group was less than 20%.

3.4 Summary

Because of the complexity and volume of new biological data, every effort must be made to maximize the power and minimize false discovery regardless of model complexity. We have been able to make a significant improvement to the current state of the art hierarchical model which accounts for possible changes in variance across the different groups within an experiment while at the same time retaining the assumption of constant variance where appropriate.

With this new model, we are able to show clear improvements in the analysis of both simulated and real experimental data. Difficulties with fitting and convergence of the Markov Chain Monte Carlo algorithm introduced by the use of a variable selection in the variance model have been overcome, and, in future sections, we will be able to show how this model can effectively identify and estimate effects on collections of pathway genes, thus creating metagenes that provide robust predictions of practically relevant phenotypes, such as survival in real clinical situations.

Chapter 4

Internal Consistency in Microarray Experiments

Advances in both computer science and biology have led to a remarkable ability to measure vast numbers of messenger RNA levels simultaneously in cell samples. Because of these advances, researchers are able to study the effects of interventions at the cellular and gene expression level. Recent studies have demonstrated the potential of gene arrays as tools for drug discovery and categorization (Lamb *et al.*, 2006), identification of cellular response to intervention, and even as predictors of clinical outcomes (Bild *et al.*, 2006; Miller *et al.*, 2005). Excitement about these advances has been tempered recently due to the difficulty of obtaining consistent results.

Through experimentation, it has become clear that the data obtained are extremely sensitive to the conditions under which the RNA is hybridized to the DNA on the micro-array chip. Minute changes in pH, temperature, or other conditions lead to differences in measured RNA levels that overshadow any difference that might be associated with treatment or intervention. Additionally, the response of a particular probe varies depending on its composition. The four different nucleic acids respond differently to environmental conditions.

This high sensitivity to experimental conditions can be accounted for by concurrently running the assay on "control" and "experimental" cell samples and examining only the difference in expression levels. Unfortunately, the need for a control data set that was taken in close proximity, spatially and temporally, to the experimental data removes most of the impetus for performing large, multi-institution studies, as the results from one group may only be compared to the controls measured by that same group and at the same time. Additionally, because the actual expression level of a particular gene is relatively meaningless, the use of microarrays in a clinical setting is precluded.

4.1 Assay Artifact Controls for Microarrays

On the Affymetrix gene array chips, there are included a number of "housekeeping" genes. These probes are designed for chip calibration and for these genes one expects to observe a constant expression level across all observations, regardless of treatment or intervention. Often, however, one finds that this is not the case. Consider the expression levels observed for the housekeeping gene "AFFX-BioB-5_at" (shown in Figure 4.1), which shows vast differences in expression level.

Let $y_{g,i}$ be the expression level of gene g from sample i, where $1 \leq i \leq N$ when there are N total samples. Also, let y_g be the row vector with elements $(y_{g,i})$. Then, for the housekeeping genes, we believe that $y_{g,i} \sim N(y_{g,i}|\mu_g, \sigma_g)$, where μ_g is a gene specific mean expression level and σ_g is a gene-specific variance of expression.

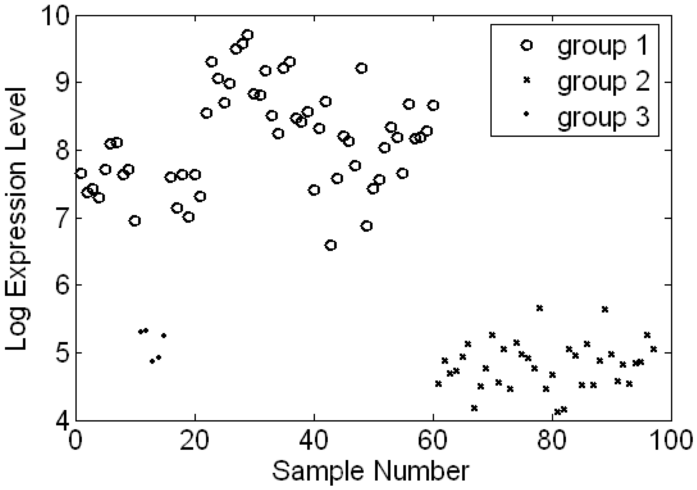


Figure 4.1: The expression levels across 97 samples of the gene "AF-FX-BioB-5_at", a housekeeping gene on the Affymetrix U133+ chip. The three different groups represent measurements that were taken at different times. What should be a constant expression level shows clear correlation with the time at which the sample was taken.

Suppose now that there are a small number of different environmental conditions, changes in which might explain variation like that seen in Figure 4.1. Additionally, due to the relative levels of nucleic acids in the DNA probes, there is a different level of response to these environmental conditions. Consider a model in which we introduce a latent $d \times n$ matrix, W, which describes these environmental conditions, and a latent d-vector, $\gamma_g \sim N(\gamma_g|0, I)$, which describes the contribution to the expression level of $y_{g,i}$ from each of the d columns of W. Our model then becomes

$$\begin{aligned} \gamma_g &\sim N(\gamma_g|0, 1) \\ y_{g,i}|\gamma_g &\sim N(y_{g,i}|\gamma_g' W + \mu_g, \sigma_g) \end{aligned}$$

It is exactly this model, given the appropriate limiting assumptions on prior hyper-parameters, that leads to classical principal components analysis (Bishop, 1999; West, 2003). Specifically, the maximum likelihood estimate (MLE) for the matrix W, assuming a fixed d, consists of the first d principal components of the matrix with element $(y_{g,i})$ in its g^{th} row and i^{th} column. Thus, under the modeling assumptions outlined above, and using an appropriate subset of genes from the gene microarray, we might hope to discover, and thereby correct for, any biases introduced by differing conditions on the lab bench.

Fortunately, the housekeeping genes provide us with exactly such a subset. There are 67 such housekeeping genes in the U133+ chip. For our experiment containing n samples ($n = 97$ in Figure 4.1), we extracted these genes. This gives us 62 observations from an n dimensional space. We center each of the n-vectors and calculate the principal components, keeping enough of them to account for 95% of the variability in expression level of the housekeeping genes. This gives us an algorithm for calculating the matrix W.

Suppose that we have some experimental $m \times d$ design matrix, H. Then we are interested in modeling $y_{g,i}$ in terms of this design matrix as:

$$y_{g,i} \sim N(y_{g,i}|\beta H', \sigma_g^2).$$

However, in the case of gene microarrays, and when the data collection process occurs across multiple labs or at different points in time, we expect significant issues with lab bias which will render this model inadequate. We have already computed the correction for this problem W, and its inclusion in the over-all model is direct:

$$y_{g,i} \sim N(y_{g,j}|H\beta_g + W\gamma_g, \sigma_g^2).$$

As we are dealing with microarrays, it is reasonable to expect that we are using some form of variable selection or sparsity inducing prior for $\beta_{g,j}$, such as

$\beta_{g,j} \sim (1-q)\delta_0(\beta_{g,j}) + qN(\beta_{g,j}|0,\tau_g)$. Because there is no clear standard, the choice of either the hyper-parameters, τ and q, or the structure of the hierarchical model describing those parameters will vary according to the experiment. However, regardless of the chosen structure, the parameters β_g and γ_g may be treated exactly the same. The caveat to this is rooted in our understanding of what these parameters describe. The parameters β_g describe the effects of an experimental intervention, and as such, are expected to be non-zero in only relatively few of the genes on the array. Conversely, the parameters γ_g describe bias that is introduced to the microarray as a whole due to conditions on the lab bench. It is expected that many, or even most, of these parameters will be non-zero. Differences or similarities in how these parameters are treated are important to consider, and the effects of the various hierarchical structures in variable selection models are a topic of considerable interest in current research.

4.2 Data Analysis Examples

The data sets we will use in this study are from Affymetrix U133+ 2.0 microarray chips. It is well established that there are results from these chips that correspond to zero expression in the cell samples under study. With this in mind, we will eliminate from the study any gene which shows an expression level lower than 6 across all samples in the study. Additionally, we are preferentially interested in genes which respond to the experimental intervention. Thus we will throw out any genes that show a maximum difference in expression across all samples of less than 2 on the $\log_2$ scale.

Our study focuses on two data sets. The first, used in Bild *et al.* (2006) for the elucidation of certain oncogenic pathways, consists of 97 samples from various experimental interventions performed on human primary mammary epithelial cells (HMEC's). This experiment was done in three stages, but all of the data was taken on Affymetrix U133+ 2.0 gene arrays. After use of the filter described in the previous paragraph, we are left with 10,777 genes in our study.

Originally, 10 samples were taken from a control group along with 7-10 samples from each of five experimental groups. In each of the experimental groups, a specific, known oncogene was artificially upregulated. The oncogenes in this group were MYC, SRC, BCAT, E2F3, and RAS. Some time later the decision was made to study the genes P63A, AKTA, E2F1, and P110. At this point it was discovered that there were systematic biases occurring between the two groups, and that the original control group was inadequate for the subsequent experiment. With this in mind, five more control samples were taken. There are potential issues here related to outcome dependent testing that are being ignored.

The results of this experiment on the measured levels of the housekeeping gene "AFFX-BioB-5_at" are shown in Figure 4.1. This is simply the first of the housekeeping genes in the list, and was not chosen for any specific features. All of the housekeeping genes show this type of variation across, and even within, the groups.

The second data set was introduced in Shi *et al.* (2006). It is publically available data, produced by the MicroArray Quality Control Project (MAQC) for the express purpose of studying the reproducability of results using microarray technology. The group performed identical assays at multiple different test sites using over 1,300 arrays from nearly all vendors. We will examine a subset of the data consisting of 120 Affymetrix U133+ 2.0 microarray chips. These were split into four experimental groups each with five replicates. The same experiment was then repeated at each of six sites. There were two cell lines used, Universal Human Reference RNA (UHRR) from Stratagene and a Human Brain Reference RNA (HBRR) from Ambion. The four experimental groups consist of various mixture ratios of RNA derived from these two cell lines (sample 1: all UHRR, sample 2: 3 to 1 UHRR/HBRR, sample 3: 1 to 3 UHRR/HBRR, and sample 4: all HBRR). The details of this data set are outlined in Shi *et al.* (2006). Note that we have reordered the labeling of these samples relative to that used in Shi *et al.* (2006). After filtering this data set, we are left with approximately 22,000 genes.

4.3 Analysis

As with any study of microarray data, we must choose a model for determining when there is a change in expression level between two study groups. We will use the hierarchical Bayesian variable selection model introduced earlier to fit our data sets. From this, we can derive full posterior marginal distributions for all relevant parameters, in particular, the posterior probability of a difference in mean and the expected difference in mean given that one exists.

Recall that we define $y_{g,i}$ to be the expression of gene g from sample i. If we let x_i be the design vector for sample i, then our full model is as follows:

$$
\begin{aligned}
y_{g,i} &\sim N(x_i'\beta_g, \psi_g), \\
\beta_{g,0} &\sim N(\beta_{g,0}|0, \tau_0), \\
\beta_{g,j} &\sim (1 - q_{g,j})\delta_0(\beta_{g,j}) + q_{g,j}N(\beta_{g,j}|0, \tau) \text{ for } j > 0, \\
1/\sigma_g &\sim Ga(\alpha_\sigma, \gamma_\sigma), \\
1/\tau &\sim Ga(\alpha_\tau, \gamma_\tau), \\
q_{g,j} &\sim (1 - r_j)\delta_0(q_{g,j}) + r_j Be(\alpha_q, \gamma_q), \\
r_j &\sim Be(\alpha_r, \gamma_r).
\end{aligned}
$$

The details of choosing hyperparameters and constructing a convergent Markov chain Monte Carlo algorithm for this model are discussed in the previous two chapters and in Lucas *et al.* (2005). There are, of course, many other variable selection models available. The PCA correction algorithm described here is directly applicable, regardless of model choice.

For the HMEC data, the experiment is designed to elucidate biological pathways. Each of the genes that are upregulated will cause a concurrent up or down regulation of the neighbors in its pathway. With this in mind, we expect relatively few genes to respond to each of the experiments. Additionally, we expect to see few genes that are in multiple pathways, and the ones that do show response to more than one treatment should be interesting, as they are at the border between the two pathways. Exceptions to this may be the MYC and RAS experiments, as it is known that MYC is directly regulated by RAS, and the E2F1 and E2F3 pathways.

We will examine the posterior distributions for the parameters $b_{g,j} = \beta_{g,j}|\beta_{g,j} \neq 0$ and $q^*_{g,j} = P(\beta_{g,j} \neq 0)$. Specifically, we will make the (admittedly arbitrary) decision that there is a change in gene expression for experimental group j and gene g when $q^*_{g,j} > .95$. Also, when there is such a change in expression, then $E(\beta_{g,0} + b_{g,j}|-)$ is the expected expression level from model fitting.

Because all of the genes in each of the pathways tested in this experiment are not known, we will look at two outcomes to argue for the efficacy of PCA correction. First, there are specific examples where gene regulation is known. Second, there is a general trend toward fewer positive results after PCA correction.

For the data from the MAQC project we are interested in RNA titration. In this experiment, samples were taken from two different sources. It is know that there are many genes that are differentially expressed between these two sources. Because we have data on different levels of mixing of the two samples, we expect to see results that are monotonic for genes that are differentially expressed.

We will determine that PCA correction is a success in this experiment by an increased probability of monotonic results. Shippy *et al.* (2006) argue for a measure of success based on the mean expression level across each sample showing a monotonic change. Also, by data exploration, they determined that one of the study sites produced outlier data, and discarded that data. In fact, by a quick study of the Affymetrix housekeeping genes (the first principal component is shown in Figure 4.2), it is clear that there are differences across all study sites. Also, the very purpose of PCA correction is to address systematic bias introduced by slightly different procedures at different labs.

Another undesirable feature of the technique used in Shippy *et al.* (2006) for evaluation of consistency is that there are 30 observations for each of the samples. This should be more than enough to determine a reasonably tight bound on the sample mean. Therefore, it is too blunt an instrument for demonstrating improve-

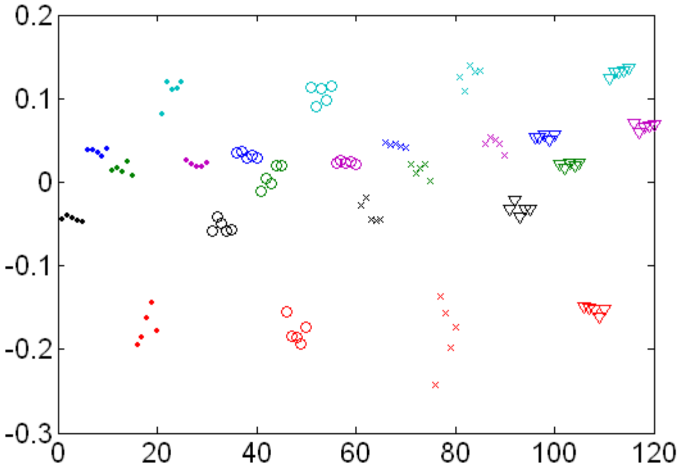


Figure 4.2: The first principal component of the housekeeping genes from the MAQC experiment. The x-axis is ordered by sample type, and results from each of the six study sites are shown in a different color.

ment in consistency before and after PCA correction. With these facts in mind we will propose a different metric. Specifically, we will use sampling to determine the probability of proper ordering if just one observation is taken from each sample type. This can be done with the raw data as well as the PCA corrected data, thereby providing a direct and relatively sensitive comparison.

Having derived an adequate measure of the success of our correction technique, we are left to decide which of the 54,000 genes on which to use our test. In particular, which of them are differentially expressed in UHRR versus HBRR. For the analysis done in Shippy *et al.* (2006), the question was sidestepped by performing the test on all genes, and plotting the results against the ratio of mean expression level of sample one to the mean expression ratio of sample four (which should show the largest difference). The choice is not so clear in our comparison study, because we must choose which mean expression ratio to use, the one obtained from the raw data, or the one obtained after PCA correction. The decision was further complicated by our results from studying the housekeeping genes. The forth housekeeping principal component is shown in Figure 4.3. There is a clear monotonicity to the results from all but one of the study sites. This leads to difficulty in interpretation of results from other genes, as it is unclear whether an observed change in expression level is due to site bias or expression differences among the four samples.

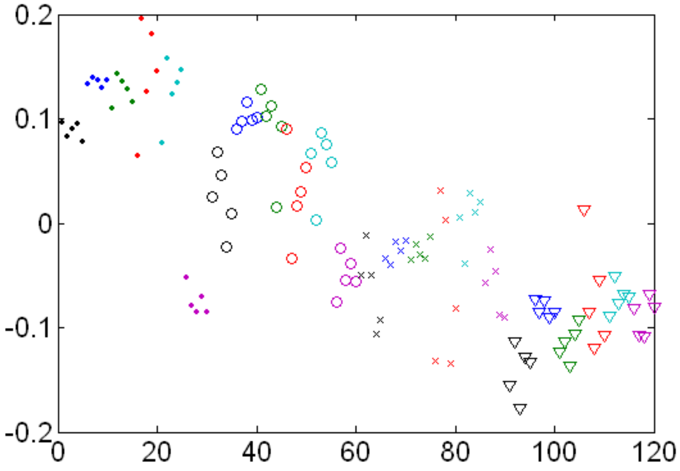


Figure 4.3: The forth principal component of the housekeeping genes from the MAQC experiment. The x-axis is ordered by sample type, and results from each of the six study sites are shown in a different color. There is a significant monotone trend across all study sites except one. This leads to significant difficulty in interpretation, as it becomes unclear whether an observed change in expression level is due to site bias or expression differences among the four samples.

With this in mind, we will use our model to determine which of the genes show significant change in expression across the samples. In Shippy *et al.* (2006), Shippy et al. examine the RNA concentration difference of RNA in the UHRR versus UBRR samples, and conclude that they are approximately 3% and 2% respectively. This leads to estimates of actual percent of UBRR RNA in samples 2 and 3 of 0.18 and 0.67 respectively. From this analysis, we add to our design matrix the column 120 dimensional vector which has 0 corresponding to sample 1, 0.18 corresponding to sample 2, 0.67 corresponding to sample 3, and 1 corresponding to sample 4. With the inclusion of a variable selection prior for this vector of the design matrix, we obtain a posterior probability of a difference in expression level. We will choose from these genes, those which show greater than a 95% posterior probability of having a sample effect on expression level.

Notice that there is the option here to include a latent variable in our model which describes the relative abundance of RNA in the two samples. We have not done this, but if we were interested in studying differential expression in these two cell lines, this would likely improve the robustness of our test.

4.4 Results

4.4.1 Temporal Separation: Gene Up-regulation in a Breast Cancer Cell Line

In the breast cancer cell line analysis, we find that there is a significant improvement in results after PCA correction. There is an overall drop of more than 60% in the total number of discovered changes in mean. Additionally, as shown in Figure 4.4, there is a drop in the expected number of pathways each gene participates in from 5.1 to 1.7 after PCA correction. Without PCA correction, we find that 65% of the 10,777 filtered genes are in 5 or more of the pathways under study. This is clearly absurd, as it is difficult to believe that even one of the pathways should include over 7,000 genes. It is, in fact, somewhat difficult to justify a number as high as 1.7. However recall that we have pre-filtered the data to screen out genes that show no response, and that all of the genes studied are know oncogenes. With this in mind, a close relationship between the gene pathways is believable.

Knowing that PCA correction has decreased our discovery rate from absurd to reasonable is encouraging, but it leaves open the question of whether we have done it in an appropriate way. To answer this question, we will look at some specific examples pulled from our data set.

Consider the expression level of the gene Nuclear factor (erythroid-derived 2)-like 1 (NFE2L1) shown before and after PCA correction in Figure 4.5. This example is a compelling endorsement for the correction algorithm. The reason for this is that NFE2L1 has a known MYC binding site in its promoter region Morrish *et al.* (2003).

More examples are shown in Figure 4.6. Parts (a) and (b) show E2F1, which is known to self promote, Parts (c) and (d) show cyclin A, which is known to be downstream of Myc Bazarov *et al.* (2001), and parts (e) and (f) show expression of RAF, which is known to be involved in replicative senescence along with p63a Jung *et al.* (2001). Notice that both E2F1 and Raf demonstrate signatures found even when the control data and the experimental intervention are done at different times.

It is difficult, without knowing the correct answers, to say for certain that the corrections are behaving as we would like. However, it is clear that the expression differences measured after correction are reasonable, and that the genes for which expression differences are found are worth studying carefully.

4.4.2 Microarray Quality Control

With the MAQC data, we know before hand that there are a collection of genes that will show a titration curve. With this in mind, and knowing the titration levels to be approximately 0.18, 0.67, and 1, we can model this data using an overall mean (which does not have a variable selection prior), a titration vector, and the housekeeping principal components.

We sampled at random 1,000 of the 22,000 genes that show both non-negligible expression levels and non-negligible changes in expression level across the study. We fit our model to these 1,000 expression vectors, and find that there is good empirical evidence that we are able to improve the titration signature. Figure 4.7 shows PCA correction when there are clear lab specific effects in the data. Notice that much of the artifacts are removed by PCA correction.

Figure 4.8 shows one of the genes overlayed with the contribution from fitting the principal components. Recall that W is the matrix with the principal component vectors W_k in its columns. Also, γ_k is the coefficient for W_k and q_k^* is the posterior probability that $\gamma_k \neq 0$. Then, the principal components points on the plot are at

$$pc_j = E(\beta_0) + \sum_k W_k \cdot \gamma_k \cdot I_{q_k^* > .95}$$

Notice that a significant part of the variance in expression of this gene is from systematic error introduced in the lab setting.

While this empirical evidence is reassuring, Figure 4.9 is more compelling. This shows that there is significant improvement in the probability of sampling one point from each of the four sample types and getting them in the correct order. The ordering of the four sampled points is improved in 77% of the tested genes, and there are no significant outliers below the diagonal (genes that are better ordered without correction).

4.5 Summary

We have demonstrated an algorithm, PCA correction, which provides a systematic technique for determining appropriate correction factors from a list of housekeeping genes. Because these genes are the only ones involved in the determination of the correction factors, and because they are subsequently thrown out of the analysis, PCA correction provides the opportunity for subtracting systematic bias without using the genes under study to define that bias. This leads to a good chance of strengthening a signal of interest without introducing biases that come from using the data itself to calculate corrections.

Our model has been shown to work on multiple different data sets. It demonstrates significant improvement in results on both a full genome scale, correcting false discovery rates, and on a gene by gene basis, leading to a much clearer distinction between signal and noise for specific genes of study such as NFE2L1.

Microarray data that has been collected at different time points or in different study sites show very clear biases, even in genes that should be expressed at a constant level. Because we are able to correct data that show these biases, we are making it much easier to compare the results of different experiments.

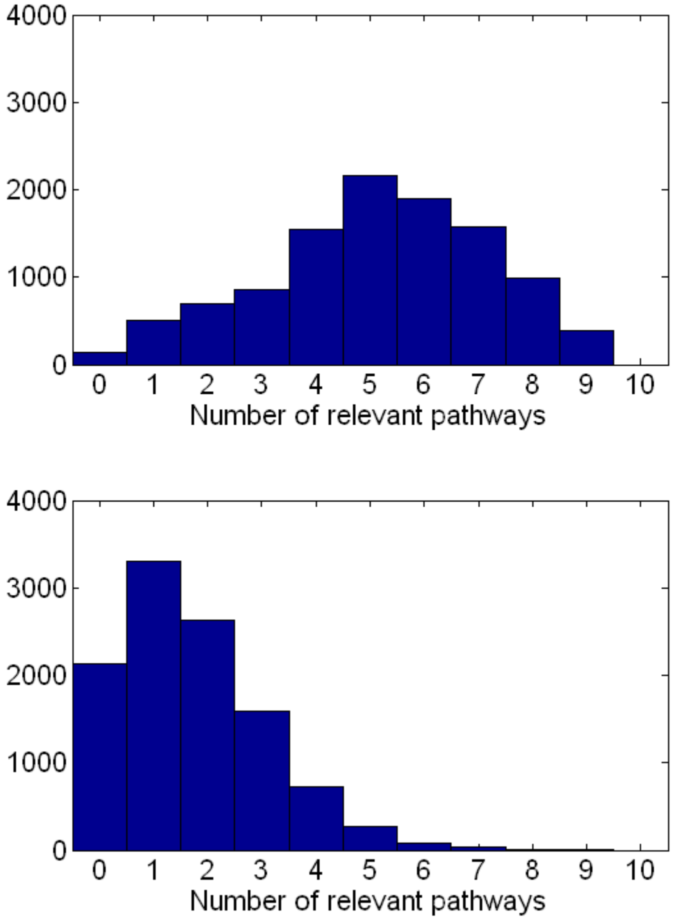


Figure 4.4: Histograms of the number of experimental groups from the breast cancer cell line example that show differential expression on a gene by gene basis. There is a drop in expected number of relevant pathways per gene from 5.1 to 1.7 after PCA correction.

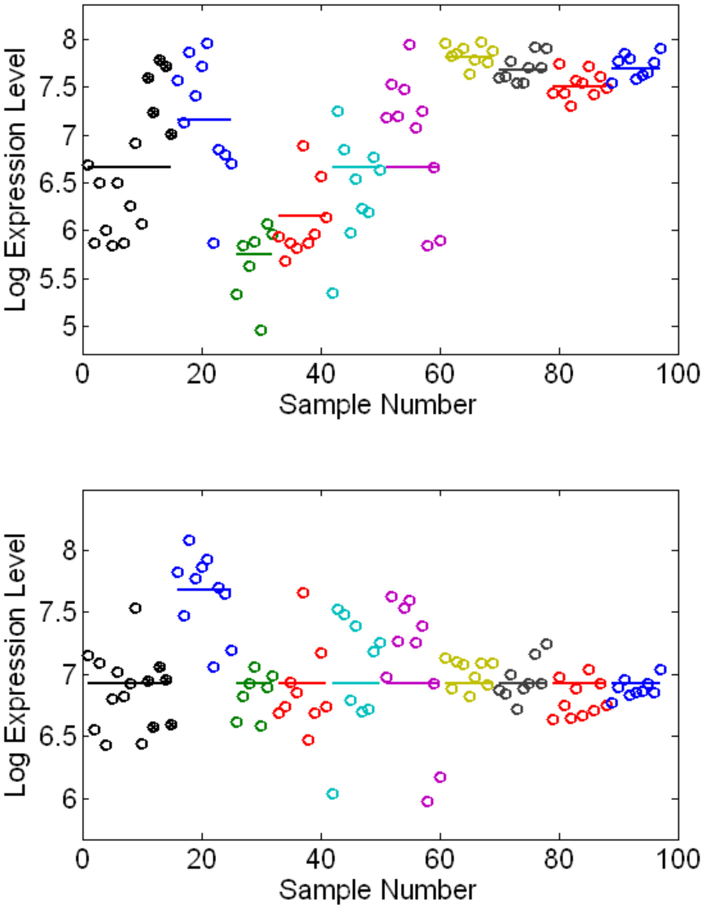


Figure 4.5: The expression levels from the breast cancer cell line study for "NFE2L1" before and after correction by PCA. The lines are drawn at the point $E(\beta_{g,0}) + E(b_{g,j}) \cdot I_{[q_{g,j}>.95]}$.

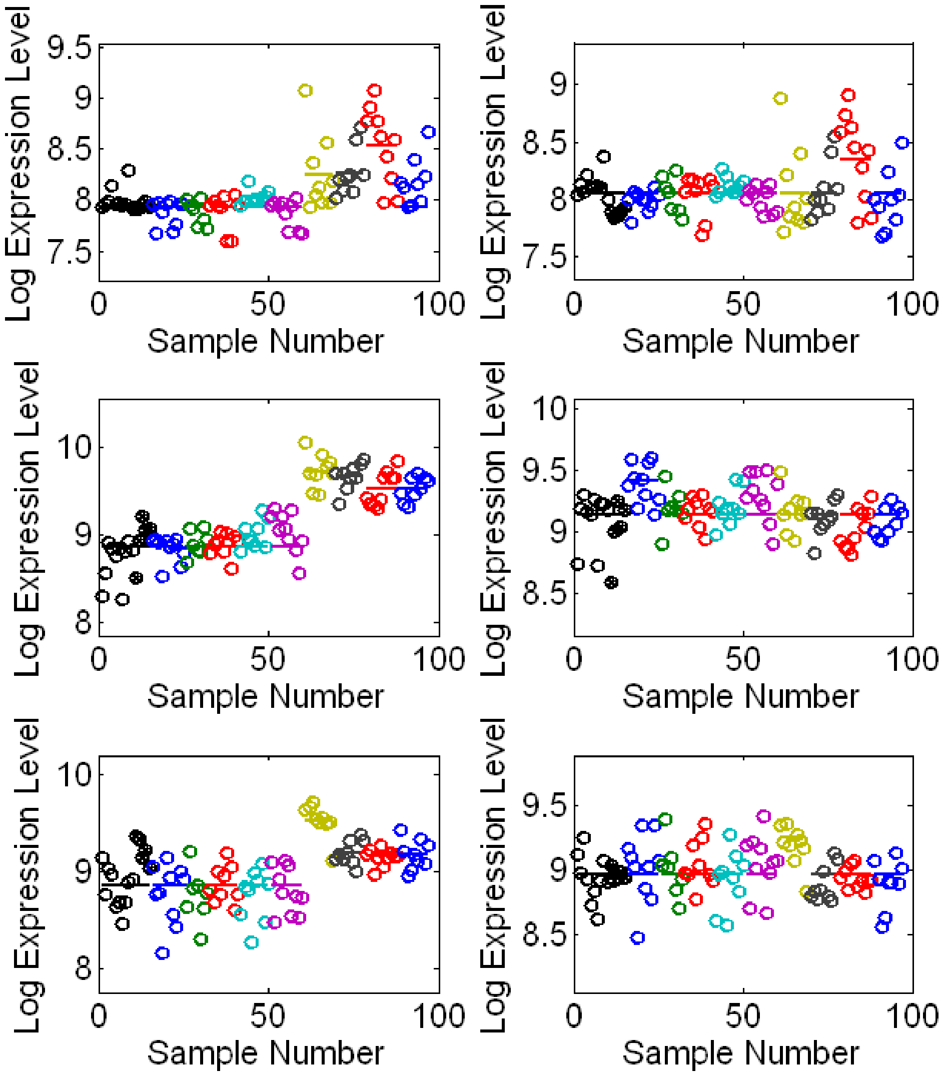


Figure 4.6: Breast cancer cell line study: expression of E2F1 (line 1), Cyclin A (line 2), and Raf (line 3) are shown to exhibit believable results after PCA correction. The left column shows before and the right column shows after correction.

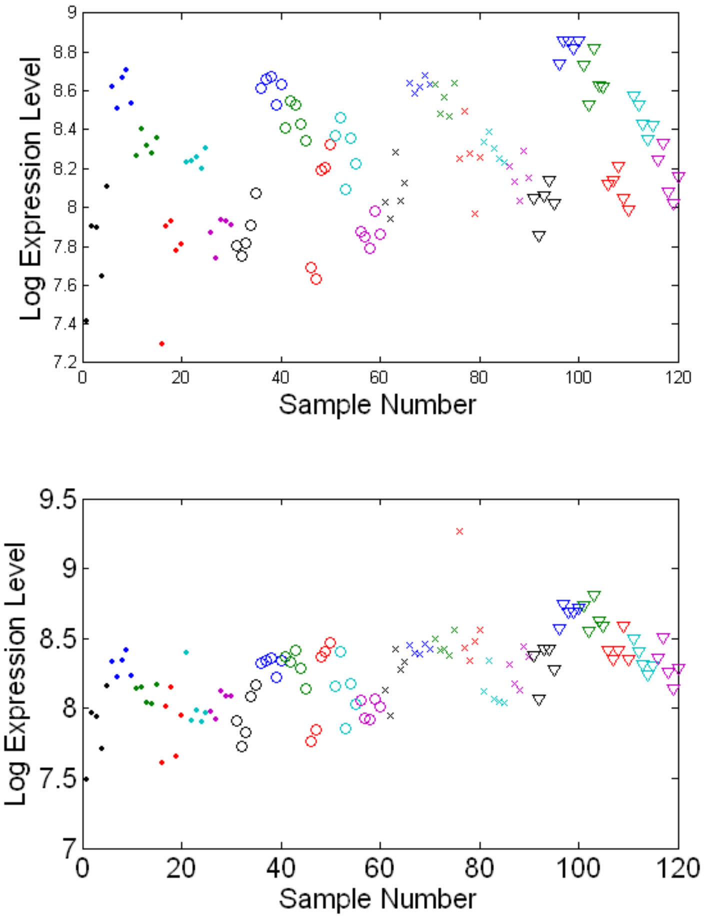


Figure 4.7: MAQC study: expression levels before and after PCA correction. Notice that most of the lab specific artifacts are corrected.

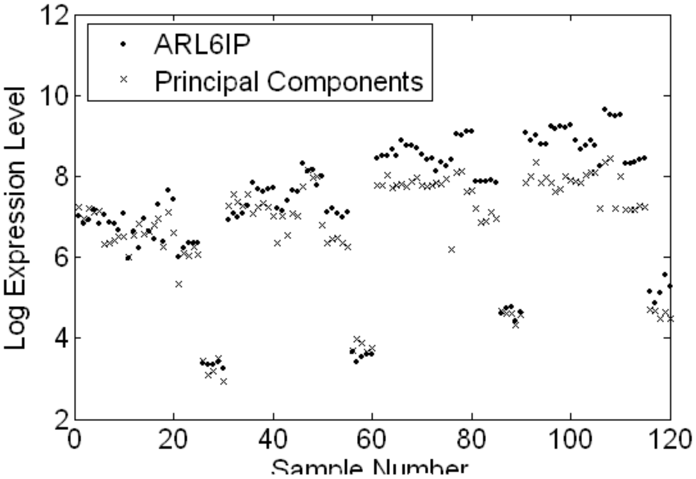


Figure 4.8: MAQC study: expression of the gene ARL6IP shown with the posterior mean of its principal component part. A significant part of the variance in expression of this gene is from systematic error introduced in the lab setting.

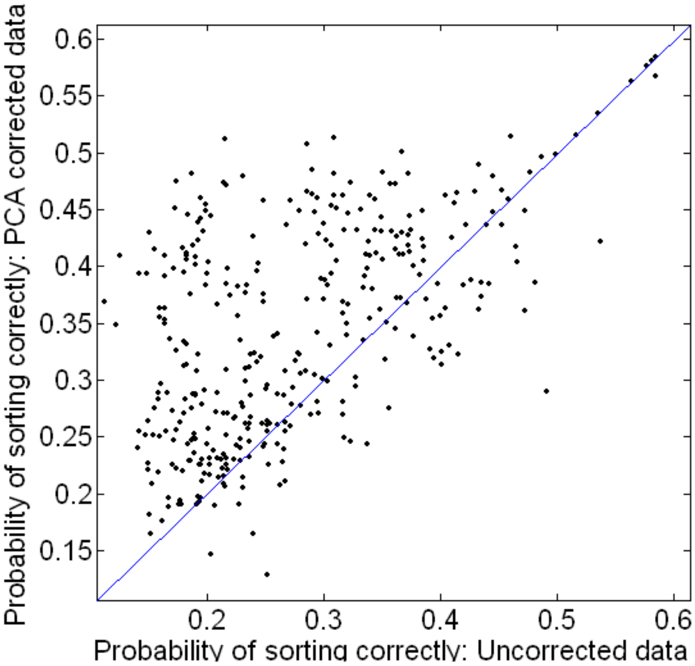


Figure 4.9: MAQC study: a comparison of the probability of obtaining the correct order when sampling one point from each of the four tissue samples. The x-axis shows these probabilities when the raw data is used, and the y-axis shows the same probabilities after PCA correction. Because a significant portion of the points are above the diagonal, we conclude that correction has improved our results.

Chapter 5

Case Study: Detection of Gene Pathways

Bayesian regression formulations of factorial designed experiments for gene expression data allow for routine analysis that addresses the core questions of identifying which model parameters - the effects of design factors and their interactions - are likely of relevance in the context of a model that has many such parameters as we move across thousands of genes. Standard shrinkage analysis, that has been the workhorse of applied Bayesian methods for decades, has been well-described and quite widely applied to two sample testing in gene expression studies (Scott and Berger, 2005) and its extension to more complex designed experiments and regression models is theoretically direct though requires some computational development (Broet *et al.*, 2002; Ishwaran and Rao, 2003; Do *et al.*, 2005). In general, these techniques have generated good results, but they do exhibit some unwanted features. First, these techniques deal with variance by either assuming it is constant across all samples, or making somewhat arbitrary variable transformations to make it so. Second, the structure of the models induces potentially unwanted correlation in posterior inclusion probabilities between genes. Finally, these models heretofore have ignored the presence of the “house keeping” genes in the data sets. These genes provide important information about the environment in which the chips were read, and can help to properly adjust the resulting data sets.

We will focus, in this chapter, on the implementation of the model innovations described in the previous chapters. They models are designed to tackle just these problems, and we will show that they perform adequately on an experimental microarray data set.

5.1 DATA

Our data consists of mRNA expression levels from a set of controls and nine different experimental tissues in which each of the genes GFP1, MYC1, SRC1, BCAT, E2F3, RAS1, P63A, AKTA, E2F1, and P110 were upregulated. Data for ten of the 15 controls was collected along with the first five experimental groups. Some time later the data for the last four experimental groups was collected. Finally, it was noted that there were systematic differences between the data obtained at the two different collection times. Thus another set of five controls was collected in the hopes that they would more closely compare to the second set of experimental data.

We set out to determine which genes are affected in each of the experimental interventions. We hope to be able to separate out the effects of the different collection times from the biological effects of interest.

5.2 The Full Variable Selection Model

5.2.1 The Complete Model Specification

We let $g \in [1, G]$ be the set of genes, $i \in [1, I]$ be the observations and $j \in [1, J]$ index the design vectors. Let $y_{g,i}$ denote the observed value of the expression of gene g on observation i. Our model is defined as follows:

$$
\begin{array}{rcl@{\qquad}rcl}
y_{g,i} & \sim & N(x_i'\beta_g + w_i'\Gamma_g, \psi_{gi}) & \log(\psi_{g,i})/2 & \sim & N(x_i'\Delta_g, \rho) \\
\beta_{g,0} & \sim & N(\mu_g, \tau_0) & \Delta_{g,0} & \sim & N(m_g, \sigma_0) \\
\beta_{g,j} & \sim & (1-\pi_{g,j}) \cdot \delta_0 + \pi_{g,j} \cdot N(0, \tau_j) & \Delta_{g,j} & \sim & (1-q_{g,j}) \cdot \delta_0 + q_{g,j} \cdot N(0, \sigma_j) \\
\mu_g & \sim & N(\mu_0, \nu_0) & m_g & \sim & N(m_0, v_0) \\
\pi_{g,j} & \sim & (1-r_j)\delta_0 + r_j Be(\alpha_{\pi_{g,j}}, \gamma_{\pi_{g,j}}) & q_{g,j} & \sim & (1-s_j)\delta_0 + s_j Be(\alpha_j, \gamma_j) \\
r_j & \sim & Be(\alpha_{r_j}, \gamma_{r_j}) & s_j & \sim & Be(\alpha_{s_j}, \gamma_{s_j}) \\
1/\tau_j & \sim & Ga(\alpha_{\tau_j}, \gamma_{\tau_j}) & & &
\end{array}
$$

Notice that the parts of the model listed in the left and right columns are mathematically similar. Thus, we will confine our discussion to the left side of the above model description with the understanding that what is said applies to the right as well.

We have split the mean gene expression, $y_{g,i}$, into two parts. $x_i'\beta_g$ and $w_i'\Gamma_g$. Here, x_i are design vectors, and w_i are principal component vectors derived from the housekeeping genes (see below). However, this distinction is purely for clarity in the description. In our implementation, once the principal component vectors have been computed, there is no difference in the treatment of the coefficients β and Γ.

We use a Bayesian shrinkage prior on the coefficients, $\beta_{g,j}$, of the linear model (except the first), with the modification that each gene has an individual prior probability that $\beta = 0$ ($\pi_{g,j}$ rather than π_j). This modification is in place to avoid over-shrinking $P(\beta_{g,j} \neq 0)$ to the mean.

The prior for $\pi_{g,j}$ is designed to allow shifts for the genes where there are detected differences while at the same time retaining a relatively high probability that $\pi_{gi} = 0$ when there are no such differences. We choose $\alpha_{\pi_{g,i}}$ and $\gamma_{\pi_{gi}}$ to put a low weight on low probabilities within the beta part of the mixture distribution to avoid model uncertainty. (For example, it is difficult to distinguish between a point mass at zero and a Be(0.1,100).) Choosing $\alpha_{\pi_{g,i}} = 3$ and $\gamma_{\pi_{g,i}} = 1$ suffices.

The constant parameters, α_{r_j} and γ_{r_j} may be chosen to encode a prior mean and some prior idea about the level of shrinkage desired for $P(\beta_{gj} \neq 0)$. If we write $\alpha_{\rho_0} = cm$ and $\gamma_{\rho_0} = c(1-m)$, then m is the prior mean, and the parameter c controls the level of shrinkage. For low values, behavior will be very similar to older models which simply use a single π_j and give it a beta prior (Scott and Berger, 2005). An example of the results of this type of model are shown in Figure 5.2. High values will behave more like a model in which each π_{gj} receives its own individual beta prior (ie. $P(\beta_{gj} \neq 0)$ and $P(\beta_{g'j} \neq 0)$ are completely independent). With $c = 200$ on a data set of size 716, we get the results shown in Figure 5.3.

At first glance, our model for the $\psi_{g,i}$ adds significant complexity. However, notice that the two columns match exactly (except for ρ), and therefore much of the code is repeated. With this richer model for variance, we can accommodate the much more realistic situation in which constant variance across all the observations of a gene is largely true, but fails for some of the genes.

The design matrix consists of the mean expression vector (all ones), and a vector for each of the experimental groups. It is important to note that there is no explicit vector that separates the two different sets of controls, nor is there a vector that separates the first five experimental groups from the last four.

As discussed in the previous chapter, Affymetrix chips contain a number of control and housekeeping genes, and it is these that we rely on to detect differences in data collection. In this example, 162 of the approximately 50,000 genes on the chip are control genes. From these 162, we calculate the first eight principal components, and add these to our design matrix, H.

5.2.2 Correlation in π and q

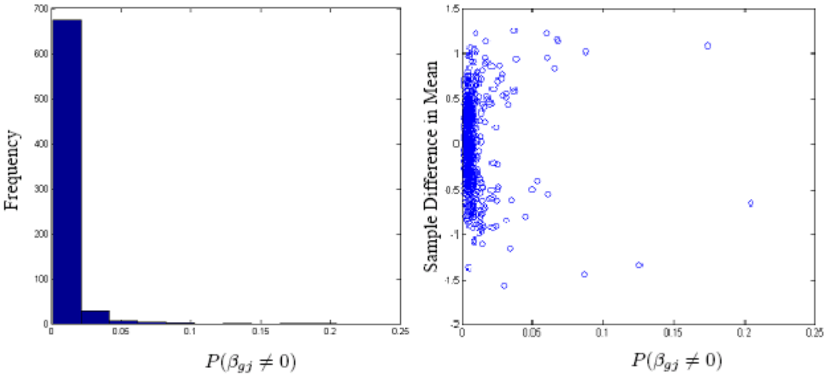


Figure 5.1: With either model, if there are no effects in the data, there are no effects detected. These are posteriors from analysing a simulated data set for which all observations were generated from a standard normal distribution. These are results from the standard variable selection model, but there is little difference from the posterior distributions of the full model of Section 5.2.1.

The unwanted correlation in the values of π and q can be seen best in Figures 5.1, 5.2 and 5.3. Since the number of genes affected by a particular experimental intervention can be a variable of interest, and this can vary significantly, we find this modification is critical to improving understanding of a given data set.

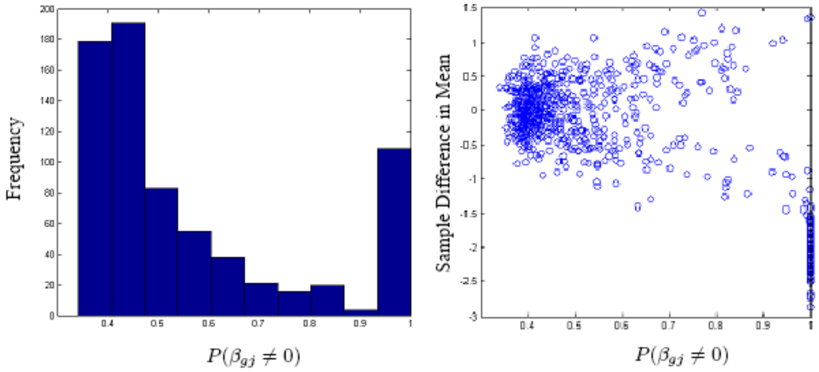


Figure 5.2: The data for the “control” group is generated from N(0,1). 616 genes in the “experimental” group were also generated from N(0,1) and the other 100 genes were generated from N(2,1). Notice that the bulk of the results are centered around $p(\beta \neq 0|-) = .4$

5.3 Some Examples and Details

5.3.1 Revisiting the Data

As described in Section 1.4, the analysis was run on RMA expression indices of $p = 10,715$ genes selected by screening all $n = 97$ samples to identify genes that varied in observed levels by at least a factor of 1 (fold) and whose median log2 expression across the samples exceeded 7 (just below the median off all genes across samples). The 97 samples were collected in three parts at three different times, and variations in lab conditions produced significant noise in the data which can be seen to correlate with the different collection times (see Figure 5.4). For this analysis of the oncogene data set, we used the first eight of the principal components of the housekeeping genes.

The specific parameter settings for prior hyperparameters required were chosen to define relatively uninformative priors consistent with the known ranges of vari-

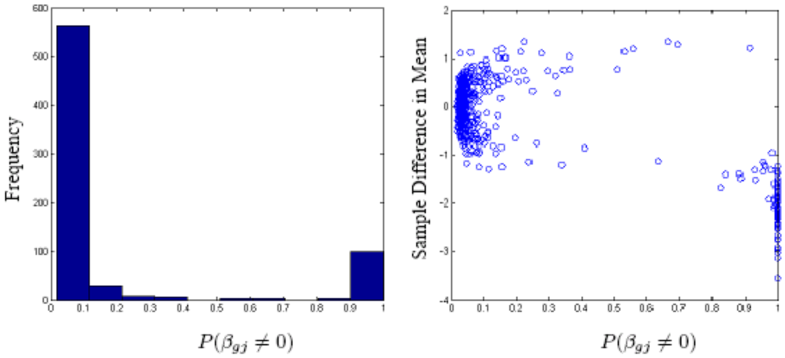


Figure 5.3: This is the same data analyzed in Figure 5.2. Notice that with the added level of hierarchy, the points that do not show strong evidence of difference between control and experimental groups have a genuinely low $p(\beta_{g,j} \neq 0|-)$

ation of gene expression (fold scale) and also know characteristics of Affymetrix data generating processes. Based on many experiments with observational and experimental data generated on the current Affymetrix platform and processed with RMA, typical ranges of residual/technical variation experienced indicate error standard deviations in the 0.1-0.5 range, so guiding our choice of $\rho = .01$. Additionally, we expect that changes in laboratory technique will produce changes across all genes while the knockout of a single specific gene will affect only those genes that are in the same pathways. Thus, in the selecting of hyperparameters α_{r_j} and γ_{r_j} we choose a relatively higher mean and allow for more shrinkage in the principle components coefficients than those associated with the experimental design.

The MCMC is initialized with values consistent with the prior and data and the simulation run for 4000 iterations to achieve nominal burn-in before saving and summarizing samples for a series of 20,000 iterations. Posterior means are then evaluated for exploration.

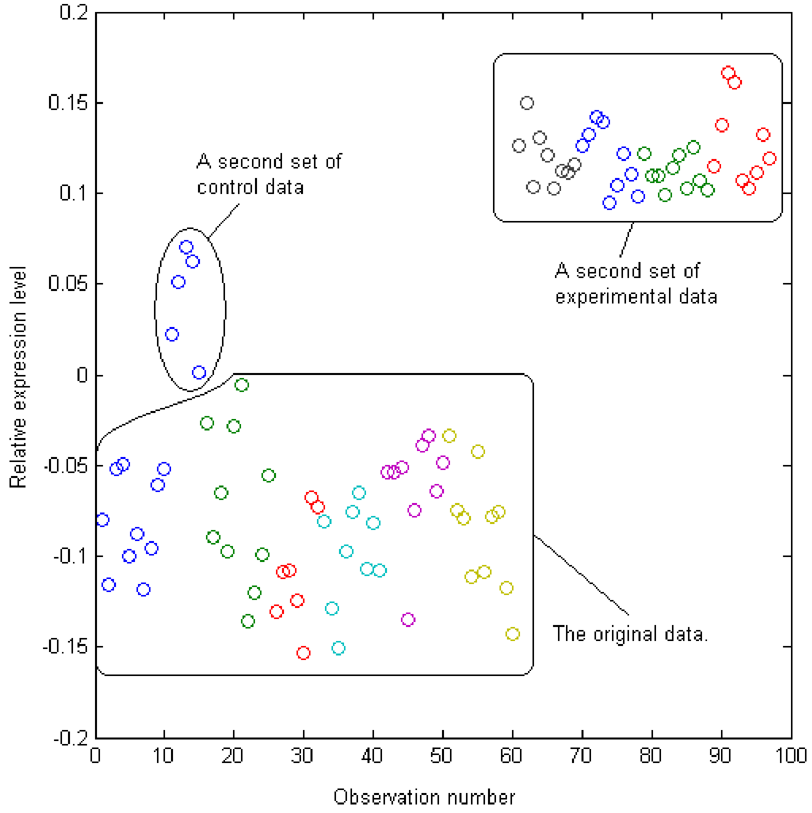


Figure 5.4: This is the first principle component of the set of 162 housekeeping genes. Notice that it nicely separates the data into three groups which correspond to the three different collection times.

The simplest aspect of our data to contrast is the difference in shrinkage between the experimental design and principal component coefficients (Figures 5.5 and 5.6). Notice that by our choice of hyperparameters, we are choosing to attribute as much of the variation in our data as possible to differences in data collection (principle components).

In our analysis, we find compelling evidence (see Figures 5.7a and b) that our use of principle components is significantly improving our ability to compare data taken at different times. Examples such as this are pervasive throughout the data

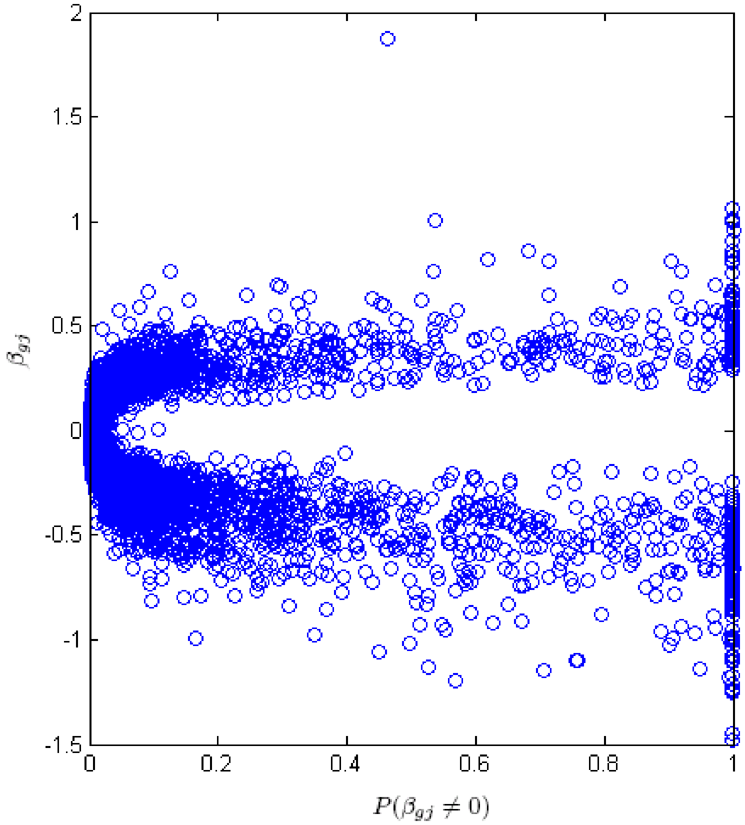


Figure 5.5: The posterior probability of $\beta_{g,j} \neq 0$ plotted with the posterior mean given it is not zero. This plot shows the coefficients associated with upregulation of the gene BCAT.

set.

In all previous models, the simplifying assumption that variance is constant across all experimental groups is made. Figure 5.8 is an example of why this is not a fair assumption. Finally, consider Figure 5.9. The first group consists of controls, and the other nine are experimental groups. Notice that experimental group six (P63A) shows a visible difference in both mean and variance. In older models, the difference in variance is undetectable, and this leads to the inability

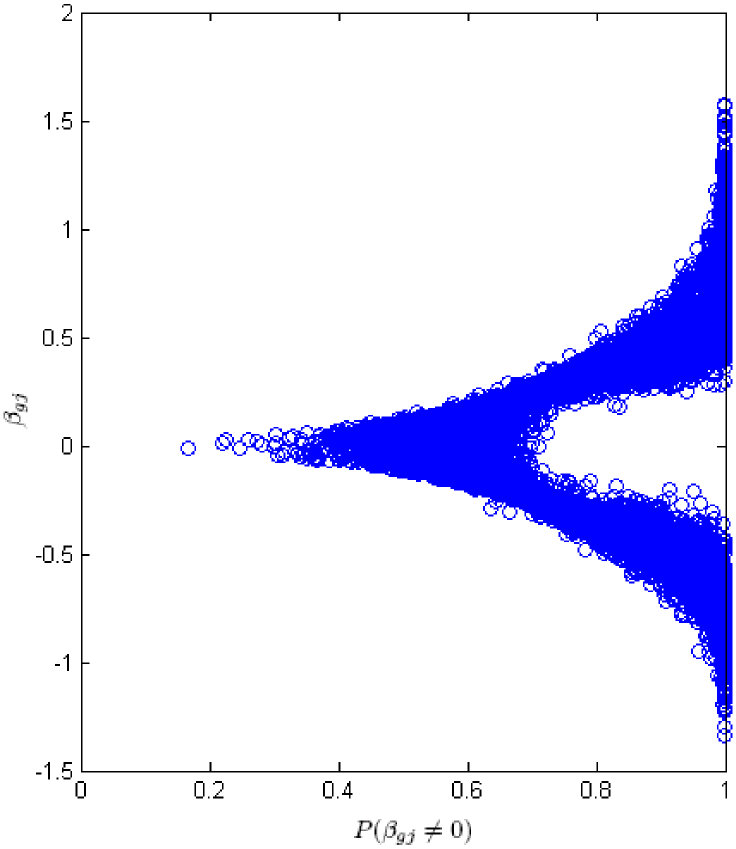
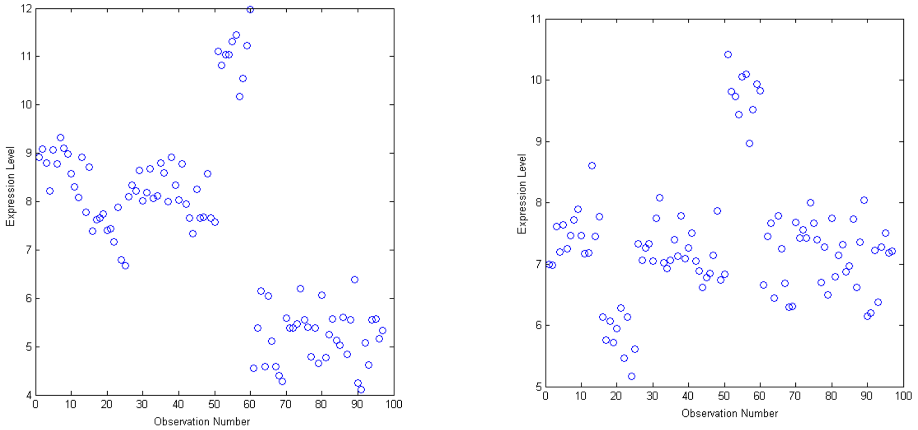


Figure 5.6: Because of our weaker prior on the hyperparameters of r associated with the principle components, we find that there is significant movement of $P(\beta_{pc6})$ toward 1.

to detect this gene in the signature for P63A.

5.4 Markov Chain Monte Carlo

The conditional densities used for the Gibbs steps in updating variables $\beta_{g,j}$, $\pi_{g,j}$, and r have been described in Sections 2.3.1, 2.3.2, and 2.3.3 respectively. Also, the Metropolis algorithm for updating $\lambda_{g,i} = \log(\psi_{g,i})/2$ has been described in



(a)

(b)

Figure 5.7: Figure (a) shows that without adjusting for differences in the data collection using the housekeeping genes, we may miss a weak signal from the Myc experimental group, and we will pick up significant type I errors from all of the data collected in the second group. Alternatively, (b) shows that after correction the Myc and Ras signals are strong, and the type I errors from the second set of data are eliminated.

Section 3.1.1. The only parts of the process that has not been described as yet are the updating of β_0 and Δ_0, and the updating of r.

5.4.1 The β_0 and Δ_0 Components

Note that, given all of the other variables, $\beta_{g,0}$ and $\beta_{g',0}$ are independent. Thus we will drop the subscripts for this subsection. Let $z_i = y_i - \sum_{j>0} x_{i,j}\beta_j$. Then

$$\begin{aligned} z_i &\sim N(\beta_0, \psi_i), \\ \beta_0 &\sim N(\mu, \tau), \\ \mu &\sim N(\mu_0, \nu_0), \\ 1/\tau &\sim Ga(\alpha_\tau, \gamma_\tau). \end{aligned}$$

This section of the model is straightforward, and so I will only state the conditional sampling distributions. To update these parameters in the MCMC, we do the

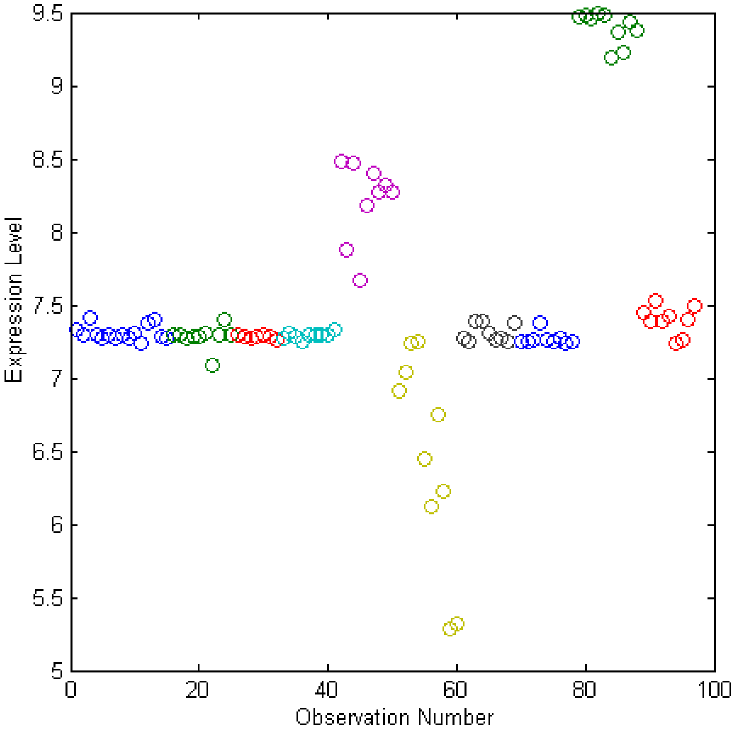


Figure 5.8: The assumption of a constant variance across all experimental groups is unreasonable.

following:

- Sample μ from a normal distribution with variance $v = \tau\nu_0/(\tau + \nu_0)$ and mean $v(\beta_0/\tau + \mu_0/\nu_0)$.
- Sample $1/\tau$ from a gamma distribution with shape parameter $\alpha_\tau + p/2$ (where p is the number of genes in the study) and rate parameter $\gamma_\tau + \sum_i z_i^2/2$.
- Sample β_0 from a normal distribution with variance $v = 1/(1/\tau + \sum_i 1/\psi_i)$ and mean $v(\mu/\tau + \sum_i z_i/\psi_i)$.

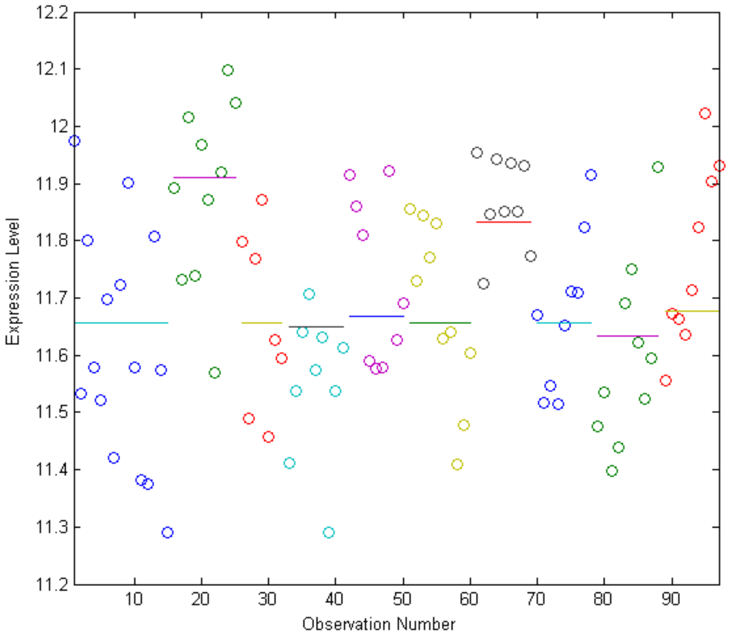


Figure 5.9: Notice that expression levels for experimental group 6 appear to be different from the control. The lines represent $\beta_0 + \beta_i \cdot \pi_i$ for gene number seven for our full model. However, because the variance differences are undetectable by older models, they are unable to find this gene in the signature for P63A.

5.5 Summary

We have demonstrated the efficacy of our model for the identification of metagenes associated with particular treatments. The model is shown to produce decreased false discovery rates compared to standard models. Additionally, where there are cases of variance changes in expression levels, Bayesian variable selection with our modifications allows for higher power in the choice of important genes.

Chapter 6

Case Study: Group Signatures and Breast Cancer

Whenever one uses microarrays to study the effects of an experimental intervention, one is essentially performing many thousands of experiments simultaneously. Because of this, it is difficult to justify singling out one gene as significant for any particular intervention in the absence of external corroborating evidence. However, the aggregation of all of the genes relevant to a particular intervention can be used to define a signature for that particular intervention. If the signature is one that is relevant for disease prognosis, such as the prediction of future heart attacks or the determination of level of malignancy in an excised tumor, then that signature has a potential use as a tool for recommending future treatment. In this way, it may be possible to avoid uncomfortable and costly treatments where there are unnecessary.

6.1 Data

In this chapter we will evaluate the use of metagenes for the prediction of breast tumor prognosis. We have three different data sets with which to work.

The first is an experimental data set. We have twelve U133+ 2.0 Affymetrix gene expression arrays that are broken into four groups of three. The same human mammary epithelial cell line was used in all of the experiments. The first group is a control group, the second set of cells were grown in the presence of lactic acid, the third was grown in low oxygen conditions, and the last was grown in both lactic acid and hypoxia. These experimental groups will be used to determine a gene signature for lactic acid and hypoxia.

Our second data set, which we will refer to as "Codex", is from a group of breast cancer patients. We have gene expression measurements, done using the U95 Affymetrix chip, from resected breast tumor cells from 436 patients with followup information on tumor recurrence. This data was collected in four groups and at three different locations. In order to partially correct for spatial and temporal biases, we will use principal components from the housekeeping genes as correction factors, as described in Chapter 4.

Our third data set comes from Miller *et al.* (2005). We will refer to this data set as "PNAS". It is also breast cancer though this study used the U133+ 2.0 chip, and it followed patient survival rather than tumor recurrence. Miller et al. use the data set to demonstrate the existence of a p53 signature in breast tumors, then show its relevance for the prediction of survival. In addition to survival data, we have a binary marker for the presence of a dangerous mutation in the p53 gene.

6.2 Modeling Gene Expression

Our first step will be the construction of lactic acidosis and hypoxia gene signatures. This will be the sole use for the experimental data set. The Codex and PNAS data sets will not be used in the construction of these meta-genes. In this way, the results obtained will be truly predictive results, with no issues of training on the same data that is used to do prediction. In this section we will discuss only the lactosis/hypoxia data set.

Many of the probes will be for genes that are not expressed, and most of the ones that are expressed will show no response to either hypoxia or lactic acidosis. Due to hybridization and measurement errors, a gene that is not expressed at all may still show some signal in the RMA data. The same errors make it difficult to distinguish between genes who's expression levels do not vary by more than one (fold). It is for these reasons that we eliminate from consideration any genes with median expression level lower than six and any gene for which the range of expression is less than one. This leaves 11,213 of the original 54,000 genes for analysis by variable selection.

One of the challenges of gene expression data is its sensitivity to the conditions under which measurements were taken. We utilize the housekeeping genes present on the U133 Affymetrix chips (which should show constant expression levels across both control and experimental groups) to build principal components. These are assumed to reflect the systematic noise, which occurs in all genes to varying degree, due to minor changes in assay conditions. We include enough principal components to account for 95% of the variation of the reference genes. In this case we found that the first four principal components, which we label $PC1$-$PC4$, were sufficient. We refer to this as principal components analysis for microarrays (HPCA) for the remainder of the paper.

Our analysis will follow the lines of Chapters 2 and 4. We will not make use of the variance model described in Chapter 3. Because we have only three observations per group, there will not generally be sufficient evidence of variance changes across groups.

6.3 Results

We find that we are generally successful in showing the biological relevance of the pathways detected in the lactosis/hypoxia experiment. Specifically, metagenes in the signatures relate to known hypoxia genes, and the scores of the tumors on these pathways are statistically relevant as parameters in a survival model of patients with breast cancer, as we will show below.

6.3.1 Principal Components

The number of genes showing a response to $PC1$ is high, but trails off to almost zero for $PC4$ (see Figures 6.1 and 6.2). This is consistent with the interpretation that $PC1$-$PC4$ are measuring decreasing levels of systematic error in the assay techniques. Empirically, we find that the results are significantly cleaner due to the inclusion of $PC1$-$PC4$ in the design matrix. Figures 6.3 and 6.4 show the RMA data before and after correction by PCAM for the gene, Androgen Receptor-Associated Protein 24 (ARA24). This is a member of class of androgen receptors, at least one of which (SRC1), has been shown to be a component of the hypoxia signaling pathway (Linja *et al.*, 2004; Carrero *et al.*, 2004).

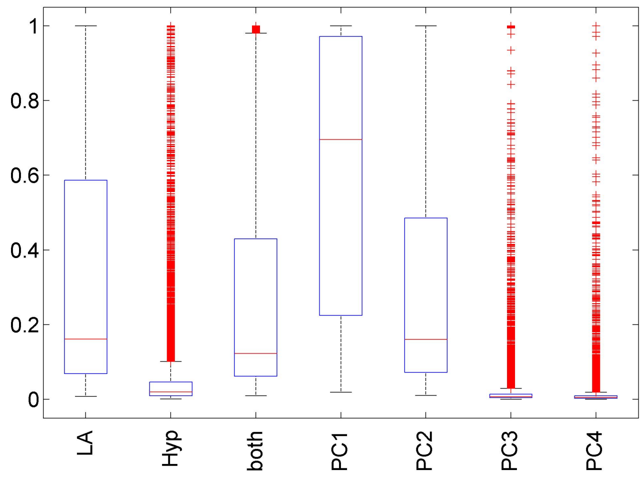


Figure 6.1: Boxplots of the posterior probability that a given factor is important in the measured expression level of the 11,213 genes. The x-axis lists the factors associated with the corresponding boxplot (hypoxia, lactic acidosis, both, and the four principal components from PCAM), and the y-axis shows the posterior inclusion probabilities.

6.3.2 Experimental Factors

We find that there are 1400 genes in the lactose response group, 152 genes in the hypoxia group, and 1100 genes that respond to the presence of both in a non-additive manner. Figure 6.5 shows the posterior probabilities of inclusion for each of the three experimental groups and all of the 11,213 genes.

We found 13 genes which show a response to hypoxia, lactic acidosis, and a synergistic response when both are present. Two of these, RRP40 and CAB56184, show a true synergistic response (Figure 6.6). Two of them, KIAA0279 and ALOX15B, show a paradoxical response in which the gene is up-regulated by either hypoxia or lactic acid, but expression returns to baseline when both are present (Figure 6.7). The remaining 9 can more accurately be categorized as

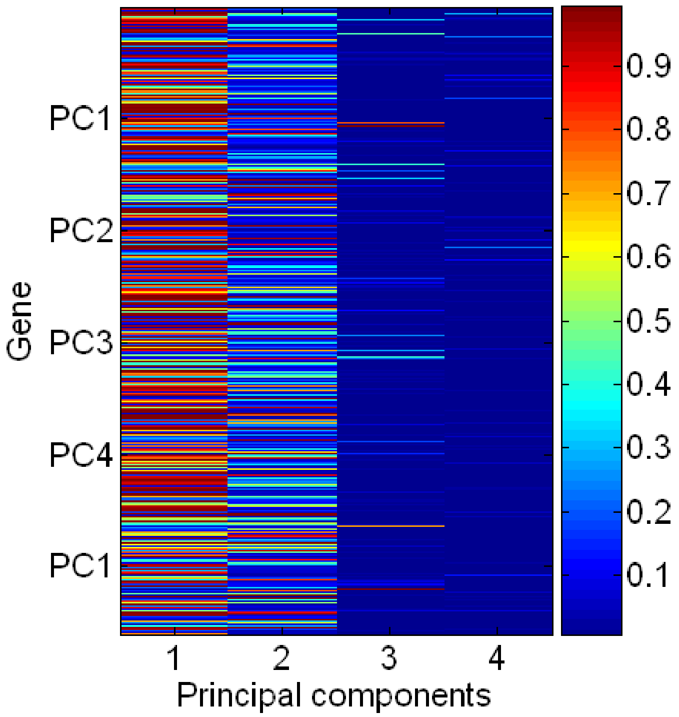


Figure 6.2: A heatmap of the posterior probability of inclusion for $PC1$-$PC4$ (x-axis) for each gene (y-axis). The low posterior probability of inclusion for $PC3$ and $PC4$ for almost all genes indicates that we have captured almost all systematic variation in the first two components.

on/off switches (Figure 6.8). In these cases, the presence of hypoxia, lactic acid, or both changes the expression level by approximately the same amount. Isolating these different behaviors, with linear models and a standard design matrix is impossible. There is some current research, built around constructing physical models for potential interactions, but this has not yet progressed to the point of fitting high dimensional data, as is seen in expression arrays.

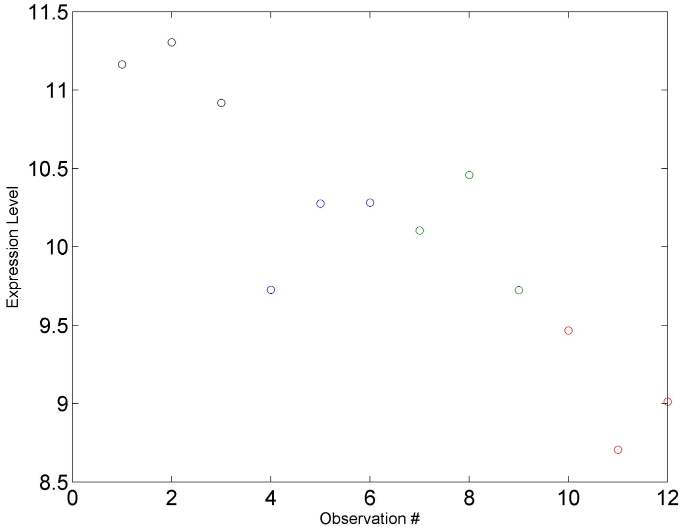


Figure 6.3: The measured expression levels before correction by PCAM. The groups are: black - control, hypoxia - blue, lactic acidosis - green, both - red.

6.4 Tumor Scoring

6.4.1 Codex Data Set

We now have gene signatures that occur in the presence of lactic acidosis, hypoxia, and under both conditions. One problem with relating these signatures to the breast tumor cells is the use of different Affymetrix chips. Affymetrix provides a list of matches from U133 to U95 that is many to one. We have arbitrarily chosen the first on the list for each U95 gene as its match.

After removing the genes from the lactic acidosis, hypoxia, and combined signatures that have no match in the U95 chip, we are left with 312, 22, and 217 genes respectively. Figures 6.9, 6.10, and 6.11 show the projections of the twelve experimental observations on the first two principal components of these three sets of genes.

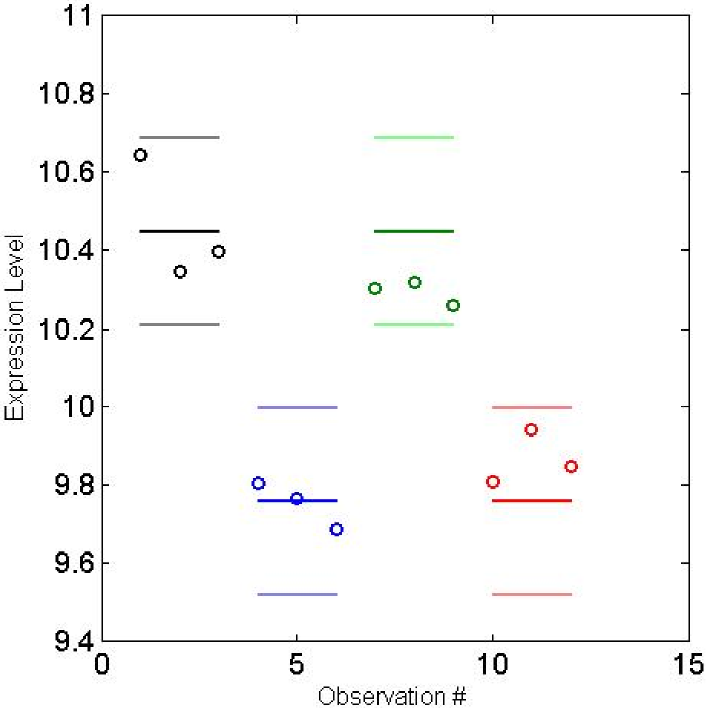


Figure 6.4: The measured expression levels after correction by PCAM. The groups are: black - control, hypoxia - blue, lactic acidosis - green, both - red. The lines represent the posterior mean plus or minus two standard deviations from fitting our hierarchical variable selection model.

Because it separates the groups appropriately, we will take the second principal component from the lactic acidosis signature group. Likewise, the hypoxia cell lines are separated from the other groups by the second principal component of the hypoxia signature.

While there is no clear predictive effect associated with the presence of a lactic acidosis signature on tumor recurrence (Figure 6.12), we find that the presence of a hypoxia signature is significant (cox survival model) in predicting a higher chance of the tumor recurring (Figure 6.13).

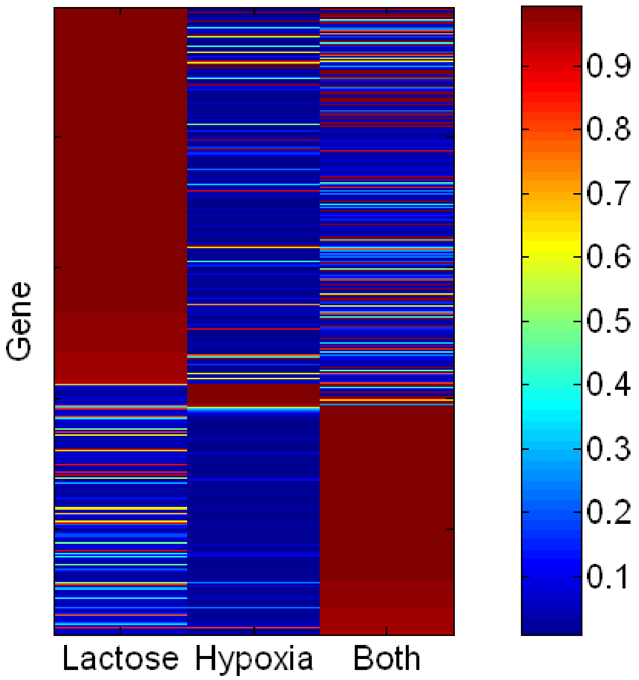


Figure 6.5: A heatmap of the posterior probability of inclusion for the three experimental groups (x-axis) for each gene (y-axis). They have been ordered to group genes with high probability of inclusion for each of the three groups together.

6.4.2 PNAS Data Set

Because the data we use to construct our meta-genes is taken on the same chip as is the tumor data, we are able to ignore the significant problem of mapping between two different chips. We project PNAS data onto the same hypoxia and lactic acidosis signatures, though now we are able to use all of the genes in the signature, rather than the subset used with Codex.

Figure 6.14 shows the first two principal components of the expression levels from the hypoxia genes of the experimental cells. There are differences in the placement of the 12 experimental cell lines on the graph due the the exclusion of

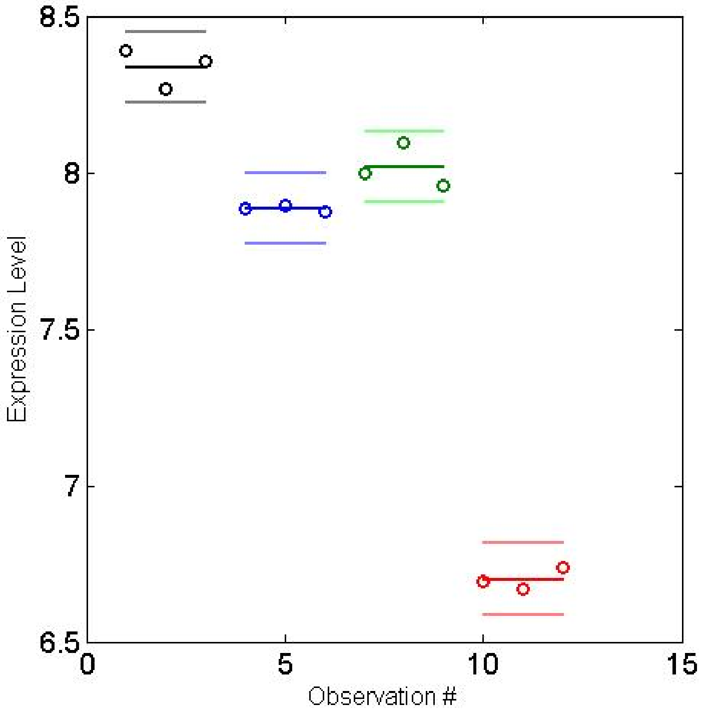


Figure 6.6: The expression levels of RRP40 with the fitted mean plus or minus 2 standard deviations. This gene is repressed significantly more when both hypoxia and lactic acid are present than would be expected if the effects were additive.

some of the genes from the meta-gene when comparing U95 and U133 chips.

Notice that there is a significant hypoxia signature in almost all of the tumors in the PNAS data set. Indeed, some of the tumor cells show a stronger hypoxia signature than even the cell lines. The experimental cell lines are homogeneous, and known to be grown in very low oxygen levels, while the tumors are necessarily heterogeneous. The level of hypoxia signature expression is extraordinary, as the tumor cell scores are averaged by the process of extraction of RNA from the tumor as a whole, and a significant part of the cells in the tumor are likely to be well perfused. Because the PNAS and experimental data sets were taken at different times and by different groups, it is difficult to determine whether the high hypoxia

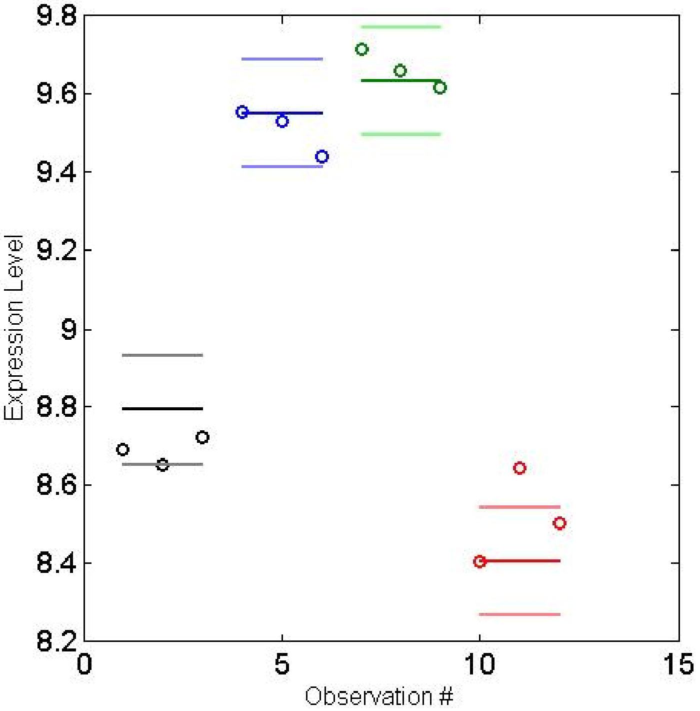


Figure 6.7: The expression levels of KIAA0279 with the fitted mean plus or minus 2 standard deviations. This gene shows a paradoxical "return to baseline" response when both hypoxia and lactic acid are present even though it is up-regulated when one or the other are present.

signature is due to difference in labs or actual elevated hypoxia signatures in tumor cells.

Splitting the patients into the top and bottom 50% of hypoxia signatures shows no obvious predictive value (see figure 6.15). However, if we separate the top 10% from the bottom 90%, we find a visible difference in the survival curves (see Figure 6.16).

Because we are able to keep all of the genes in the lactic acidosis meta-gene, we have a more sensitive measure of tumor response to lactic acidosis than we did with the codex data set. Interestingly, lactic acidosis signature and the p53

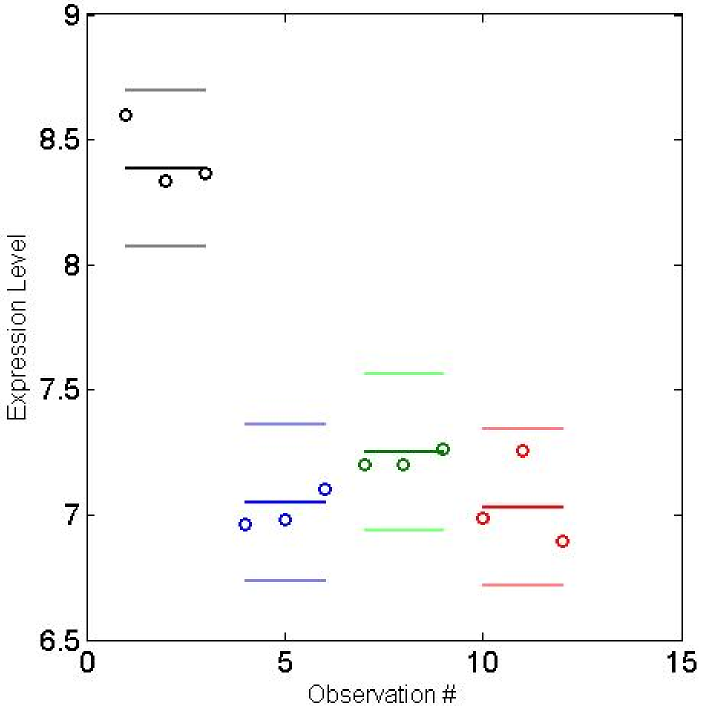


Figure 6.8: The expression levels of SLC7A11 with the fitted mean plus or minus 2 standard deviations. This gene behaves like an on/off switch in the presence of hypoxia, lactic acid or both. Isolating this behavior, with linear models and a standard design matrix, from paradoxical or synergistic responses is impossible.

wild type versus mutant type are visibly correlated (see Figure 6.17). It is already known that this particular p53 mutant is indicative of a poor prognosis, thus it is no surprise, given the correlation, that low lactosis score in a tumor is also indicative of poor prognosis (Figure 6.18). It has been shown previously that lactic acidosis is required for tumor cell death from hypoxia, and that p53 plays an important role in this process (Schmaltz *et al.*, 1998). Thus, our analysis adds support for this connection, and its relevance for disease prognosis.

As is to be expected, we find that p53 status in the PNAS data set is indicative of survival, however, our lactic acidosis signature is a much stronger predictor.

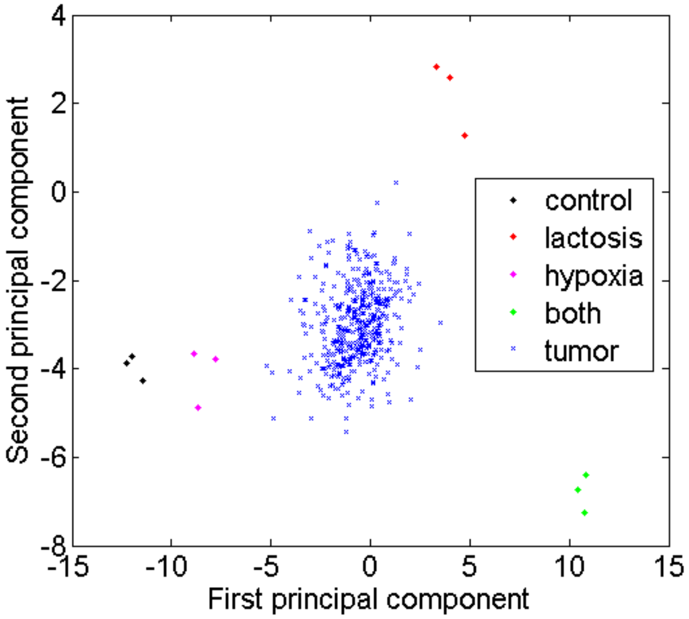


Figure 6.9: The scores of the twelve experimental observations on the first two principal components of the lactic acidosis signature genes. Control is black, lactic acidosis is red, hypoxia is magenta, and both is green. Notice that it is the second principal component that separates the lactic acidosis cells from the rest. The scores of the 436 cancer cell lines are overlayed using 'x'.

We find that when lactic acidosis score is taken into account, p53 status becomes insignificant while lactic acidosis signature remains significant.

It has become common knowledge within the biological community that the presence of lactic acid in tumor cells is indicative of poor cancer prognosis (Walenta and Mueller-Klieser, 2004). It is important to note here that, while the presence of lactic acid and the cellular response to that presence should obviously be correlated in normal cells, it is exactly the derailment of cellular responses that leads to tumor growth. With this in mind, it should not be overly surprising that one might see an abnormal response to the presence of lactic acid in a tumor, and that this abnormal response might be indicative of a poor prognosis. Indeed, one should

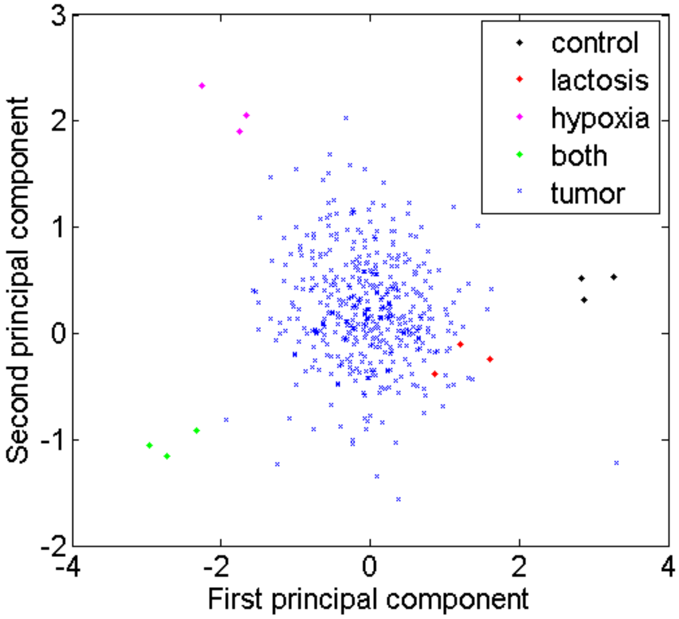


Figure 6.10: The scores of the twelve experimental observations on the first two principal components of the hypoxia signature genes. Control is black, lactic acidosis is red, hypoxia is magenta, and both is green. Notice that it is the second principal component that separates the hypoxia cells from the rest. The scores of the 436 cancer cell lines are overlayed using 'x'.

expect the presence of lactic acid to correlate more with the cause for its presence (machinery for hypoxia response) than with the response to it.

6.5 Summary

We have been able to use the work, outlined in previous chapters, to construct gene signatures for cells grown in hypoxic envrionments, cells grown in the presence of lactic acid, and cells under both conditions. Even acknowledging the probability that some genes have been incorrectly identified, we are able to measure these signatures in real tumor cells taken from patients. The level of expression of both hypoxia and lactic acidosis signatures was shown to be significant in disease

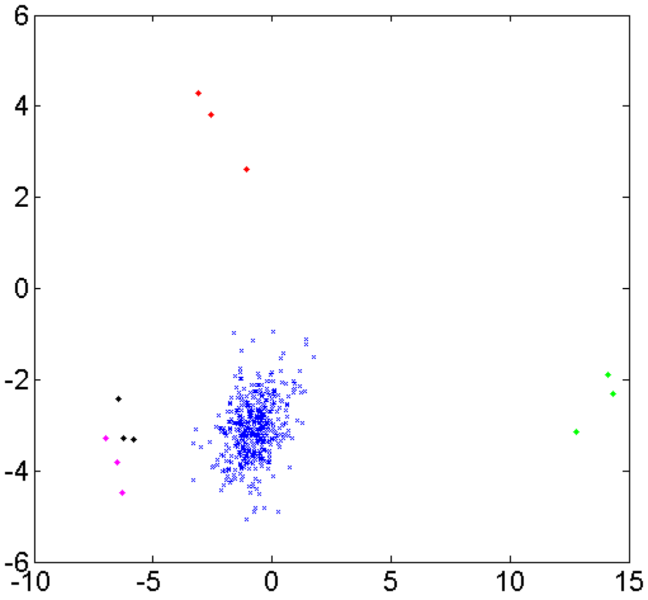


Figure 6.11: The scores of the twelve experimental observations on the first two principal components of the ”both present” signature genes. Control is black, lactic acidosis is red, hypoxia is magenta, and both is green. Notice that it is the first principal component that separates cells that are grown with both lactic acidosis and hypoxia from the rest. The scores of the 436 cancer cell lines are overlayed using ’x’. Notice that there is very little evidence of the presence of a ”both” signature in the tumor cells.

prognosis in the patient.

These gene signatures have the potential to identify which patients will respond to specific types of treatments. This will allow patients and their doctors to skip treatments that will do more harm than good in favor of those that have been shown to work on the patients tumor type.

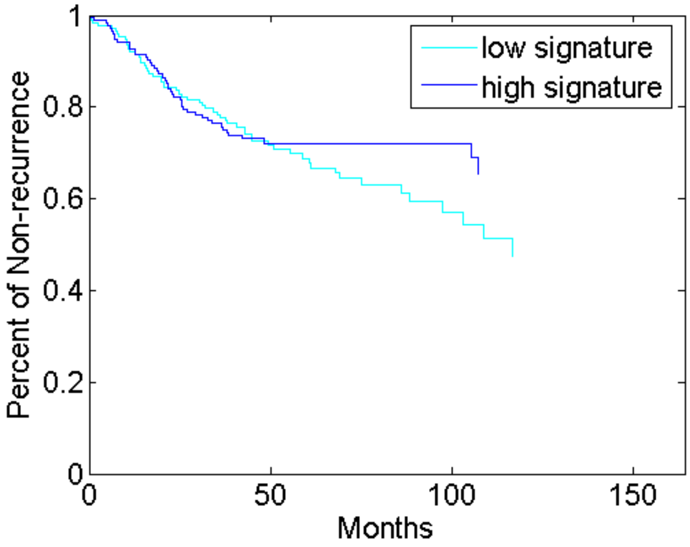


Figure 6.12: The survival curve for the patients in the study split according to whether they were in the top 50% in scoring on the lactosis signature (cyan) or the bottom 50% (blue). There is no clear affect of a lactosis signature in the cancer cell on tumor recurrence.

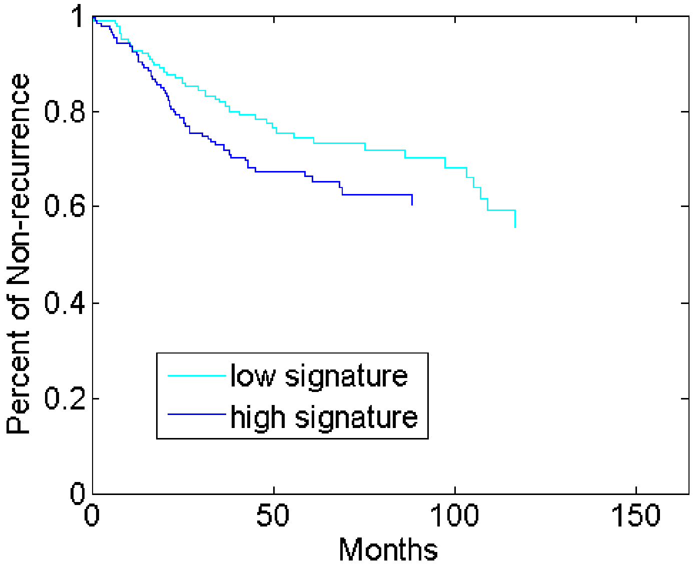


Figure 6.13: The survival curves for patients with a strong hypoxia signature (cyan) versus a weak hypoxia signature (blue). The survival curves are visibly separated, and the value of the hypoxia score is relevant with a p value of .043.

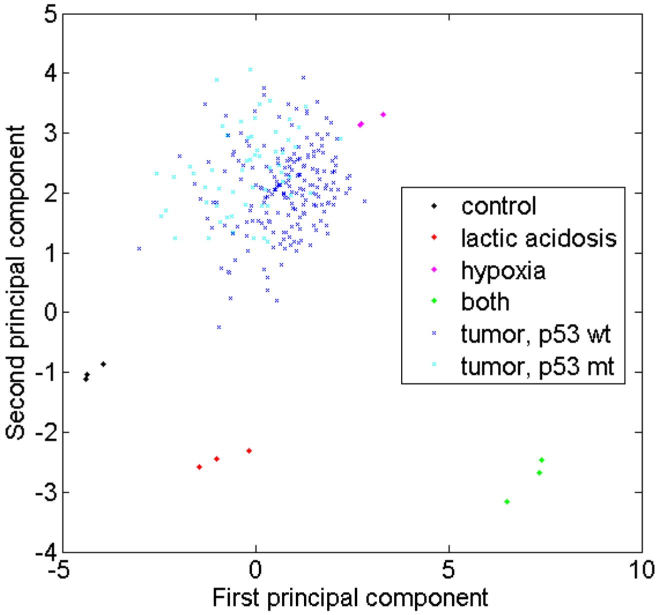


Figure 6.14: The hypoxia signature for the data sets that are used to define the meta-genes along with the values from projecting the PNAS data onto the corresponding principal components from expression of the genes in the meta-gene. Notice that the tumor cells almost universally exhibit a strong hypoxia signature, with some having a higher score than even the experimental cells that were grown in known hypoxic conditions.

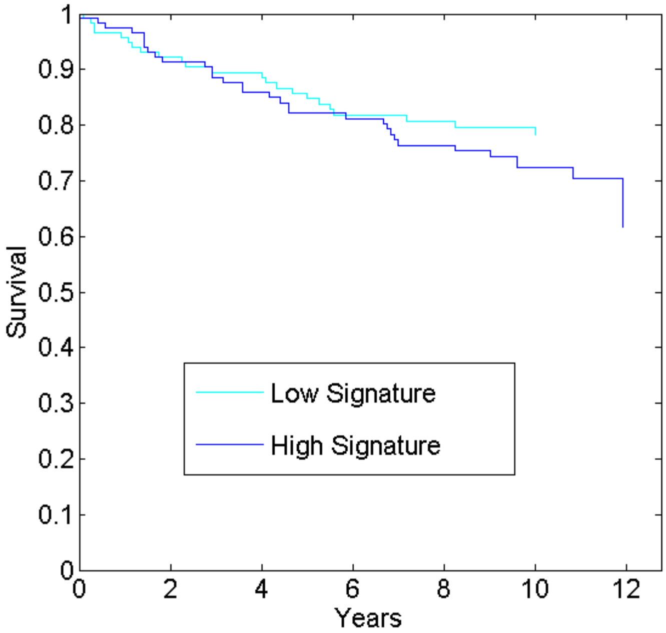


Figure 6.15: There is no obvious difference in survival when patients are split equally according to the level of hypoxia signature in their tumors.

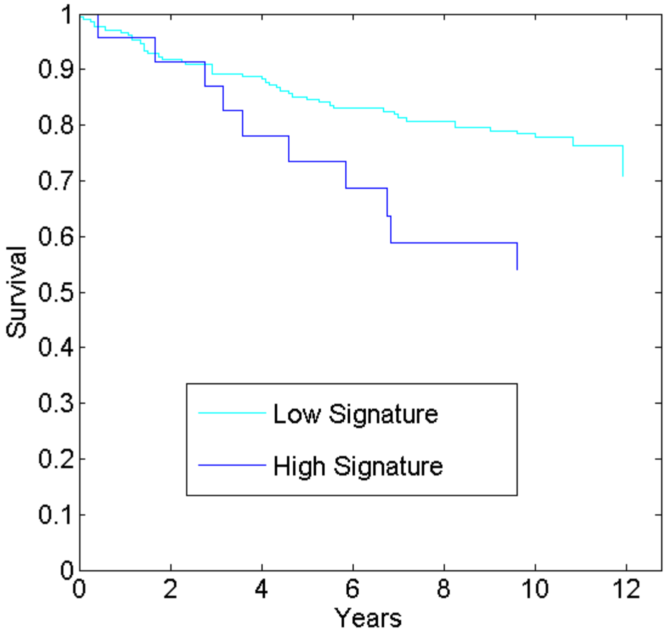


Figure 6.16: There is a visible difference in survival between patients with the top 10% of tumor hypoxia scores and those in the bottom 90%.

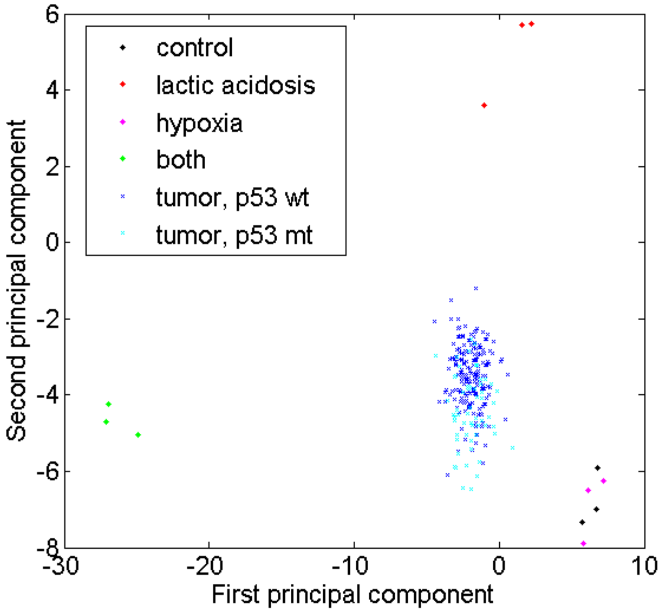


Figure 6.17: Notice the correlation between p53 mutant type vs wild type cells and lactic acidosis score.

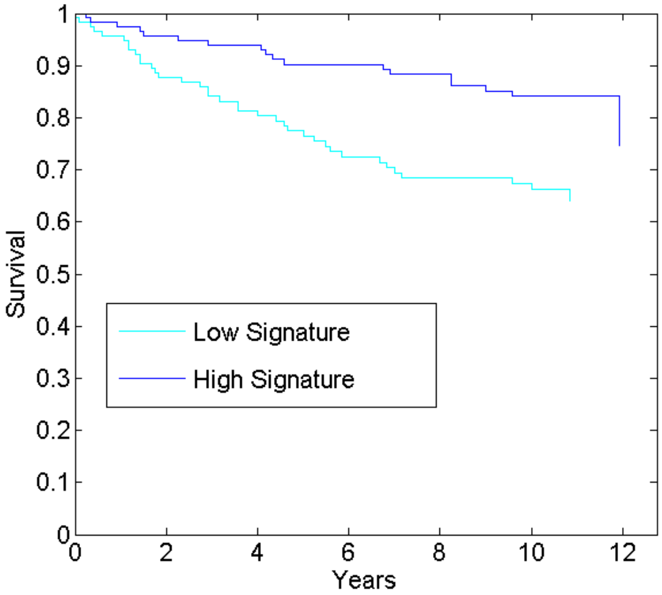


Figure 6.18: High expression of the lactic acidosis signature in tumors is indicative of a better survival outcome.

Chapter 7

Variable Selection in Gibbs Energy Models

Aside from the standard linear models used in earlier chapters, Bayesian variable selection has been specifically formulated for many types of generalized linear models, including log-linear models (Albert, 1995), logistic regression (Chen and Dey, 2003; Viallefont *et al.*, 2001), time series (Ibrahim *et al.*, 2000), and Poisson regression (Clyde and DeSimone-Sasinowska, 1997). Additionally, implementations have been described for time frequency surface estimation (Wolfe *et al.*, 2004) and certain non-parametric models (Kohn *et al.*, 1999). The feature that all of these models have in common is the ability to write down and sample from the conditional densities of the variables that have been assigned variable selection priors. It is the absence of this feature that has so far precluded the use of variable selection priors in complex Gibbs energy models.

In the case of most hierarchical models, if one can not draw from the conditional density, then it is usually easy to switch to Metropolis steps in the Markov chain Monte Carlo. Unfortunately, that is not straightforward in the case of point mass mixtures. Because the density function associated with a point mass distribution is infinite at the point mass (the function must integrate to 1 at that point), if we start a random walk at zero, it will never leave. Also, if the proposal distribution has no point mass at zero, and doesn't star there, it will never visit zero. For this reason, use of a random walk Metropolis-Hastings algorithm directly is impossible.

If we write the likelihood of data Y given parameters x as $L(Y|x)$, the prior distribution as $\pi(x)$, and the proposal distribution for x' starting from x as $Q(x'|x)$, then the probability of accepting a move from 0 to x is:

$$\frac{L(Y|x)}{L(Y|0)}\frac{\pi(x)}{\pi(0)}\frac{Q(0|x)}{Q(x|0)}$$

For a random walk, the ratio of proposal distributions is 1. We know, $\pi(0) = \infty$, so we are left with zero probability of accepting the move.

In this chapter we will focus on approximations of variable selection priors that will allow them to be used outside of the framework of generalized linear models.

7.1 Proposal Distributions

The most direct solution to this problem is to use the prior distribution as a proposal distribution. In this case, the prior and proposal distributions cancel each other, leaving only the likelihood in the calculation of acceptance probability. In cases where priors are well informed, this may be an acceptable alternative, however it is often the case that the use of the prior in this context will lead to very low acceptance rates. For example, if we take 50 observations from a standard normal distribution and use a mean zero normal prior with standard deviation of four for the mean, we will have an acceptance rate of just over 4%. Because acceptance rates become low very quickly, the use of any form of diffuse prior in this situation is precluded.

What we need to fix this acceptance rate problem is some ability of the proposal distribution to adapt to information gained about the posterior as the MCMC progresses. Suppose that, instead of using a prior that is a mix of zero with a continuous distribution, we approximate the continuous distribution with a finite set of points, $\{x_i\}_{i=1}^N$. In this situation, we may clearly choose a proposal distribution with support on our discrete set and with non-zero probability on every point. If we write $\pi(x_i) = \pi_i$ and $Q(x_j|x_i) = q_{i,j}$, then our acceptance function for a move from x_i to x_j now looks like

$$\frac{L(Y|x_j)}{L(Y|x_i)} \frac{\pi_j}{\pi_i} \frac{q_{j,i}}{q_{i,j}}$$

By use of discretization, we may now use any sampling method we choose, including a random walk through the points. However, because we are approximating a point mass mixed with a continuous distribution, we should assume that the prior probability attached to zero is much higher than that attached to the other points in the discrete set. Indeed, if we begin to consider larger and larger discrete sets (as a better and better approximation of a continuous distribution), then the ratio of those probabilities will limit to infinity.

With this in mind, we need to ensure that, if our proposal distribution is a random walk, then it has a reasonably high chance to move both to and away from zero.

7.2 A Random Walk Analogue

Let $g(x|\mu)$ be a family of distribution functions (on the real numbers) with location parameter μ. (For example, to produce a random walk, $g(x|\mu)$ might be normal.) Define $g_i(x_j)$, a probability density function on $\{x_j\}_{j=1}^N$, as $g(x_j|x_i)/\sum_j g(x_j|x_i)$.

We define our proposal distribution, $Q(x_j|x_i)$, to be $g_i(x_j)$ with probability q and 0 with probability $1-q$. With this proposal distribution, we will return to zero at random intervals in the MCMC chain. In order to simplify calculation, we might assume that the points $\{x_i\}_{i=1}^N$ extend far out into the tails of the posterior distribution, and that $g_i(\cdot)$ is symmetric around x_i. Under these conditions, $g_i(x_j)$ and $g_j(x_i)$ will be approximately equal (with any difference originating from mass in the tail of both the proposal and the posterior distributions). This means that the acceptance probability in this situation will be

$$\frac{L(Y|x_j)}{L(Y|x_i)}\frac{\pi_j}{\pi_i}.$$

Without these conditions, we will need to calculate the integration constant, $\sum_j g(x_j|x_i)$ for each point x_i (or at each step of the MCMC) and include the ratio $g_j(x_i)/g_i(x_j)$ in our calculation, as this will affect the transition probabilities.

For moves between non-zero points, we are using a standard random walk. Therefore we need only calculate the acceptance probability for the case of $x_i \neq 0$ and $x_j = 0$ (the opposite direction will have an inverse acceptance probability). For simplicity, let us assume that the point 0 is treated separately from the remaining point masses $\{x_i\}_{i=1}^N$. That is to say, $g_i(0) = 0$ and moves to zero only occur through a separate point mass at zero. Finally, define $1-r$ to be the prior probability that $x = 0$. Acceptance of a transition to zero is then

$$\frac{L(Y|0)}{L(Y|x_i)} \frac{\pi(0)}{\pi_i} \frac{Q(x_i|0)}{Q(0|x_i)} \quad = \quad \frac{L(Y|0)}{L(Y|x_i)} \frac{(1-r)}{r\pi(x_i)} \frac{qg_0(x_i)}{(1-q)}$$

Suppose that $g(x|\mu)$ is normal with a variance that has been calibrated during burnin to achieve an acceptance rate of between 15% and 50% (standard practice in random walk MCMC). If our posterior distribution is unimodal with mode at zero, or if it is bimodal, but the modes are close (zero will almost always be a mode), then this technique should converge to our posterior distribution. By allowing our set of discrete points, $\{x_i\}_{i=1}^N$, to tend to infinite size (or at least to the number of numbers representable by a computer), we recover our original mixture prior.

Unfortunately, it is precisely this desired unimodality that is likely to be destroyed by our prior. If there is indeed evidence that x is non-zero, then the posterior will have a mode that is separated from zero by an area of low posterior probability. This will lead to poor or no convergence.

7.3 Random Walks and Multi-modality

Now we will return to treating our variable on continuous space. Suppose that we are able, by judicious use of the information obtained about our posterior during burnin, to design $g(\cdot)$ so that it is approximately proportional to our posterior conditional on $x \neq 0$, that is $p(x|Y, x \neq 0)\pi(x|x \neq 0)$ (perhaps with some overdispersion in order to avoid under-sampling the tails). If this is possible, then we might use a proposal distribution that mixes a point mass at 0 with $g(\cdot)$. This would lead to reasonable acceptance rates with periodic jumps to and from 0.

We propose to learn, in a second burnin phase, about our distribution through the following non-Markovian chain:

- Suppose we are at step t in the chain. Let $n \leq t$ be the largest value such that $x_n \neq 0$.

- If $x_t \neq 0$ then

  choose $y_t \sim N(x_t, \sigma)$ with probability q. Set $x_{t+1} = y_t$ with the acceptance probability

  $$\frac{L(y)}{L(x_t)}\frac{\pi(y)}{\pi(x_t)}$$

  choose $y_t = 0$ with probability $1 - q$. Let $x_{t+1} = y_t$ with acceptance probability

  $$p_{t,n} = \frac{L(0)}{L(x_t)}\frac{1-r}{r\pi(x_t)}\frac{qN(x_t|x_n, \sigma)}{1-q}$$

- If $x_t = 0$ then set $x_{t+1} = 0$ with probability q. Otherwise, let $y_t = N(x_n, \sigma)$ and accept the move $x_{t+1} = y_t$ with probability $1/p_{t,n}$.

Note that this is not Markovian because it relies on knowledge of values from the chain (x_n) that are potentially arbitrarily far back in history. However, as we will show, it can perform well for the purposes of learning about the distribution of interest.

If we allow this algorithm to run for a burn in period, then begin collecting mean and variance data, with which we can construct a normal approximation to $P(x|Y, x \neq 0)\pi(x|x \neq 0)$. Our true Markov chain then proceeds with a mix of this Gaussian and zero.

7.4 A Toy Example

We will use a set of data points that are known to be drawn from a normal distribution with variance one, and we want to decide whether the mean of the distribution is zero. Under these conditions, we may use a point mass prior for the mean of the distribution

$$\mu \quad \sim \quad (1-q)\delta_0 + qN(0,\sigma)$$

From section 2.3.1 we can calculate the posterior distribution exactly. Thus, this example provides a tool for testing our sampling mechanism.

We will use a sample of 25 observations with mean and variance artificially set to .75 and 1 respectively. Additionally, we will set the prior standard deviation to $\sigma = 100$ and the prior probability that $\mu = 0$ to 0.5. Under these conditions, and using the formulas in section 2.3.1 we obtain an approximate posterior distribution for μ of

$$\mu \sim (.3065)\delta_0 + .6935N(.75,.04)$$

From the burnin, using a normal random walk with standard deviation of .7, we obtain a distribution for $\mu|(\mu \neq 0)$ with a mean of .7504, a variance of .0395, and $P(\mu = 0) = .3250$. When we then shift to sampling from $N(.7504,1)$ (overdispersed to ensure that the sampler will sample from the tails) with probability .5 and zero otherwise we obtain the slightly more accurate estimate of $P(\mu = 0) = .3057$. Note that we have chosen a mean of .75 for this example to induce a posterior probability that is relatively far from both 0 and 1. For means that are higher that .75, the posterior probability quickly converges to 1 and for lower means it quickly converges to 0.

Note that there are significant questions opened up by the use of this scheme. First, it is unclear whether the secondary burnin algorithm will converge to any distribution. Second, we have demonstrated good behavior in the case of estimation of the mean of a collection of observations. This is a one dimensional variable selection problem, and the behavior of this algorithm may not be as simple in a larger, more complex model. This algorithm is presented as an ad hoc method that produces good behavior in our situation, but opens up a potentially fruitful area of research.

One important issue that may need to be addressed is the problem introduced by posterior correlation. If two (or more) variables which both have variable selection priors are highly correlated, then there are likely two modes, one at $(0, 0)$ and another at (x, y). In this circumstance, it will be very difficult to move between the two modes with proposals in one dimension. If such behavior is extreme, then convergence may be very slow. It is fair to note that this is a problem with Gibbs sampling in one dimension in the case of linear models as well.

Now let us focus on a physical model, for which we can not calculate exactly the conditional distribution of the model parameters.

7.5 A One Dimensional Multi-scale Model

A graphical model is a set of random variables together with the dependencies between those variables. The name is derived from the fact that one can represent the model pictorially by assigning the variables to the nodes of the graph and drawing edges to show dependency. The statistical concept of dependency is often one way, and so the edges of a statistical graphical model are often (but not required to be) arrows.

Consider the model depicted in 7.1, and suppose that $x_i \in X$ and $y_i \in Y$, and let α index all of the possible n-vectors of states, $\bar{x}_\alpha$. Finally, assume that we have an "energy" function defined by, $U(\bar{x}) = \sum_{i=1}^{n} U_0(x_i|x_{i-1}, x_{i+1})$ for some function $U_0 : X \to \mathbb{R}$.

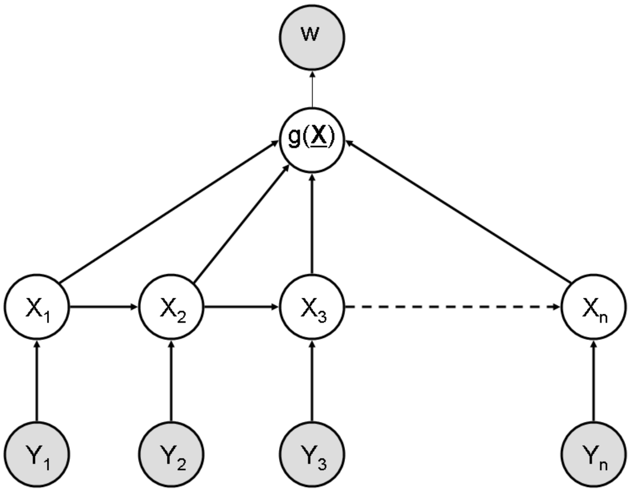


Figure 7.1: A hidden Markov model in which there is imperfect information, z, about a summary statistic, $f(x)$. Shaded nodes are observed, and clear nodes are unobserved.

We can then define a density on the space of possible configurations such that

$$f(\bar{x}_\alpha) = \frac{1}{Z} e^{-U(\bar{x}_\alpha)/T}$$

where Z is the integration constant:

$$Z = \int_\alpha e^{-U(\bar{x}_\alpha)/T} d\alpha$$

The constant Z, known as the "partition function" or "evidence", is generally intractable to computation. (Except in the special cases of the Kalman filter or when there are a finite number of states.)

Additionally, because the conditional independence is destroyed by the additional variable w (see figure 7.1), the backward-forward algorithms from traditional hidden Markov models can not be used. We will, however, be able to calculate the partition function directly in the case where X is a finite space.

Let $X = \{a_1, \cdots, a_m\}$ be the finite set of states in X. Also, define $\bar{x}_k$ to be an n-vector of states in the space X (corresponding to the n elements in the chain), and let $x_{i,k}$ be the i^{th} element of the vector $\bar{x}_k$. This leaves us with a total of m^n states in the system. We have

$$\begin{aligned} Z &= \sum_k e^{-U(\bar{x}_k)} && (7.1) \\ &= \sum_k \prod_i e^{-U_0(x_{i,k}|x_{i-1,k},x_{i+1,k})} && (7.2) \end{aligned}$$

Let M_i be the matrix with elements $u_{b,c} = exp(-U(x_i = a_b|x_{i-1} = a_c))$. Additionally, let $v = (e^{-U_0(x_1=a_1)}, \cdots, e^{-U_0(x_1=a_m)})$ be proportional to the density on the possible states of x_1 and $w = (e^{-U_0(x_n=a_1)}, \cdots, e^{-U_0(x_n=a_m)})'$ be the same for x_n. Then

$$Z = v \cdot \prod_{i=1}^{n} M_i \cdot w.$$

Keeping in mind the fact that one need only multiply an m-dimensional vector with an $m \times m$ matrix a total of n times, the calculation of Z takes place in order $m^2 n$ time. In cases where some transitions are impossible, there will be zeros in the matricies, M_i, and the order may be less.

Once it is possible to calculate the evidence, inference becomes possible. Suppose, for example, that the function, $g(x)$ in Figure 7.1 is the average probability of finding any of the x_i are in state a_1. Let u_i be the column vector $exp(-U_0(a_1|x_{i+1,k}, x_{i-1,k})$ where k runs from 1 to n. Then define $M_i^* = (u_1, \bar{0}, \cdots, \bar{0})$. The quantity $Q_i = vM_1 \cdots M_{i-1}M_i^*M_{i+1} \cdots M_n w/Z$ is the probability that variable x_i is in state a_1. Thus, $1/n \sum_i Q_i$ is the quantity of interest. Notice that the calculation of Q_i is almost identical to the calculation of Z, and therefore requires very little additional work.

7.6 Summary

We have introduced a novel technique for fitting a point mass prior when there is no possibility of calculating the conditional probability of the variable in question. In the following chapter we will explore further use of this model in the prediction of α-helicity in short polypeptides.

Chapter 8

Calibration and Variable Selection in a Physical Model for α-Helicity

The development of new pharmaceutical therapies has become the newest booming economy, but at its heart, it is still a process of "guess and check". Almost all drugs work on the binding sites of proteins, but the folded structure of the proteins themselves is difficult to determine. Because of this, it is difficult to design a drug with a specific protein in mind. Additionally, the drugs to be designed may be proteins themselves. Thus, even if the geometry of a target binding site is known, directly designing a protein to fit the target is beyond current modeling technology.

One of the main secondary folded structures one finds in proteins is the α-helix. As suggested by the name, the structure is formed when the backbone of a protein forms a helix. The structure is stabilized by hydrogen bonds between certain carboxy and amino groups in the backbone.

Because the α-helix structure is linear, and because the potential for long range interactions (interactions between amino acids that are separated by more than a few residues) is relatively low, this structure lends itself to modeling by the model outlined in chapter 7.

In this chapter, we will describe the modeling of α-helices with the one dimensional multi-scale model described previously. Additionally, we will demonstrate inference using that model and show that the posterior fit is better than the current state of the art in modeling helices.

8.1 Data

We have collected 1187 different polypeptide helicity measurements which are broken into 360 distinct polypeptide chains, some of which have been measured at different temperatures and pH's. All have been collected from the biological literature. To date, biochemical researchers have focused on singling out specific interactions which has led to a data set that is by no means a random sample from the space of polypeptides. Discovering what biases are inherited due to this process will be part of future work.

8.2 A Physical Model for α-Helix Formation

Consider the polypeptide backbone shown in 8.1. We will define an amino acid to be in a helical conformation when the ϕ and ψ dihedral angles are within a particular range (figure 8.2). Thus our model for helix formation at its simplest has hidden states, $x \in \{h, c\}$ (helix and coil states respectively), with $y \in \{G, A, V, L, I, M, F, W, P, S, T, C, Y, N, Q, D, E, K, R, H\}$ (these are the 1-letter abbreviations for the 20 amino acids).

With this model we can take into account the fact that each amino acid has a different energy level associated with being in α-helix conformation. However, this does not allow for the inclusion of any long distance interactions (as one would

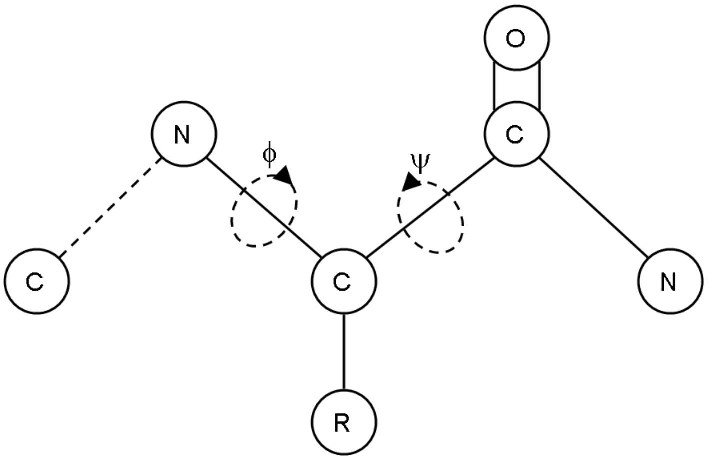


Figure 8.1: Definition of the ϕ and ψ dihedral angles in a polypeptide backbone.

expect from a first order Markov model).

In truth, one would expect to have to input energy to constrain any amino acid to its α-helix conformation. Unfortunately, if we use this prior information, we will find that, in fact, α-helix conformation will be unfavorable in all situations. The stabilizing interaction in an α-helix is a hydrogen bond that occurs between an amino acid and the amino acid that is three steps away. This interaction is available when there are three amino acids in a row that are in helical conformations. We may keep our first order model and incorporate this interaction energy by expanding our state space. Now we assume that our state space consists of pairs of helices/coils. That is $x \in \{hh, hc, ch, cc\}$. It becomes immediately clear that our transition matrix will contain a number of zeros, as transitioning from state $h\bar{h}$ to state $\bar{c}h$ requires that the amino acid in position 2 in the first state and in position 1 in the second state be simultaneously in both helix and coil formations (here the bar signifies the positions that are represented in both states).

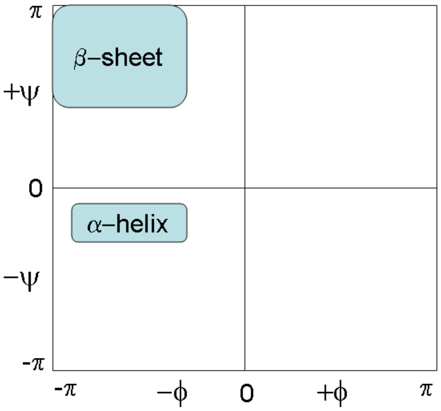


Figure 8.2: An example of a Ramachandran plot. The colored areas represent rotational angles corresponding to α-helix and β-sheet conformations. This is a stylized example. In truth, these plots vary with the amino acid.

The full transition matrix is shown in Figure 8.3.

Note that our work is adapted from the work of Chakrabartty *et al.* (1994).

$$\begin{array}{c} \\ h\bar{h} \\ h\bar{c} \\ c\bar{h} \\ c\bar{c} \end{array} \begin{array}{c} \begin{array}{cccc} \bar{h}h & \bar{h}c & \bar{c}h & \bar{c}c \end{array} \\ \begin{bmatrix} w_j & v_j & 0 & 0 \\ 0 & 0 & 1 & 1 \\ v_j & v_j & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

Figure 8.3: The transition matrix for the hidden Markov model in which each state contains helix/coil configuration for two amino acids. The overbars are meant to mark an amino acid position that is recorded twice. Thus if three amino acids are in positions 'hch', then that would be labeled '$h\bar{c}$' together with '$\bar{c}h$'.

As described in section 7.5, we are using a Gibbs energy formulation for our model. The coil position is defined to have energy 0 (which, when exponentiated gives us the 1's in the transition matrix). Consider the meaning of w_j and v_j. Notice that transition weight, w_j, corresponds to three consecutive amino acids in helical configuration. Using a standard Gibbs energy (Munoz and Serrano, 1995a),

$$\Delta G_j = \Delta H - T\Delta S_j \tag{8.1}$$

where j ranges over the distinct amino acids, we define $w_j = e^{-\Delta G_j/rT}$. The contribution to stability due to hydrogen bonding is measured by the strength ΔH (which is constant across all amino acids). An amino acid in a helical conformation with at least one coil neighbor is assigned $v_j = e^{-\Delta S/R}$, corresponding to the entropy loss of helix formation without the enthalpy of H-bond formation between helical residues.

8.2.1 Model For Temperature

It has been shown (Munoz and Serrano, 1995b) previously that the temperature dependence in 8.1 is inadequate. We are modeling a phase transition between random coil and α-helix. In such a transition there will be changes in solvent accessibility due to the exposure of groups of atoms in one state that are unavailable to the solution in the other state. This involves a change in heat capacity which we include in our model as a temperature dependence in both the energy and enthalpy terms in 8.1.

$$\begin{aligned}\Delta H &= \Delta H_{tref} + \Delta C_p(t - t_{ref}) \\ \Delta S_r &= \Delta S_{r,tref} + \Delta C_p \log\left(\frac{t}{t_{ref}}\right)\end{aligned}$$

Notice that any change in t_{ref} can be incorporated in the constants C_p, ΔS_r, and ΔH_{tref}. In order to stay consistent with previous work, we set $t_{ref} = 273$ Kelvin. These equations result from thermodynamic considerations, and are proposed for α-helix formation in Munoz and Serrano (1995b).

8.2.2 Interaction Parameters

In an α-helix, amino acids that are separated in the polypeptide chain are potentially brought into close proximity by the geometry of the helix. Specifically, once the chain is in helical formation, an amino acid at position i is nestled between those at positions $i+3$ and $i+4$. This leads to long range interactions that, depending on the amino acids involved, may have significant effects on the stability of the α-helix. We model these by inclusion of new interaction parameters, $\Delta\Delta G_{a,b,3}$ and $\Delta\Delta G_{a,b,4}$, representing sidechain interactions between amino acids a and b in positions i to $i+3$ and i to $i+4$ respectively. Thus the our energy function becomes:

$$\sum_i \Delta G_i + \Delta\Delta G_{i,i+3,3} + \Delta\Delta G_{i,i+4,4}$$

In order to accommodate these interactions, we must have sixteen possible states at each position of the Markov model with a corresponding 16x16 transition matrix. The transition matrix shown in Figure 8.3 is inadequate to the task. Instead we will use the following transition matrix.

Figure 8.4 shows a matrix of transition weights expanded to include enough states to accommodate interaction energies in positions separated by as many as three other amino acids. Listed along the left side and across the top are the sixteen different states at each point in the Markov model, with un-normalized transition probabilities in the matrix. The zeros in the matrix represent impossible states, where some amino acid would have to be simultaneously in helical and coil positions. Consider the top left position of the matrix, $h\overline{hhh} \rightarrow \overline{hhh}h$. This implies that there are five amino acids in a row that are in helical position. We are assigning the i to $i+4$ interaction energy between the amino acids in the first

$$\begin{array}{c|cccccccccccccccc}
 & \overline{hhh}h & \overline{hhh}c & \overline{hhc}h & \overline{hhc}c & \overline{hch}h & \overline{hch}c & \overline{hcc}h & \overline{hcc}c & \overline{chh}h & \overline{chh}c & \overline{chc}h & \overline{chc}c & \overline{cch}h & \overline{cch}c & \overline{ccc}h & \overline{ccc}c \\
\hline
h\overline{hhh} & x_j & w_j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h\overline{hhc} & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h\overline{hch} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h\overline{hcc} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
h\overline{chh} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 & 0 & 0 \\
h\overline{chc} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 \\
h\overline{cch} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
h\overline{ccc} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
c\overline{hhh} & y_j & w_j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c\overline{hhc} & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c\overline{hch} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c\overline{hcc} & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
c\overline{chh} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 & 0 & 0 \\
c\overline{chc} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_j & v_j & 0 & 0 & 0 & 0 \\
c\overline{cch} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
c\overline{ccc} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}$$

Figure 8.4: An un-normalized transition matrix that is large enough to account for interactions between amino acids that are separated by 4 positions.

and last positions and the i to $i+3$ interaction energy between the first and forth amino acids to the amino acid at position three. This is equivalent to a model in which we split the energy between the amino acids actually involved in the interaction because it is the total energy of the polypeptide that is relevant for the calculation of percent helicity. Additionally, in order to split the energy as described we would have to significantly increase the size of our state space.

8.2.3 Blocking Groups

Recall the partition function for this Markov model:

$$Z \;=\; v' \cdot \prod_{i=1}^{n} M_i \cdot w$$

The first and last amino acids in the polypeptide chain will obviously have neighbors to only one side. The first state cannot be $h\overline{hhh}$ as this would imply that the first amino acid has neighbors in helical position preceding it. Also, without a group preceding it the first amino acid cannot benefit from the stabilizing effects of hydrogen bonding. Similarly, the final amino acid cannot benefit from hydrogen bonding without a group following. In fact there are two types of polypeptides in our data. Those with "blocking groups" in the first and/or last positions, and those without. The blocking groups serve the purpose of allowing the first (or last) amino acid to participate in the formation of hydrogen bonds. In order to exclude impossible situations and accommodate blocking groups, we assign

$$\begin{aligned} v' &= (0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1) \\ w' &= (0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1) \end{aligned}$$

when the polypeptide is blocked, and

$$\begin{aligned} v' &= (0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1) \\ w' &= (0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1) \end{aligned}$$

when it is unblocked.

8.2.4 Capping and Positional Parameters

There are some reasons to believe that certain amino acids are seeds for α-helix formation. For example, Proline has a ring structure that incorporates its backbone. Because of this, its ϕ and ψ angles are severely restricted, causing this amino acid to be an α-helix breaker. It is, however, known statistically to be strongly associated with the amino terminus of α-helices.

Also, an α-helix creates an electric dipole that has the potential to interact with charged amino acids differently depending on whether they occur early or late in the helix. A negatively charged amino acid at the amino terminus would tend to favor the formation of an α-helix because its charge would complement the dipole which would put a positive charge at this end. For the same reason, it would discourage α-helix formation if it were located at the carboxy terminus.

With no modifications to our model, we have the ability to define $\Delta\Delta G$ parameters for: 1) the N-capping position (the first coil position before a helix corresponding to $**ch*$), 2) the C-capping position (the first coil after a helix corresponding to $*h\bar{c}**$), 3) the first and second positions of the α-helix ($*ch**$ and $chh**$ respectively), and 4) the last two positions of the alpha helix ($**hc*$ and $**hhc$). This is done by modifying the corresponding values in the transition matrix above.

There are also experiments which suggest that there are capping effects associated with the blocking groups. We treat amino capping by modifying v' and w' so that

$$\begin{aligned} v' &= (0,0,0,0,0,0,0,0,0,0,0,0,n_{cap},n_{cap},1,1) \\ w' &= (0,0,0,c_{cap},0,0,0,1,0,0,0,c_{cap},0,0,0,1) \end{aligned}$$

8.2.5 Model for pH

There are nine amino acids with ionizable side groups. Each of which occurs in protonated and unprotonated states. There is the potential for these state changes to cause significant differences in the probability of α-helix formation. The nine amino acids and their corresponding pK's are listed below (Creighton, 1993).

Amino Acid	side-chain pK
Asp	3.9
Glu	4.1
Arg	12.5
Lys	10.8
His	6.0
Cys	8.3
Ser	13
Thr	13
Tyr	10.1

We assign to each a ΔS_r associated with the unprotonated state. In addition, each of these five amino acids has a separate $\Delta\Delta S_r$ which is added when the amino acid is in its protonated state. We use this formulation so that we may use a variable selection prior on the parameter $\Delta\Delta S_r$ that is centered at zero.

We use the standard formula for calculating the fraction of an acid or base that is protonated, f_p, at a known pH (Munoz and Serrano, 1995b):

$$f_p = \frac{1}{1 + 10^{pH - pK}}$$

Thus the Gibbs energy associated with being in helical position for one of these nine amino acids is:

$$v_r = e^{f_p \Delta S_r + (1-f_p)(\Delta S_r + \Delta\Delta S_r) + C_p \log(t/t_{ref})}$$

In addition to the calculation of ΔS_r, we use split parameters according to the fraction of these amino acids that are protonated in the calculation of capping, positional, and interaction parameters. This allows us to obtain a reasonable fit for most pH curves in our data set. An exception is the polypeptide AETAAAK-FLRAHA, a chain that contains an $i - i + 8$ salt bridge between Glutamate at position two and Arginine at position 10. Our current model contains $i - i + 3$ and

$i - i + 4$ interactions. Expanding it to include $i - i + 8$ interactions would require that we use 128 states at each position in our Hidden Markov Model, significantly more than the sixteen we have now.

8.3 Prior Specification and Variable Selection

Consider what happens when we use a flat prior for ΔS_r, the energy required to constrain the amino acids in α-helical formation. Figure 8.5 shows the posterior mean distribution of, $exp(-T\Delta S_r/RT)$, for the 20 amino acids. Notice that Arginine is calculated to be an order of magnitude more helical than the other amino acids. In fact, it is highly likely that this reflects some bias in the data, though the origin of the bias is unclear.

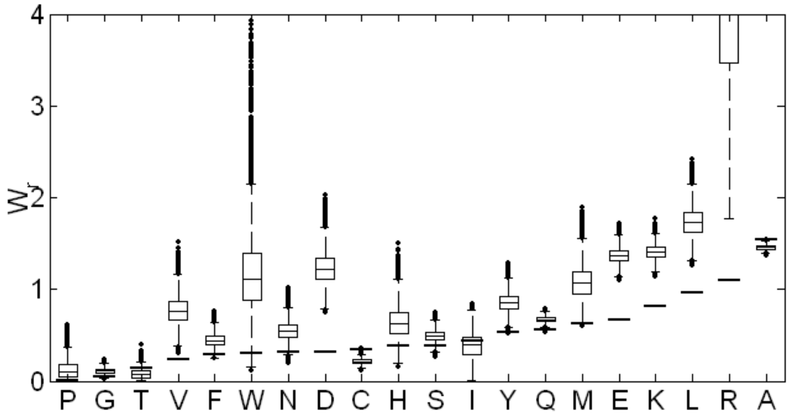


Figure 8.5: The calculated inherent helicity of each amino acid (box plots) is plotted with the biologically estimated helicity (lines). The biologically estimated values may differ from our calculated values because our statistical model can account for confounding effects which were not accounted for in the experiments. However, such a significant difference in the inherent helicity of Arginine is more likely due to bias in the data.

Significant work has been done experimentally to obtain estimates of many of the parameter values in this model (Chakrabartty *et al.*, 1994). We will try to avoid situations such as those shown in Figure 8.5 by capitalizing on this previous work. We define our prior distributions for ΔS_r and ΔH to be normal with a standard deviation of around 200 cal/mol and centered around the estimates provided in (Chakrabartty *et al.*, 1994). This allows us to somewhat alleviate the overfitting associated with over-parameterized models in general, and our model in particular.

Consider the temperature curves shown in 8.6. Notice that there is general disagreement in the slopes of the temperature curves predicted by Agadir and those measured in the lab (if the shapes were consistent, the sequential points would follow straight line paths). It is this disagreement that leads us to choose uninformative priors for both the specific heat parameter, C_p and the hydrogen bonding parameter ΔH.

We are faced with a data set of around 1100 data points and a model with the potential for over 900 parameters. Thus over-fitting is a significant concern. Because of this, we have chosen to implement variable selection in two different ways. The first is by the use of a point mass mixture prior of the form

$$(1-\pi)\delta_0 + \pi N(0, \sigma^2)$$

. For comparison, we will also implement a Laplacian (reflected exponential) prior. The MAP estimate from this prior is equivalent to the lasso method for shrinkage in linear regression Hastie *et al.* (2001). This prior distribution has the property that the maximum a posteriori (MAP) estimate of the parameters, in cases where there is little evidence of effect, is exactly zero.

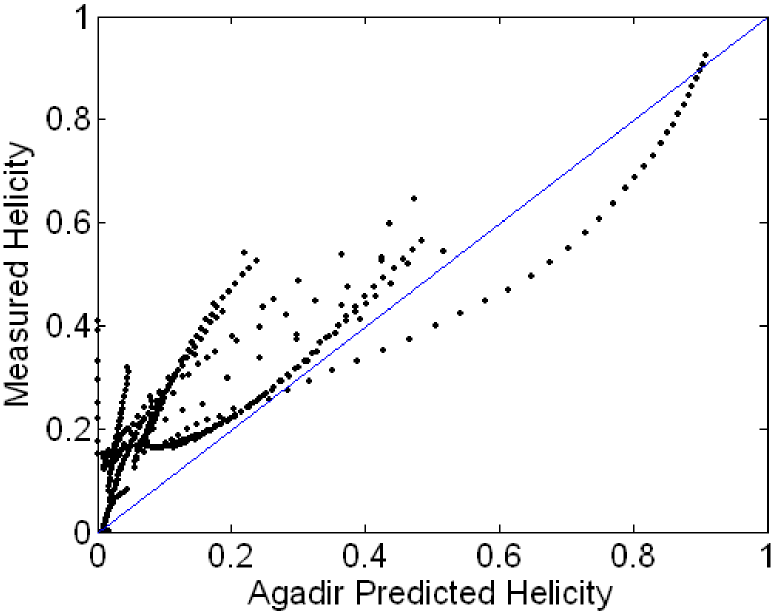


Figure 8.6: Chemically measured helicity values versus those predicted by Agadir. The points shown here are taken only from the temperature curves in the data. Because the shape of the Agadir curves is evidently off, we have chosen not to use informative prior distributions for either the specific heat or hydrogen bonding energy.

8.3.1 Markov Chain Monte Carlo for Variable Calibration

If we label the model for helicity described above as M, then for peptide x_i we have a statistical model for measured helicity h_i defined as follows:

$$\begin{aligned} h_i &= M(x_i) + \epsilon \\ \epsilon &\sim N(0, \sigma) \\ 1/\sigma &\sim Ga(\alpha, \beta) \end{aligned}$$

We have over 1100 data points, so any reasonably uninformative prior for σ will suffice to avoid biasing our results. We have chosen $\alpha = \beta = .01$.

Assuming an algorithm for using M to calculate the helicity of a given polypeptide, this is a very simple model with conditional posterior for σ of

$$\sigma|\{h_i\}, \{M(x_i)\} \sim Ga(N/2 + \alpha, \frac{1}{2}\sum(h_i - M(x_i))^2 + \beta$$

It is the details of calculation of helicity from M that cause issues.

For the cases where we have either normal or reflected exponential priors, we proceed by Metropolis steps in a Markov chain Monte Carlo algorithm. The calculation of helicity has been outlined in previous sections. Given a set of model parameters, θ, we may calculate a predicted helicity, $h_{i,\theta}$ for each polypeptide. If h_i is our measured helicity, then the conditional posterior of $h_{\theta,i}$ is:

$$P(\theta|\{h_i\}, \sigma) = \prod N(h_i|h_{\theta,i}, \sigma)p(\theta)$$

We propose a new set of model parameters, $\theta' \sim N(\theta, \Sigma)$ and accept them with probability

$$min\left(1, \frac{P(\theta'|\{h_i\}, \sigma)}{P(\theta|\{h_i\}, \sigma)}\right).$$

This produces a Metropolis Markov chain with stationary distribution equal to the posterior of interest. When we are using a point mass prior, we will proceed with the sampling algorithm outlined in section 7.3.

There are 658 model parameters, and we update them individually with a one dimensional random walk. As part of the burn-in process, we update the variance of the random walk individually for each parameter to keep the acceptance rate between 15% and 50%. This is done by keeping track of the number of accepted proposals for each 500 steps and incrementing or decrementing the variance accordingly.

8.3.2 Convergence

We run four separate chains with random start points, and use the squared error of prediction as a convergence diagnostic statistic. Figures 8.7 and 8.8 show the posterior mean squared errors from four different runs of the algorithm. There is close agreement among the chains, which suggests that the chains have converged. We have also monitored the traces of all individual parameters for agreement across all models.

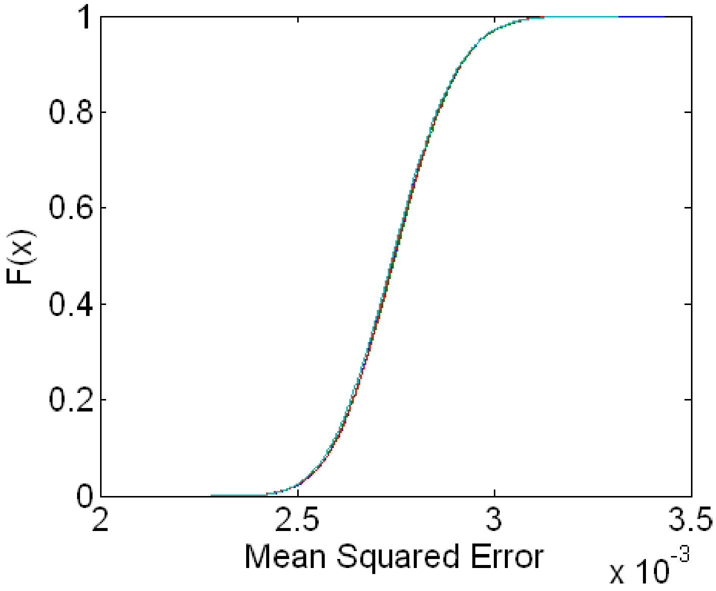


Figure 8.7: The cumulative distribution functions computed from the data for each of four different runs shows reasonable agreement.

8.4 Results

Dataset We have collected a database of peptide helicity measurements drawn from the published literature on helical peptide studies. In doing so we have incorporated most of the papers cited by Agadir (Munoz and Serrano, 1994) in

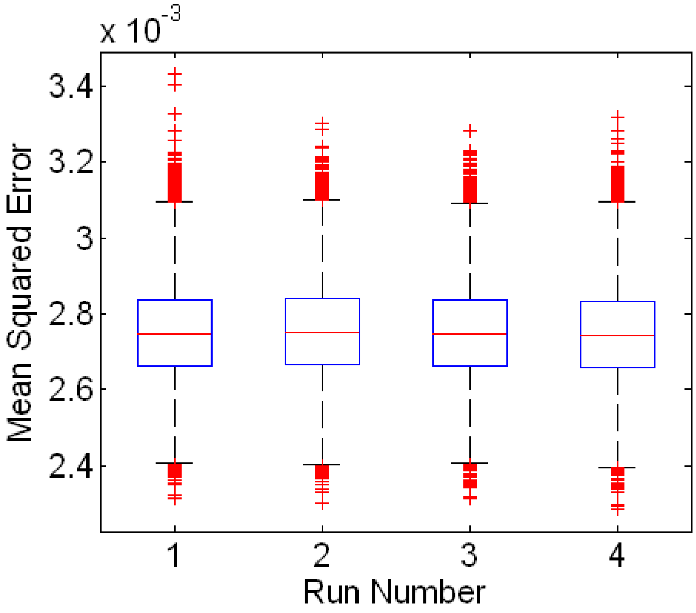


Figure 8.8: Boxplots from four different runs show close agreement in posterior mean squared error values.

an attempt to recreate as closely as possible the dataset used there. We have also added to our dataset a number of peptide helicity measurements published after the Agadir publication and 3 unpublished measurements, to obtain a "test set" on which neither Agadir nor our model have been trained for purposes of comparing prediction accuracy.

The dataset used in this paper contains 1187 peptide helicity values measured by circular dichroism (CD). The set contains 300 distinct peptides, including 142 designed and 218 natural sequences. The remainder of the data consists of repeated measurements on these peptides under various perturbed conditions, including 22 pH curves and 19 temperature curves. Most of the designed sequences are alanine-based peptides, and a large fraction of both the designed and native sequences are single mutations of other sequences in the database. Thus coverage of sequence space is far from uniform, and significant bias exists in amino acid composition.

Standard Deviation of Error Rate It is accepted that the error associated with the measurement of helicity is dependent on the technique used, but that it is around 5%. This is in line with the fitted error rate shown in 8.9.

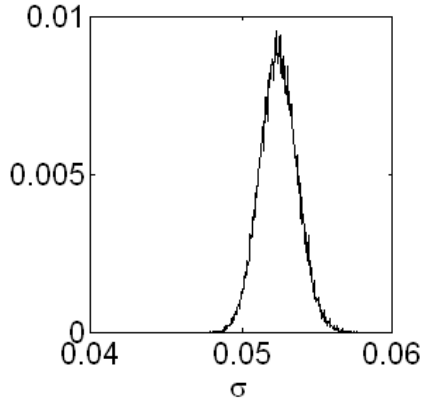


Figure 8.9: The posterior distribution of the standard deviation of the measurement error. There is close agreement between this value and the expected experimental error of around 5%.

Estimated parameters Figure 8.10 shows the resulting posterior intervals for the ΔS_R) parameters for each sidechain. Also shown for comparison are the parameters obtained from experimental measures Chakrabartty *et al.* (1994). Posterior intervals for the ΔS_r's were obtained by fitting the model described in the beginning of the chapter via the Bayesian inference procedure described in Section 8.3.1 using MCMC, and then transforming the Lifson-Roig w_r parameters at a temperature of 273 Kelvin. Sidechains are ordered according to experimentally measured helicities to enable comparison of trends. We see that Ala is estimated to one of the most helical residues, with Pro and Gly the lowest, as expected. We see that a number of residues encompass values higher than Ala. We are exploring the possibility that this may be due to sequence composition bias in the database arising from overrepresentation of designed peptides relative to native ones. However it is important to note that the experimental values do not necessarily reflect the true physical values; indeed, we expect to observe some differences and that these differences will improve the predictive ability of the model.

Figure 8.11a shows the posterior distribution obtained for ΔH. The Agadir value of -1.4kcal mol^{-1} is well within the posterior distribution for this parameter. The specific heat parameter, shown in Figure 8.11b, is positive, indicating an interaction favorable for melting of α-helices at higher temperatures.

Some posterior distributions for the subset of sidechain-specific i to $i+4$ interactions described in Section 8.2.2 obtained using laplace shrinkage priors are shown in Figures 8.12a and 8.12b. We found that complementary charges yield generally favorable interaction energies, while same charges result either in repulsion or no interaction, the latter occurring when sidechains are free to rotate away to avoid unfavorable interactions. Figures 8.12c and 8.12d show how these interactions can depend on how the side groups are charged. Several interactions are seen to be asymmetric, or dependent on the N/C-terminal ordering of the sidechains.

We estimate the N-Capping parameters of all amino acids as well as the favorable effects of being in either position 1 or position 2 of a helix. 8.13 shows the N-Cap parameter estimates for all 20 amino acids as well as that for the blocked amino terminus of the polypeptide chain.

All told, there are 67 of the 669 variables in the model which show evidence of being non-zero. These include the ΔH parameter, the heat capacity parameter, all of the ΔS_r parameters, one shift in pK (that of Glu), two shifts in ΔS for pH changes, 16 positional parameters and 26 interaction parameters. The Agadir model retains all of these parameters, and performs more poorly in both fitting and prediction.

Cross-validation results In order to evaluate predictive accuracy of the model in an unbiased fashion, it is important to look at test-set predictions on datapoints which were not used in estimating the model parameters. Many results reported to date for helix-coil theory models have been given in terms of accuracy in reproducing the data used for parameter determination, and results on out-of-sample peptides have been primarily anecdotal rather than across large samples. Such evaluations do not provide an accurate picture of the model's ability to predict the helicity of new peptides which were not present in the dataset at the time of parameter fitting. Without out-of-sample test set evaluations, "testing on the training set" as it is sometimes called is well-known to yield upwardly biased estimates of model accuracy.

One approach to obtaining unbiased estimates of predictive accuracy is the use of leave one out cross-validation. The resulting accuracy is estimated as the mean across the validation points, ensuring that every datapoint is available for validation, while avoiding training on the test set. Special care must be taken in performing cross-validation on the peptide helicity dataset described above, due to the high near-redundancy in the dataset. We partition the peptides into subsets according to sequence uniqueness, placing all repeated measurements, temperature curves, pH curves, and single-site mutations of each sequence in the same cross-validation subsets.

Figures 8.14a and 8.14b show the helicities predicted by our model versus experimental values for the training set and test set, respectively. The overall mean-squared-error (MSE) is .0021 for the training set and .0107 for test set, compared to .0178 and .0176 for Agadir. These results indicate that our parameter estimation and selection scheme simultaneously improves the model fit to the training data and the predictive accuracy on the test set.

Figure 8.15 shows real and predicted test-set temperature and pH curves for several peptides in the database. We see that even where the model curves depart from the experimental curves somewhat, the experimental data lies within the posterior predictive intervals of the model. This suggests that the error arises from lingering uncertainty in the model parameters, rather than a poor fit of the helix-coil model itself, and highlights another advantage of our statistical approach over previous algorithms such as Agadir (shown for comparison). In most curves the model correctly predicts the general shape of the curve.

For comparison we have included in Figures 8.18 to 8.29 all of the posterior distributions for the interaction parameters under both reflected exponential and point mass mixture priors. The histograms of the point mass mixture priors have been truncated on the y-axis in order to show the shape of the histogram away from zero.

8.5 Correlation in the Posteriors

One of the nice features of the posterior distribution for this model is the fact that most variables are close to independent of each other. Figure 8.16 shows the pairwise correlations of all 67 of the variables included in the model. There is generally low correlation with a few exceptions. As can be seen in Figure 8.17, five of the 15 variable pairs that have greater than 50% correlation are of the form $\Delta\Delta S_x$ vs ΔS_x. This is not surprising, because these parameters are unidentifiable for every ionizable amino acid for which has no pH curve in the data set that covers the range of its pK.

One of the features of variable selection techniques discussed thus far is the independence of the prior distributions. This structure precludes the use of prior knowledge of dependence structure, and can lead to overestimation of the significance of a particular variable. Figure 8.10 shows that the ΔS parameter for all five of these pairs is within a close range of its prior. This together with the fact that the two variables are known, a priori, to be confounding leads us to propose a new form of non-independent prior.

We have an informative prior for ΔS and, additionally, we propose to use a prior hierarchy that tells us that ΔS is a more important and more relevant parameter than $\Delta\Delta S$. This leads us to the following hierarchical prior structure:

$$\begin{aligned}
\Delta\Delta S &\sim \frac{(1-r)N(\Delta S|\mu,\sigma)}{(1-r)N(\Delta S|\mu,\sigma)+r}\delta_0 + \frac{r}{(1-r)N(\Delta S|\mu,\sigma)+r}N(m,v) \\
\Delta S &\sim N(\mu,\sigma)
\end{aligned}$$

That is to say, the farther the parameter ΔS ranges from its prior distribution, the more likely we are to allow the inclusion of the less important variable, $\Delta\Delta S$.

8.6 Conclusions

We have presented an approach to combining statistical mechanical models based on biophysical theory with databases of experimental measurements via Bayesian parameter inference and model selection. We have applied this approach to a frequently used model for biopolymer sequence and structure analysis, the helix-coil model. Our approach allows the incorporation of previous experimental and theoretical knowledge in the form of prior information on model parameters and model structure, and combines this with a large dataset to obtain posterior distributions on model parameters. This general approach has applications to a wide variety of problems in biostatistics and biophysics. Our approach may be applied directly to the problem of protein secondary structure prediction using helix-coil models (Froimowitz and Fasman, 1974; Qian and Chan, 1996; Misra and Wong, 1998), with little modification. Of broader interest may be problems in empirical forcefield parameterization for protein structure prediction by threading and homology modeling, fragment reconstruction, and empirical energy minimization, as well as problems in protein-protein and protein-ligand docking.

In this chapter we have applied our approach to model selection and parameter estimation in the context of peptide helicity prediction, and shown that this approach provides both improved fit to the training data and improved test-set predictive performance. The use of shrinkage priors enables the inclusion of a large number of potential sidechain interaction and capping parameters, yet still induces a sparse model structure by retaining only those parameters for which significant evidence exists in the data set. Upon study, we find that the error in fitting helicity has heavier tails than a normal distribution. There is the opportunity to improve fitting and prediction by using heavier tailed T-distributions. This extension can be accomplished by mixtures of normal distributions, and therefore would require little modification of our treatment.

In the following chapter, we will discuss how our approach enables the application of statistical design of experiments to select future helical peptide studies which will be most informative in improving model predictive accuracy, reducing uncertainty in parameters, and examining hypotheses about model structure. This approach represents a promising new way to bridge the gap between detailed physical modeling done in computational biophysics and computational chemistry, and more traditional data-driven bioinformatics algorithms.

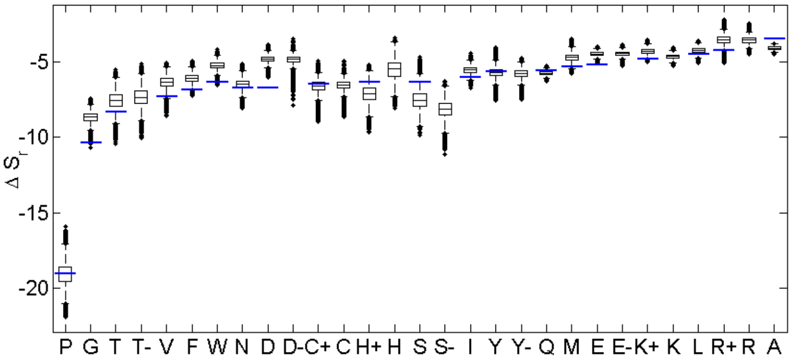
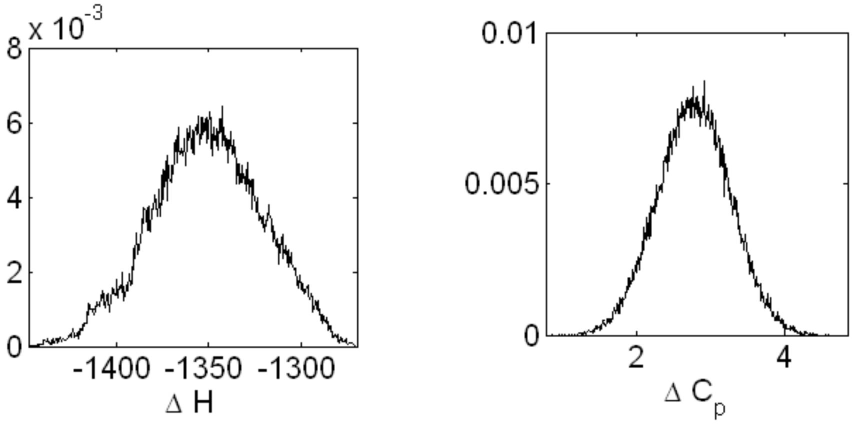


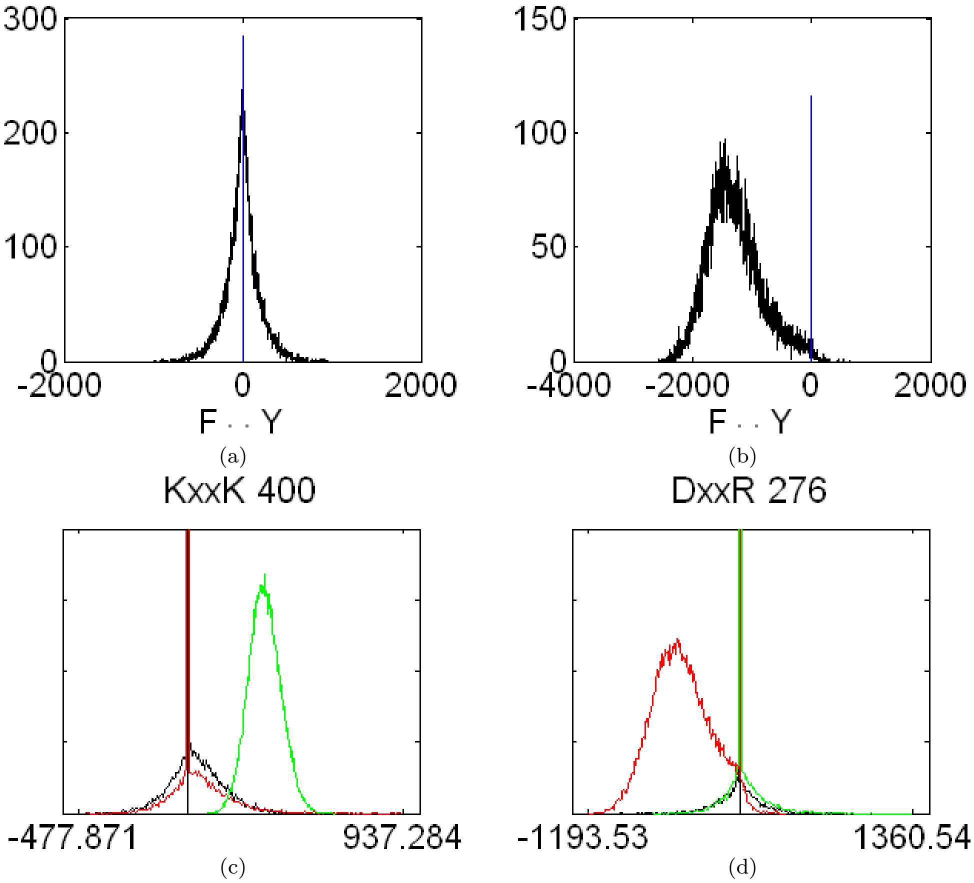
Figure 8.10: Posterior distributions for ΔS_r parameters. Boxplots show posterior means, quartiles, and 95% posterior intervals. Amino acids are ordered by LR w values obtained previously Chakrabartty *et al.* (1994). These values, converted to ΔS_r scales, are shown in red. Differences between experimental values and posterior intervals are discussed in text.



(a)

(b)

Figure 8.11: The temperature parameters show reasonable agreement with the standard values. The Agadir value, $\Delta H \approx -1400$cal/mol, is well within the posterior distribution. ΔC_p is in units of cal/(mol Kelvin).



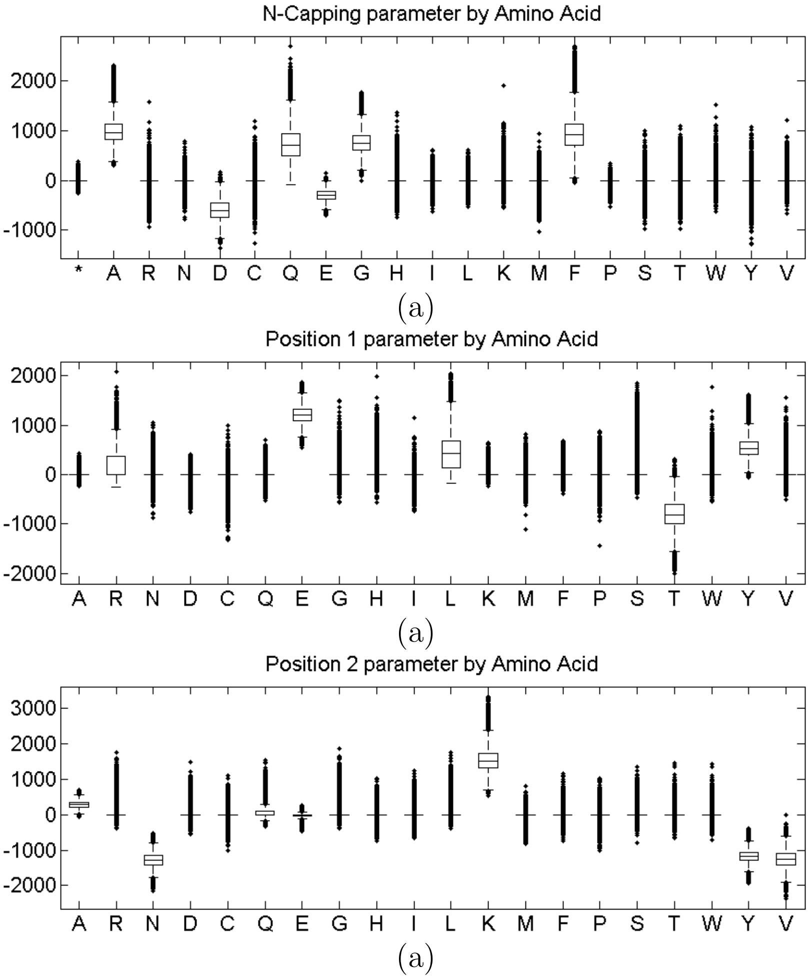
(a)

(b)

(c)

(d)

Figure 8.12: When there is no evidence for a non-zero parameter in the data, we get back the prior distribution. In (a), there is only one observation of F - Y in the i to $i+3$ position in the data set. The corresponding posterior using the point mass prior puts a 93% probability that the parameter is zero. In (b), there are 133 observations in the data set, and this allows us to determine that this parameter is likely non-zero. The point mass prior puts a 35% probability that this parameter is non-zero. Our model distinguishes between charged and uncharged versions of particular side groups. For Figures (c) and (d), the black curve corresponds to the absence of protons (high pH), the red curve corresponds to the presence of one proton among the pair, and the green curve corresponds to the presence of two protons. For figure (c), we see that there is an unfavorable interaction only when both Lys side groups are charged. For Figure (d) we see evidence of a favorable interaction only when both E and R are charged (oppositely in this case).


N-Capping parameter by Amino Acid

(a)

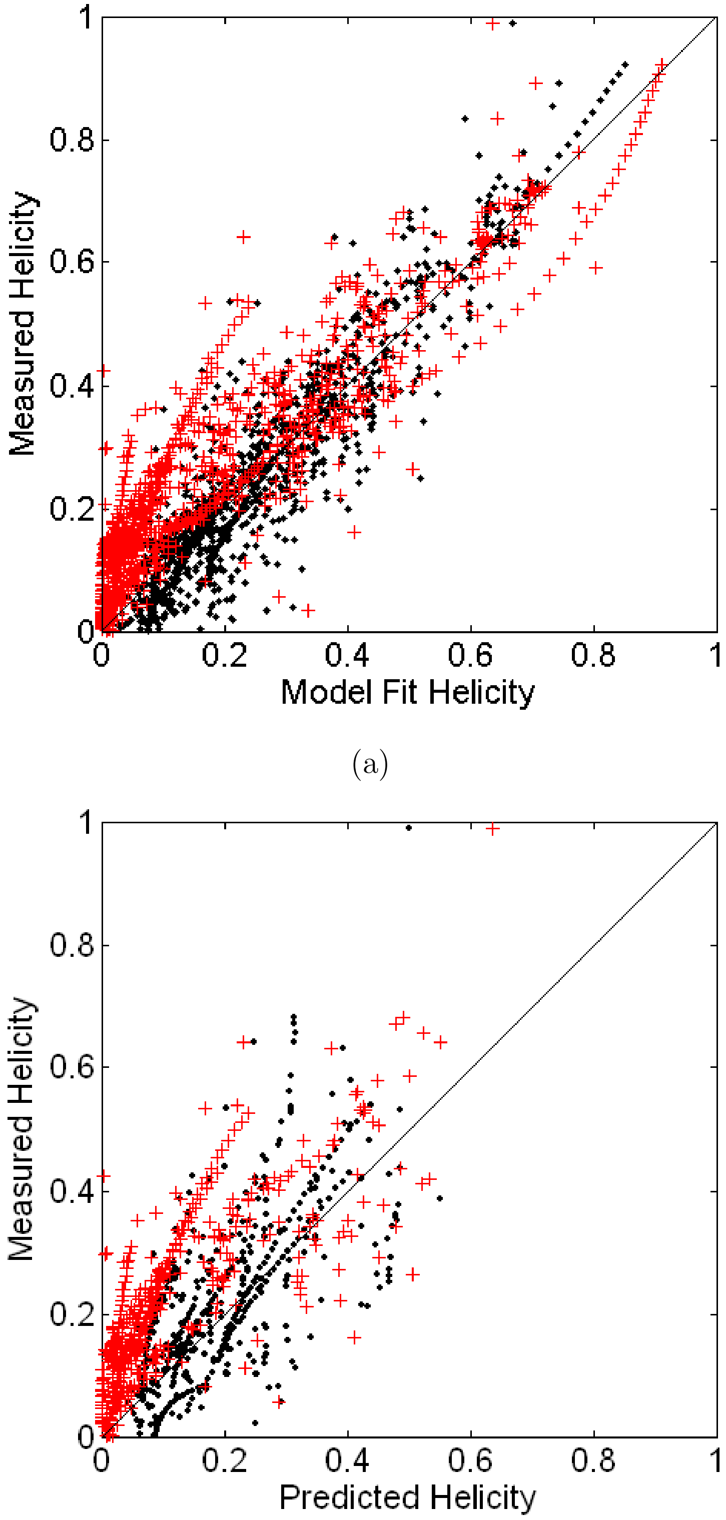
Position 1 parameter by Amino Acid

(a)

Position 2 parameter by Amino Acid

(a)

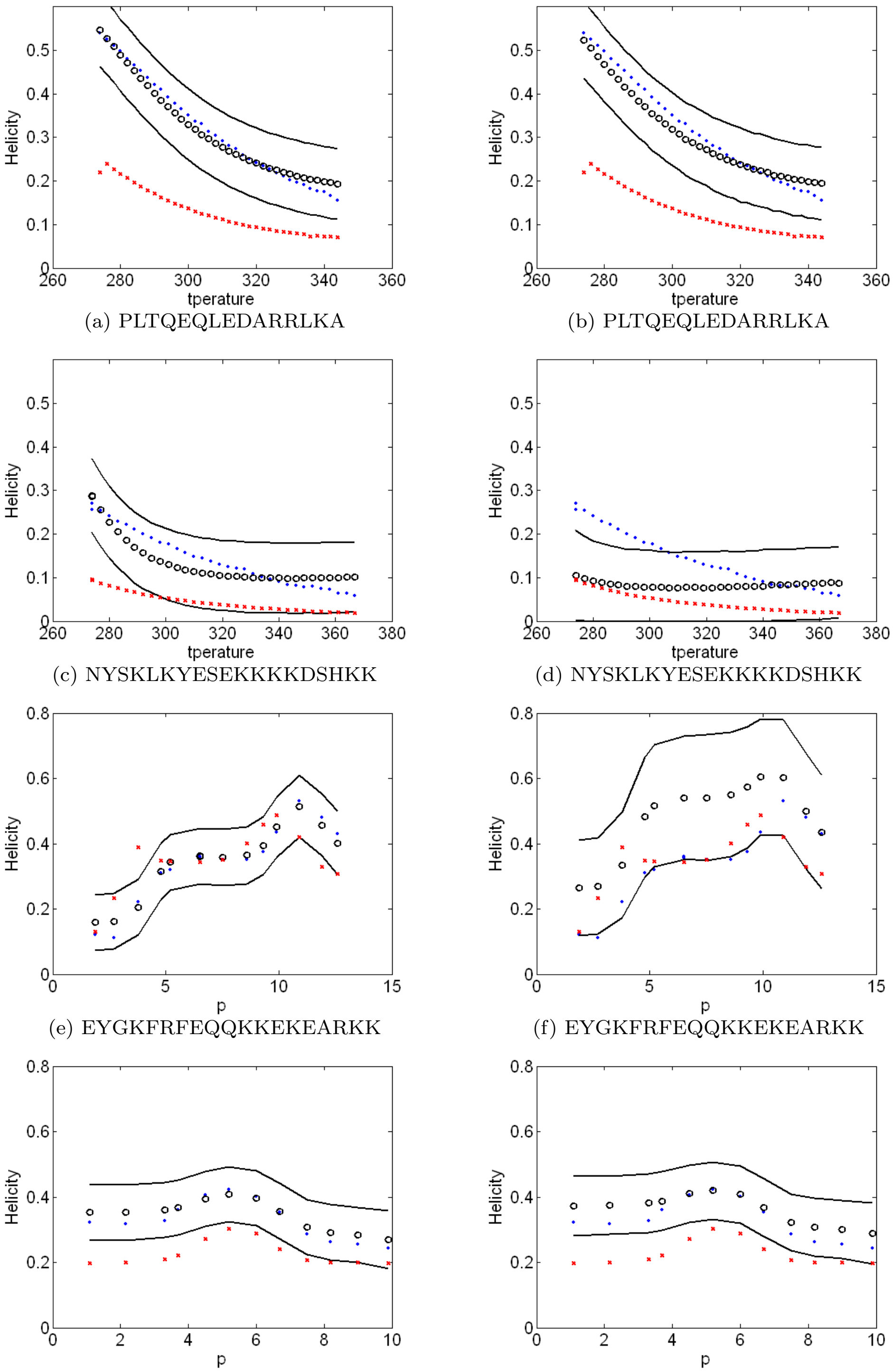
Figure 8.13: Many of the N-Cap parameters (a) are non-zero. There is some cause for concern here, as these parameters are generally considered to be of lesser importance. One of these that is known to be important is the parameter for the amino terminus blocking group (labeled * in the figure). Some positional parameters (b) and (c) are also strongly non-zero



(a)

(b)

Figure 8.14: There is good agreement between known helicities and both fit (a) and out of sample (b) model values. Also, there is a significant improvement over the Agadir model.



(a) PLTQEQLEDARRLKA

(b) PLTQEQLEDARRLKA

(c) NYSKLKYESEKKKKDSHKK

(d) NYSKLKYESEKKKKDSHKK

(e) EYGKFRFEQQKKEKEARKK

(f) EYGKFRFEQQKKEKEARKK

(g) AETAAAKFLRAHA

(h) AETAAAKFLRAHA

Figure 8.15: Predicted helicity versus experimentally measured temperature and pH curves for several peptide sequences. Predictions by our model (open circles) are shown with 95% posterior predictive intervals. Fitted curves are shown in the left column and predictions from cross-validation are shown in the right column. Agadir is shown in red and measured data values are in blue.

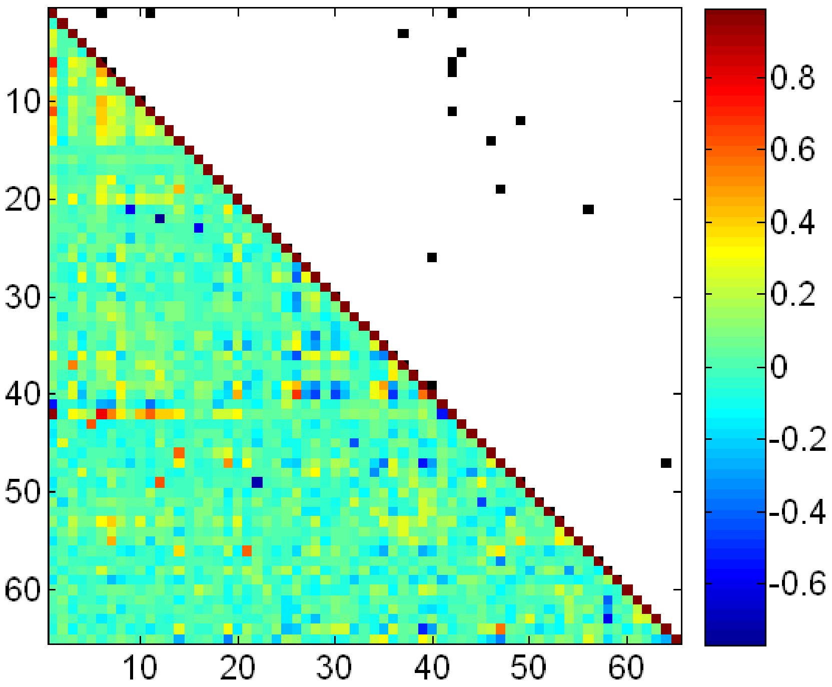


Figure 8.16: A heatmaps of the pairwise correlations between the 67 variables that are included in the model. Below the diagonal shows a standard heatmap, and above the diagonal highlights those pairs that have correlation greater than 50%. There is generally low correlation between the variables.

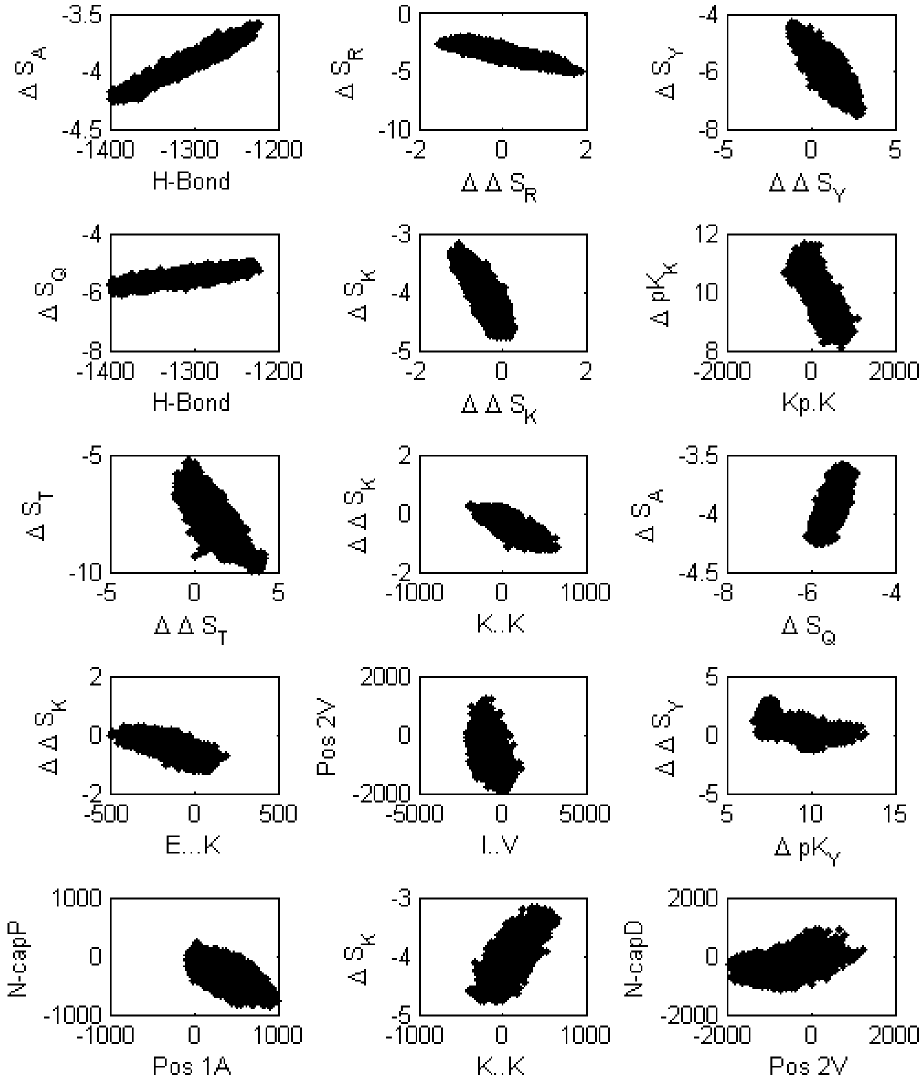


Figure 8.17: Scatterplots of the 15 pairs of variables that have higher than 50% correlation. Five of the 15 are of the form $\Delta\Delta S_x$ vs ΔS_x.

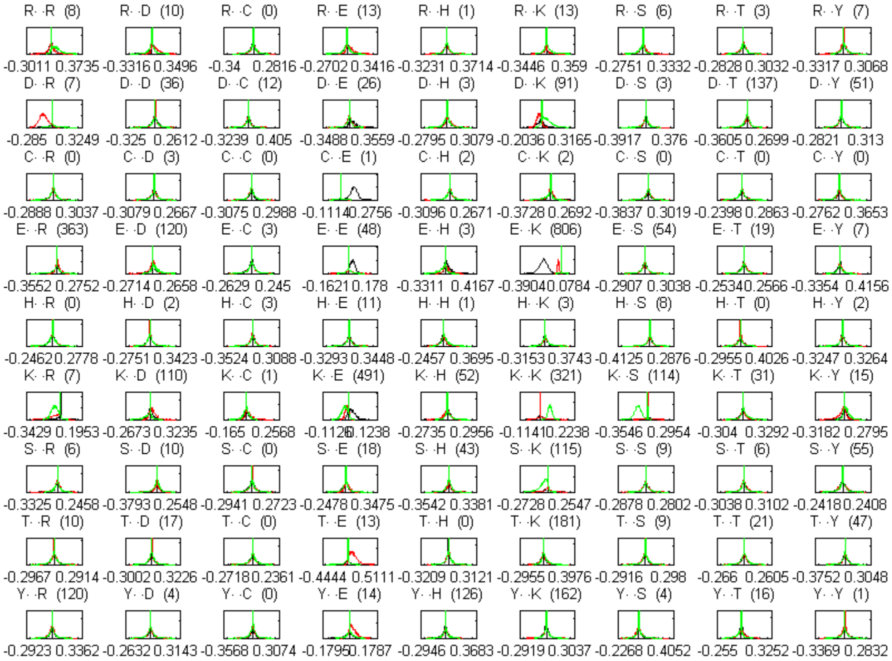


Figure 8.18: The posterior distribution for the i to $i+3$ charged group parameters under the point mass mixture prior.

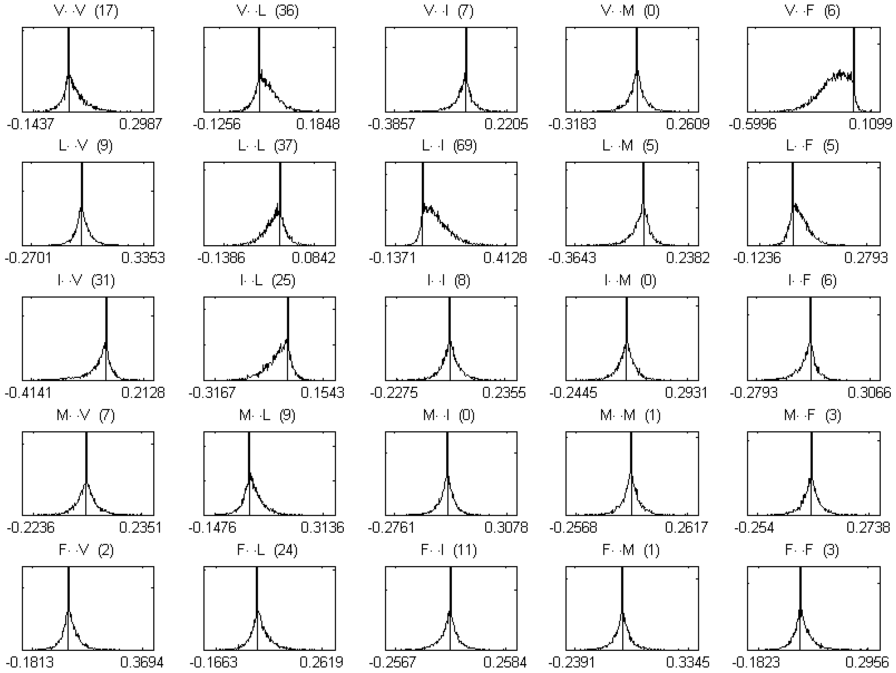


Figure 8.19: The posterior distribution for the i to $i + 3$ hydrophobic group parameters under the point mass mixture prior.

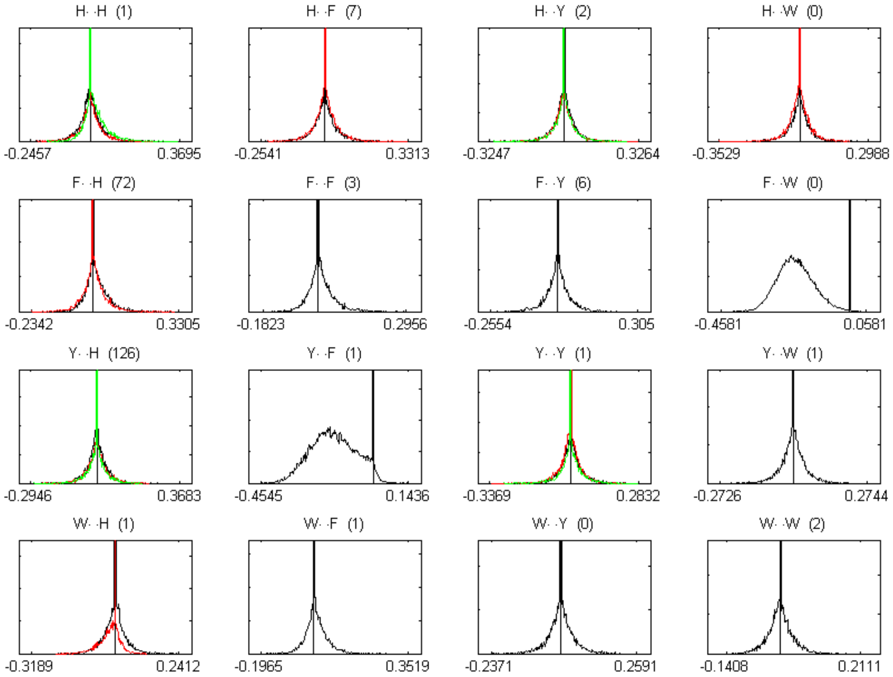


Figure 8.20: The posterior distribution for the i to $i+3$ ring structure group parameters under the point mass mixture prior.

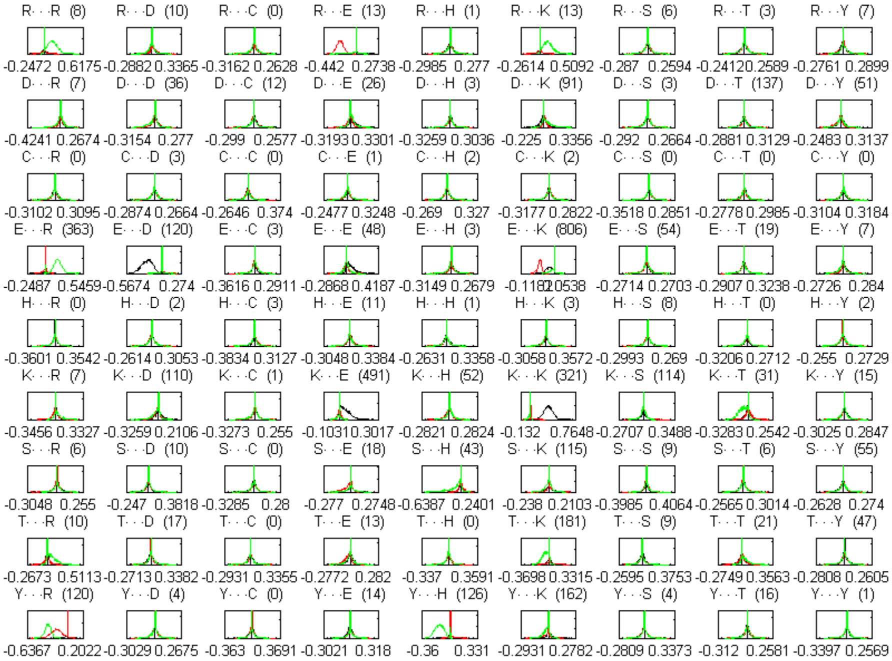


Figure 8.21: The posterior distribution for the i to i+4 charged group parameters under the point mass mixture prior.

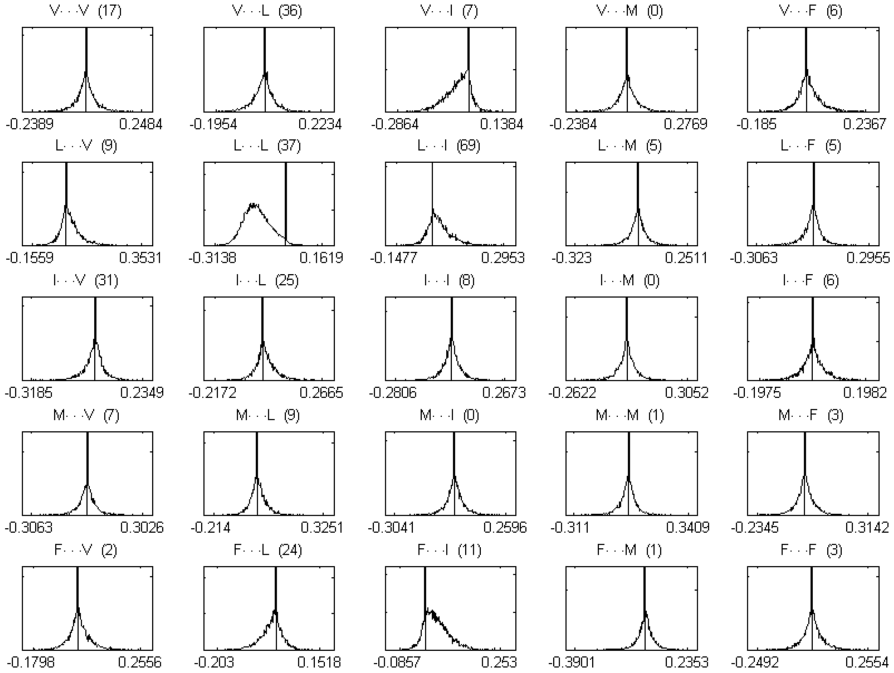


Figure 8.22: The posterior distribution for the i to $i + 4$ hydrophobic group parameters under the point mass mixture prior.

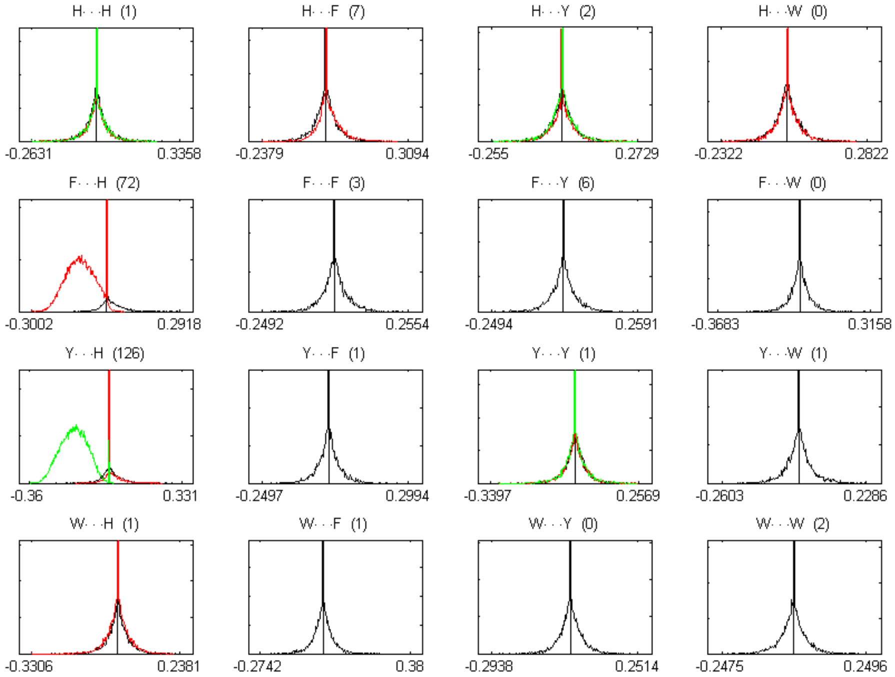


Figure 8.23: The posterior distribution for the i to $i+4$ ring structure group parameters under the point mass mixture prior.

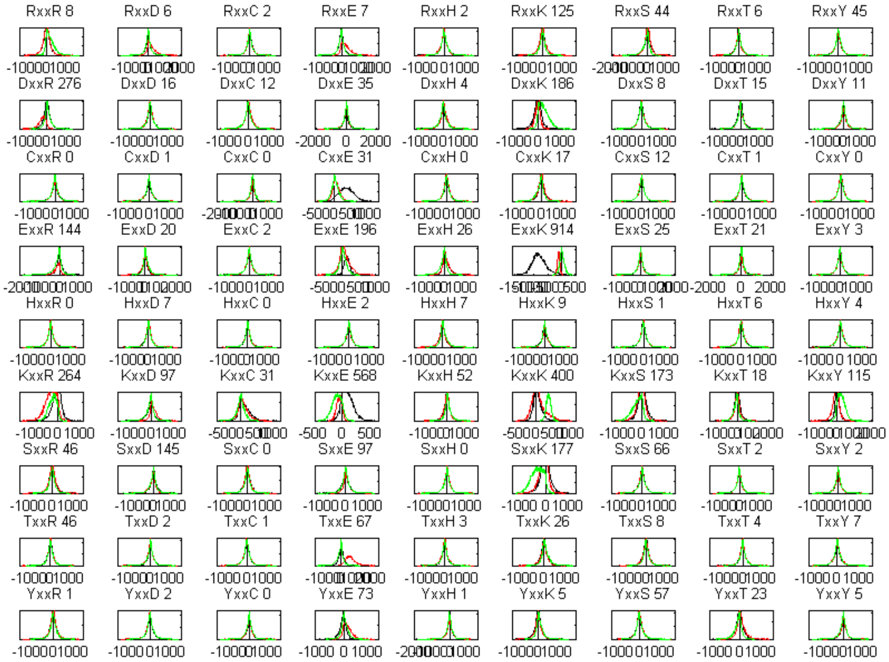


Figure 8.24: The posterior distribution for the i to i+3 charged group parameters under the Laplacian prior.

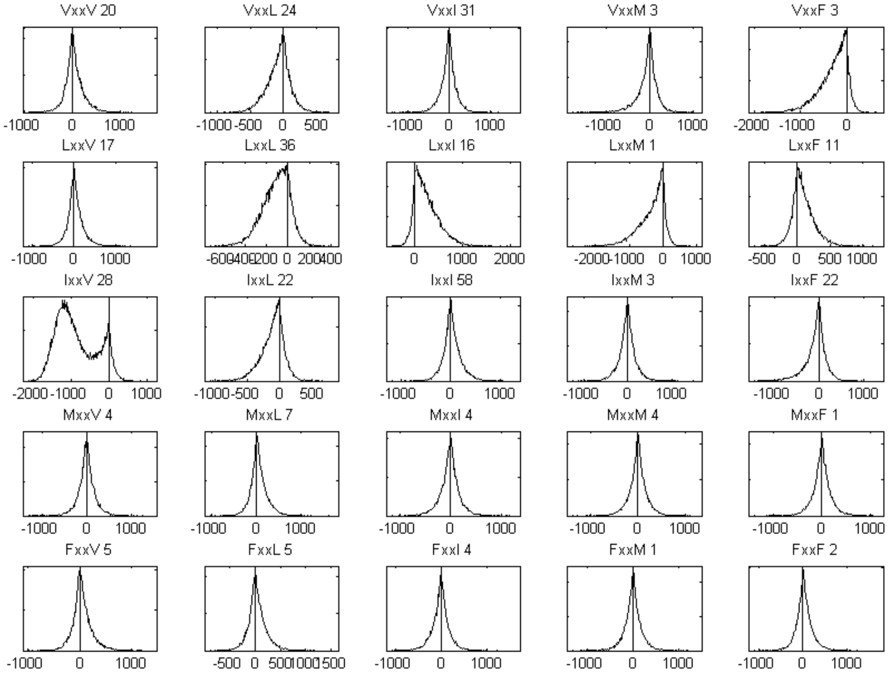


Figure 8.25: The posterior distribution for the i to $i + 3$ hydrophobic group parameters under the Laplacian prior.

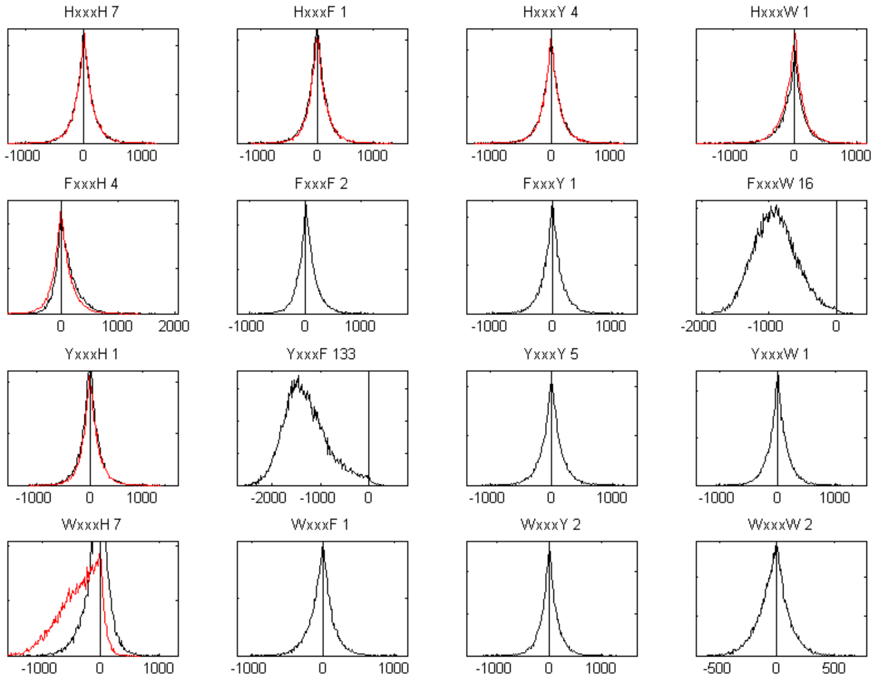


Figure 8.26: The posterior distribution for the i to $i+3$ ring structure group parameters under the Laplacian prior.

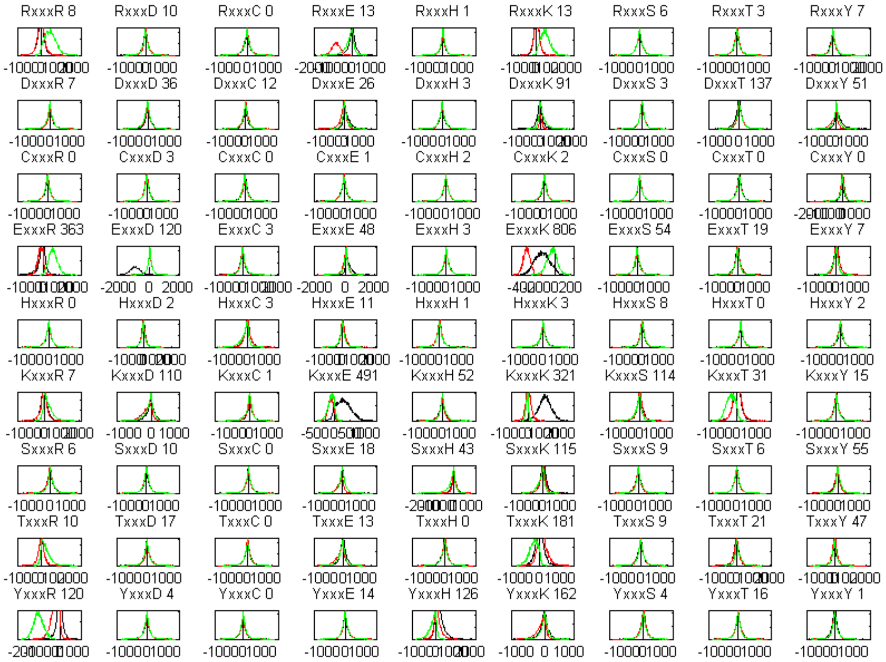


Figure 8.27: The posterior distribution for the i to i+4 charged group parameters under the Laplacian prior.

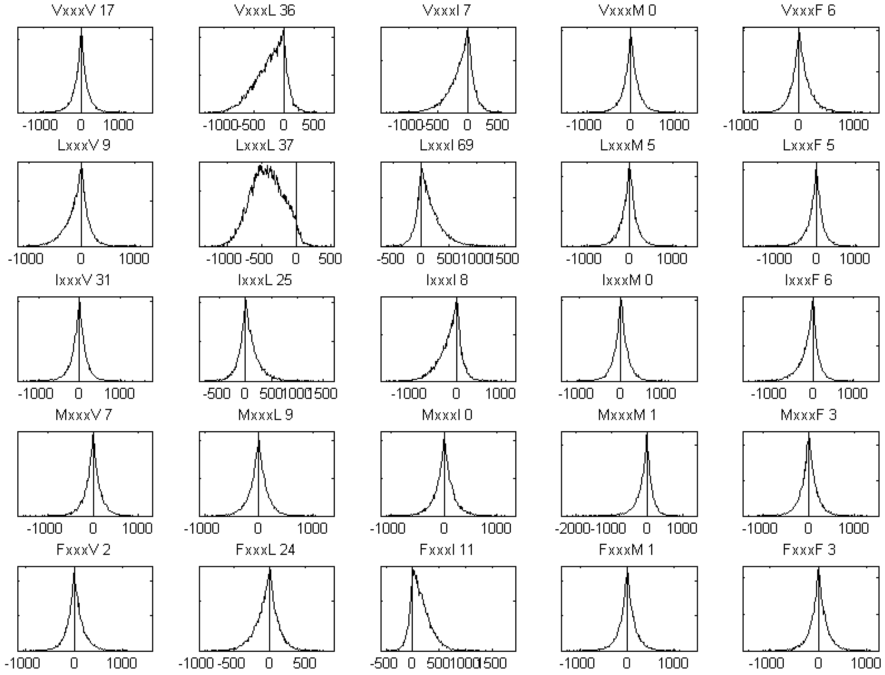


Figure 8.28: The posterior distribution for the i to $i + 4$ hydrophobic group parameters under the Laplacian prior.

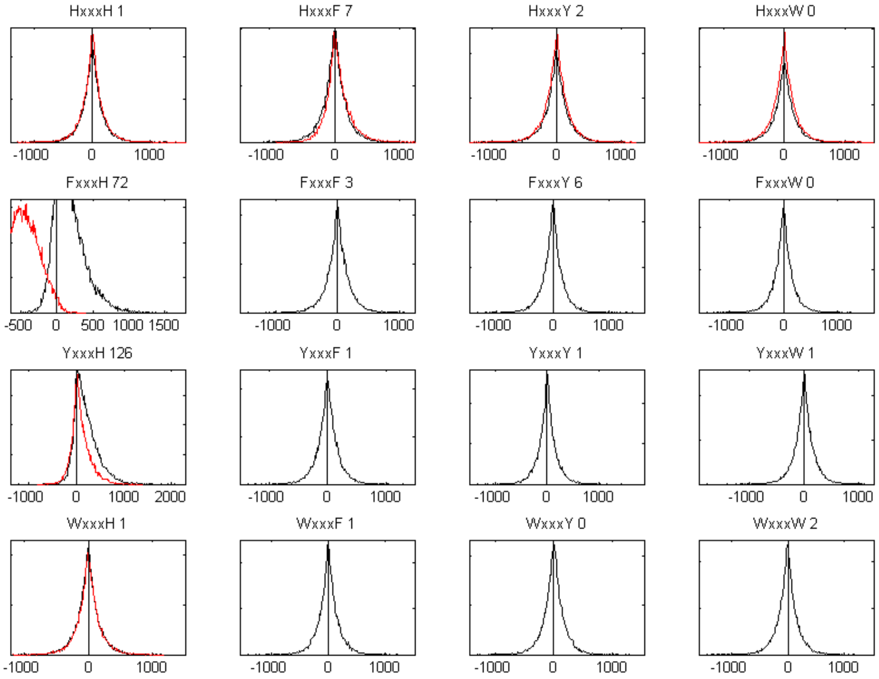


Figure 8.29: The posterior distribution for the i to $i + 4$ ring structure group parameters under the Laplacian prior.

Chapter 9

Experimental Design for Future Peptide Helicity Prediction

The array of different potential polypeptides is vast. Because of this, randomly choosing one to study is unlikely to significantly improve our ability to predict helicity in future polypeptides. Unfortunately, experimentation is time consuming and expensive, so there is a need to maximize, as much as possible, the information obtained. Our model along with the posterior distribution on the parameters derived in chapter 8 provide us with enough information to give some direction for improving the model by future study.

The idea of making choices about the next avenue of study based on previous data is not a new one. In 1948, Dixon and Mood (1948) described an iterative 'up and down' technique for estimation of median effective dose. Other work on this problem includes the stochastic approximation method of H. and Monro (1951) and a Bayesian technique introduced by Freeman (1970).

There is similar work on the problem of sequential clinical trials, in which subjects are added to the trial as they come. If one wants to estimate a particular parameter in a model, for example the difference in expression levels of a particular gene between control and experimental groups, then one must make a trade off between bias and variance. In general, it is desirable to have balance between control and experimental groups, but by deterministically assigning new patients to groups (rather than random assignment) one runs the risk of introducing bias. Some examples of this work include a variance minimization rule (Pocock and Simon, 1975) and a 'biased coin' rule (Efron, 1971). A comparison of these and other experimental design techniques is presented in (Atkinson, 2002).

One difficulty with developing a general theory of sequential design is the variety of potential models and potential parameters of interest within a particular model. An attempt at generalizing the approach is to try to maximize the determinant of the information matrix (Wynn, 1970). Experimental designs which attempt do this are called D-optimal designs. (In fact, there is a whole series of related optimality criteria, called alphabetical optimality criteria.) A description of such a technique for generalized linear models is given by Dror and Steinberg (2006), and for certain semi-parametric models by Verotta (1990). Additionally, sequential D-optimal designs are used in a number of applied papers including Coffey *et al.* (2005), Berger (1994), and Fujiwara *et al.* (2005). This technique is helpful when one wants to learn generally about the parameters in a model.

The general Bayesian approach to experimental design, and the approach we will take, is to specify a utility function, then choose the experimental design which maximizes the expected utility. (This is equivalent to designating a loss function and minimizing expected loss.) To convert this to sequential experimental design, one simply recalculates at each step the maximum expected utility experiment given any new data that has been discovered. For the case of median effective dose, the loss function is the difference between the estimated and actual median dose. For the case of clinical trials, the loss may be the variance of the estimator of the difference in response between experimental and control groups. Also, D-optimality has been show to be a special case of the Bayesian approach which assumes a particular utility function (Bernardo, 1979). A comprehensive survey of Bayesian experimental design and decision theory is given in Chaloner and Verdinelli (1995).

In our application we have a clear choice of loss function. It is the helicity of polypeptides that we want to predict. While it is true that we have a physical interpretation of our model parameters, it is predictive accuracy that drives our study. Thus, we choose to make expected predictive error our loss function and define our utility function to be $-1\times$ our loss.

9.1 Notation

For the purposes of this chapter, let $\mathbf{X}$ be the set of all polypeptides. Let the "true" helicity of $x_i \in \mathbf{X}$ be h_i and the experimentally measured helicity be $\bar{h}_i$. Suppose that we have a set of parameters for the model described in the previous chapter, $\theta \in \Theta$. Define $\hat{h}_{i,\theta} = h(x_i, \theta)$ to be the model predicted helicity of polypeptide x_i given model parameters θ. Then, in this notation and fixing parameters θ, our model is of the form $\bar{h}_i \sim N(\hat{h}_{i,\theta}, \sigma)$. Let $R(\theta)$ be the log of the prior probability of θ. If we have helicity measurements on a subset, $\{x_n\}_{n=1}^{N}$, of polypeptides, then the conditional log-density of $\theta|\{\bar{h}_n\}, \sigma = R(\theta) + \sum_n -\frac{(\bar{h}_n - \hat{h}_{n,b})^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$. In order to limit the excess of notation, we will denote the probability density function for θ as $f(\theta|\{\bar{h}_n\}, \sigma) = f(\theta)$ (with an implicit assumption of dependence on σ and the data). Our model estimated helicity for polypeptide x_i, from the previous chapter, will be

$$\hat{h}_i|\{\bar{h}_n\} \quad = \quad \int_{\Theta} \hat{h}_{i,\theta} f(\theta) d\theta$$

In the coming sections, we will approximate this integral by Monte carlo integration (see chapter 8).

Recall that our general goal with this work is the prediction of helicity, so it is intuitive that we will use the squared error of prediction as our loss function. The squared error of prediction for polypeptide x_i is simply $(\hat{h}_i - h_i)^2$. We will also refer to this as the predictive accuracy.

We are faced with a somewhat ill-defined question. We are generally interested in improving the predictive accuracy of our helicity model through better understanding of its parameters. However, this begs the question, "Predictive accuracy over what set of polypeptides?".

It is possible to answer this question in various ways, for example, predictive accuracy for all polypeptides of length k. (Which would be a set of order 20^k.) However, it is probably true that not all of these are equally interesting. Additionally, fixing a length, k, is too restrictive.

Define a distribution $\omega(x)$ on all polypeptides (potentially zero at some points). By a choice of ω an experimenter may define which polypeptides are considered important. Possibilities may include high or low helicity polypeptides, naturally occurring polypeptides, polypeptides that are resistant to temperature or pH changes, etc. Having defined ω, we will define our expected predictive accuracy over ω to be

$$A_{\omega,\{\bar{h}_n\}} = \sum_{\mathbf{X}} (h_i - \hat{h}_i)^2 \omega(x_i).$$

We are ultimately interested in determining what polypeptide to study next. We have defined a ranking of important polypeptides by way of ω, and an expected predictive accuracy. Our change in predictive accuracy from studying polypeptide x_i is then

$$A_{\omega,\{\bar{h}_n\}} - A_{\omega,\{\bar{h}_n\}\cup\bar{h}_i}.$$

Because we do not have $\bar{h}_i$ available when making the decision whether or not to study x_i, we must integrate across its possible values. Recall that the density for $\bar{h}_i$ is $N(\hat{h}_{i,\theta}, \sigma)$. We define our expected change in expected predictive accuracy for polypeptide x_i to be

$$E_{\omega,\{\bar{h}_n\}}[x_i] = \int_\Theta \int_{[0,1]} A_{\omega,\{\bar{h}_n\}} - A_{\omega,\{\bar{h}_n\}\cup\bar{h}_i} N(\bar{h}_i|\hat{h}_{i,\theta}, \sigma) d\bar{h}_i d\theta.$$

This is our expected gain from studying polypeptide x_i.

We are considering the decision theoretic statistic of expected utility, with a loss function defined by predictive accuracy error. In fact, there are other considerations when deciding what polypeptide to study. First, synthesis of longer polypeptides may be more expensive than shorter. Also, introducing a series of 20 point mutations to a particular polypeptide may be less expensive than studying 20 completely distinct polypeptides. If we let $C(x_i)$ be our expected cost for polypeptide x_i, then our expected utility is

$$E(U(x_i)) = E_{\omega,\{\bar{h}_n\}}[x_i] - C(x_i).$$

Costs may vary from lab to lab. Because of this and because adding cost to the calculation can be added on after calculation of $E_{\omega,\{\bar{h}_n\}}[x_i]$, we will ignore cost in what follows. Additionally, given a distribution, ω, on $\mathbf{X}$ we may approximate the expected utility of any polypeptide by sampling from ω and averaging the expected utility of x_i across the sample. Finally, for the process of averaging expected utility, we need only calculate the pairwise expected utility for each element of the sample with x_i. Thus, for the remainder of the paper, we will concentrate on the calculation of a simple pairwise expected utility. Specifically, we will answer the question "If I study polypeptide x_a, what is the expected improvement in predictive accuracy for polypeptide x_b?" With the ability to answer this question, one might perform a random search of the polypeptide space (for example, the Shotgun search algorithm introduced by Hans *et al.* (2007)) in order to find good candidates for future study.

9.1.1 Expected Loss

Suppose we are interested in a particular polypeptide, x_a. We obtain the expected squared error, Λ_a, by integrating over the distribution of h_a. We know $h_a \sim N(\hat{h}_{a,\theta}, \sigma)$.

$$\begin{aligned}
\Lambda_a &= \int_\Theta \int_{[0,1]} (\hat{h}_a - h_a)^2 N(h_a|\hat{h}_{a,\theta}, \sigma) dh_a f(\theta) d\theta \\
&= \int_\Theta \int_{[0,1]} [(\hat{h}_a - \hat{h}_{a,\theta}) + (\hat{h}_{a,\theta} - h_a)]^2 N(h_a|\hat{h}_{a,\theta}, \sigma) dh_a f(\theta) d\theta \\
&= \int_\Theta \Big[E_{h_a}[(\hat{h}_a - \hat{h}_{a,\theta})^2] + E_{h_a}[(\hat{h}_a - \hat{h}_{a,\theta})(\hat{h}_{a,\theta} - h_a)] + E_{h_a}[(\hat{h}_{a,\theta} - h_a)^2]\Big] f(\theta) d\theta \\
&= \int_\Theta \Big[(\hat{h}_a - \hat{h}_{a,\theta})^2 + \sigma^2\Big] f(\theta) d\theta
\end{aligned}$$

This is our expected predictive error (expected loss) for polypeptide x_a. As described in the previous chapter, we have an algorithm for sampling from the distribution, $f(\theta)$. If $\{\theta_i\}_{i=1}^{N}$ is such a sample, we may approximate Λ_a as

$$\begin{aligned}\Lambda_a \approx \hat{\Lambda}_a &= \frac{1}{N}\sum_i (\hat{h}_a - \hat{h}_{a,\theta_i})^2 + \sigma_i^2 \\ &= \frac{1}{N}\sum_i (1/N\sum_j \hat{h}_{a,j} - \hat{h}_{a,i})^2 + \sigma_i^2 \\ &= \sigma_{h_a}^2 + \sigma_i^2\end{aligned}$$

Let $\bar{h}_a$ denote the experimental measurement that would be obtained if we were to study peptide x_a. Let $g(\theta) \propto p(\theta|\{\bar{h}_n\}, \bar{h}_a)$ be the distribution on the parameters of our model when we include the "measurement", $\bar{h}_a$. Then the updated loss for polypeptide x_b that occurs due to the study of polypeptide x_a is

$$\Lambda_b^* = \int_\Theta (\hat{h}_b' - \hat{h}_{b,\theta})^2 + \sigma^2 g(\theta) d\theta$$

where $\hat{h}_b'$ is calculated over the density $g(\theta)$. As we are interested in discovering polypeptides for potential study, we must assume that we do not know $g(\theta)$. We may integrate over potential values for $\bar{h}_a$. This gives us an expected loss of

$$\Lambda_b^* = \int_{t\in[0,1]} \int_\Theta (\hat{h}_b' - \hat{h}_{b,\theta})^2 + \sigma^2 g(\theta|\bar{h}_a = t) d\theta \int_{q\in\beta} N(t|\hat{h}_{a,\theta}, \sigma^2) f(\theta) d\theta dt$$

This can not be calculated analytically, so we will approximate it numerically. One approach is through numerical integration over t. If we discretize t, then we must re-calculate $g(\theta)$ for each point in the discretization. Even supposing some form of quadrature rule that allows us to approximate the integral with only a very rough discretization, we must redraw the set $\{\theta_i\}$ multiple times for each value of t and each pair of polypeptides we want to compare. Since such draws from $f(\theta)$ involve a large MCMC simulation requiring multiple days, this sort of brute force calculation is prohibitive for more than a few pairs of points.

Alternatively, if we assume that the addition of a single point, $\bar{h}_a$ to our list of known helicities will not have a drastic affect on the posterior for our parameter space $(p(\theta|\{\bar{h}_n\}, \bar{h}_a) \approx P(\theta|\{\bar{h}_n\})$ then our new estimate of the expected squared error of helicity estimation for x_b can be obtained from the same sample used before, $\{\theta_i\}_{i=1}^N$ where $\theta_i \sim f(\theta) \forall i$, via importance sampling:

$$\Lambda_b^* \approx \hat{\Lambda}_b^* \quad = \quad \sum_i \frac{g(\theta_i)}{f(\theta_i)} [(\hat{h}_b' - \hat{h}_{b,\theta_i})^2 + \sigma_i^2] \tag{9.1}$$

*Note: $\hat{h}_b'$ is the model estimated helicity calculated over the distribution $g(\theta)$.

In order to calculate Λ_b^*, we must be able to quickly evaluate $\frac{g(\theta_i)}{f(\theta_i)}$. Recall that, given a set of polypeptides $\{x_n\}$ with measured helicities, $\{\bar{h}_n\}$, we have

$$\log(f(\theta_i)) \quad = \quad R(\theta_b) + \sum_n -\frac{(\hat{h}_{n,i} - \bar{h}_n)^2}{2\sigma_i^2} - 1/2 \log(2\pi\sigma_i^2)$$

and therefore

$$\begin{aligned} \log(g(\theta_i)) - \log(f(\theta_i)) \quad &= \quad -\frac{(\hat{h}_{a,\theta_i} - \bar{h}_a)^2}{2\sigma_i^2} - 1/2 \log(2\pi\sigma_i^2) \\ \frac{g(\theta_i)}{f(\theta_i)} \quad &= \quad \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(\hat{h}_{a,\theta_i} - \bar{h}_a)^2/(2\sigma_i^2)} \\ &\propto \quad N(\hat{h}_{a,\theta_i}|\bar{h}_a, \sigma_i) \end{aligned}$$

Let w_i be the re-weighting of the sample draw θ_i. Then

$$w_i \quad = \quad \frac{N(\hat{h}_{a,\theta_i}|\bar{h}_a, \sigma_i)}{\sum_{k=1}^N N(\hat{h}_{a,\theta_k}|\bar{h}_a, \sigma_k)} \tag{9.2}$$

Replacing $g(\theta_i)/f(\theta_i)$ in 9.1, we see that

$$\begin{aligned}\hat{\Lambda}_b^* &= \sum_i w_i[(\hat{h}_{b,\theta_i} - \hat{h}_b')^2 + \sigma_i^2] \\ &= \sum_i w_i \left[\left(\hat{h}_{b,\theta_i} - \sum_k w_k \hat{h}_{b,\theta_k} \right)^2 + \sigma_i^2 \right]\end{aligned}$$

Notice from equation 9.2 that the form of w_i implies that any re-weighting from a new helicity measurement, $\bar{h}_a$ should place greater weight on parameter sets, θ_b, that lead to small differences between $\hat{h}_{a,b}$ and $\bar{h}_a$, as expected.

Now we are in a position to decide how much information might be obtained from the study of a particular polypeptide, x_a. We know that an unknown helicity, t, for polypeptide x_a has a distribution of $f_{x_a}(t) = \int_\Theta N(t|\hat{h}_{a,\theta}, \sigma) \cdot f(\theta) d\theta$. Again, if we have a sample $\{\theta_k\}$ from the distribution on our model parameters, $f_{x_a}(t) \approx \frac{1}{N} \sum_k N(t|\hat{h}_{a,\theta_k}, \sigma_k)$. Notice that, if we are interested in the distribution of the error in helicity, then we may generate samples from this distribution easily by sampling first from $\theta \sim f(\theta)$, then sampling $y_j \sim N(t|\hat{h}_{a,\theta}, \sigma)$.

$$\begin{aligned}
\Lambda_b^* &\approx \int_{[0,1]} \sum_i w_i(t) \left[\left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 + \sigma_i^2 \right] f_{x_a}(t)dt \\
&\approx \int_{[0,1]} \sum_i w_i(t) \left[\left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 + \sigma_i^2 \right] \frac{1}{N} \sum_k N(t|\hat{h}_{a,\theta_k}, \sigma_k)dt \\
&\approx \frac{1}{N} \int_{[0,1]} \sum_i N(\hat{h}_{a,\theta_i}|t, \sigma_i) \left[\left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 + \sigma_i^2 \right] dt \\
&= \frac{1}{N} \int_{[0,1]} \sum_i N(\hat{h}_{a,\theta_i}|t, \sigma_i) \left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 dt + \frac{1}{N} \int \sum_i N(\hat{h}_{a,\theta_i}|t, \sigma_i)\sigma_i^2 dt \\
&= \frac{1}{N} \int_{[0,1]} \sum_i N(\hat{h}_{a,\theta_i}|t, \sigma_i) \left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 dt + \frac{1}{N} \sum_i \sigma_i
\end{aligned}$$

Notice that both $\hat{\Lambda}_b$ and $\hat{\Lambda}_b^*$ contain mean($\{\sigma_i\}$). This stems from our assumption that the addition of a single new point will have little affect on the parameters, σ_i. Therefore, our estimator for the change in expected squared error in estimating the helicity of x_b due to studying polypeptide x_a is

$$\hat{\Lambda}_b - \hat{\Lambda}_b^* = \frac{1}{N} \sum_i (\hat{h}_b - \hat{h}_{b,\theta_i})^2 - \frac{1}{N} \int \sum_i N(\hat{h}_{a,\theta_i}|t, \sigma_i) \left(\hat{h}_{b,\theta_i} - \sum_k w_k(t)\hat{h}_{b,\theta_k} \right)^2 dt$$

Suppose that we are studying a subset of $\mathbf{X}$ that is of order M, and that we are interested in choosing from among A possible polypeptides to study. Further define N to be the number of samples from our parameter distribution and T to be the number of steps in the calculation of the integral over t. We must calculate $\sum_k w_k(t)\hat{h}_{b,\theta_k}$ for every combination of these parameters, so brute force calculation takes place in order $M \cdot A \cdot N \cdot T$ time.

9.2 Points with High Potential Information

With the techniques outlined in the previous section, we are now able search for influential points in the study of a particular subset of polypeptides. We have taken a random sample of polypeptides from our data set and used our algorithm to calculate the expected gain in predictive accuracy for each of them assuming that just one of them is studied. Figure 9.1 shows the results of this experiment. Notice that the diagonal is generally darker, which signifies that the greatest improvement in predictive accuracy for a particular polypeptide can most often be had by studying the polypeptide itself. Also, There are blocks visible on the

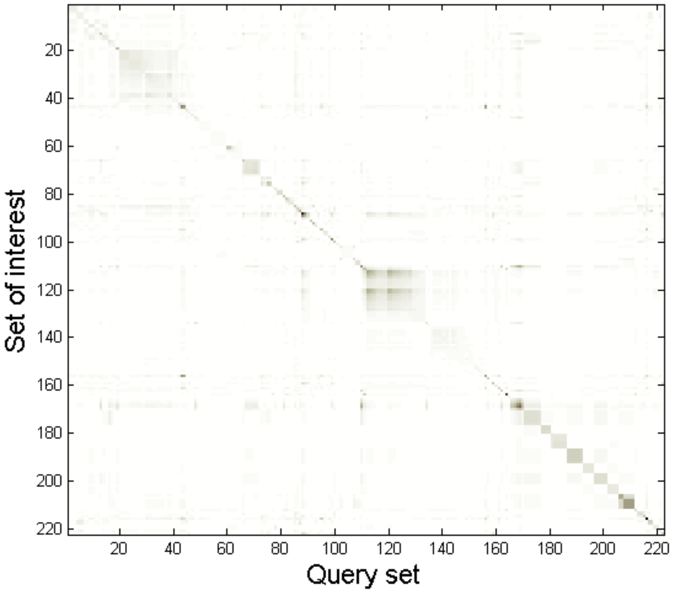


Figure 9.1: Improvement in predictive accuracy of a polypeptide in question (y-axis) from obtaining the helicity of another polypeptide (x-axis). Notice that, not surprisingly, the diagonal is generally darker. This signifies that the greatest improvement in predictive accuracy for a particular polypeptide can most often be had by studying the polypeptide itself. There are blocks visible on the diagonal due to the alphabetical ordering of the polypeptides on the x and y axes.

diagonal due to the alphabetical ordering of the polypeptides on the x and y axes. These were randomly chosen from our data set, and therefore contain subsets of temperature and pH curves. It is not surprising that one can learn about how x_i behaves at a given pH or temperature by studying it at different pH's or temperatures.

9.2.1 Case Study: Influential Points

One of the dark points that is not on the diagonal in figure 9.1 corresponds to the polypeptide 'PANLKALEAQKQKEQR'. We find that the expected change in squared error of helicity prediction from studying this polypeptide is .022, while if we study 'Y(AEAAKA)^{8}F' (where (AEAAKA)^8 means that the sequence 'AEAAKA' is repeated eight times) we get an expected improvement in the prediction of the helicity of 'PANLKALEAQKQKEQR' of over a third of that. Ignoring the temperature parameter, 'PANLKALEAQKQKEQR' has only the individual amino acid parameters and the following i-(i+3) interaction parameters in our model: K-E, E-K, K-R, L-L, and of these, K-E and E-K are repeated seven and eight times respectively in the longer polypeptide.

Figures 9.2, 9.3, 9.4, and 9.5 show the histograms of the interaction parameters from 'PANLKALEAQKQKEQR' that are included in our model for polypeptide helicity. Of the four, the L-L interaction parameter is arguably zero. Thus it is no surprise that the study of a polypeptide with multiple repeats of important parameters would lead to improved understanding of 'PANLKALEAQKQKEQR'.

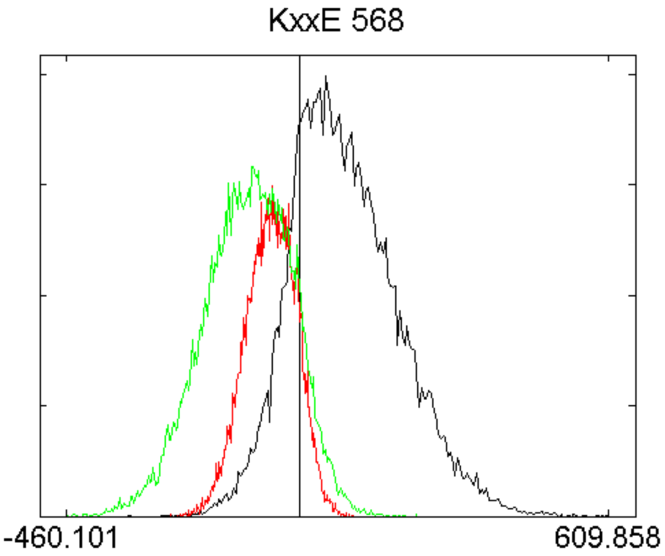


Figure 9.2: A histogram of the K-E interaction parameter at the i to i+3 distance.

9.2.2 Case Study: A Proposed Point Mutation

Suppose that one is particularly interested in polypeptides with low helicity, and is looking for one or some of 20 possible point mutation experiments to perform. From our database of 1187 polypeptides we extracted all with a temperature between 273 and 280 (because all polypeptides have low helicity at high temperature) which also have a helicity $< .3$. This leaves us with 335 polypeptides which we will use as our measure set. We will examine point substitutions at the position labeled ‘X’ in the polypeptide ‘APAELKAAXAAFKRHGPY’ at pH’s of 4, 7, and 10. This gives a “query set” of order 60.

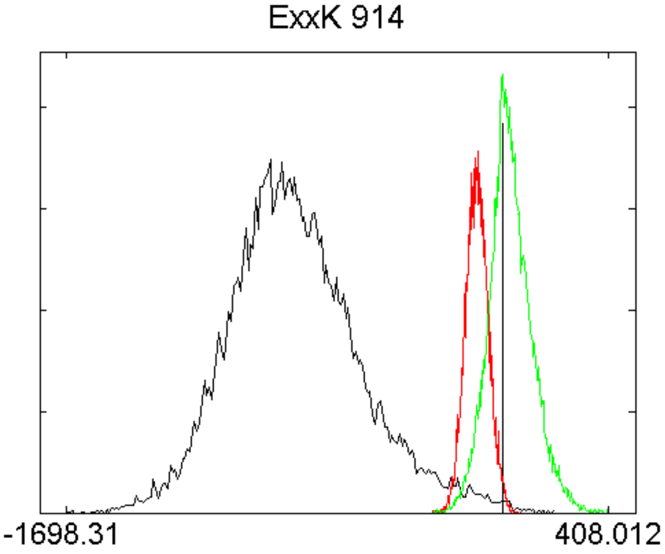


Figure 9.3: A histogram of the E-K interaction parameter at the i to i+3 distance.

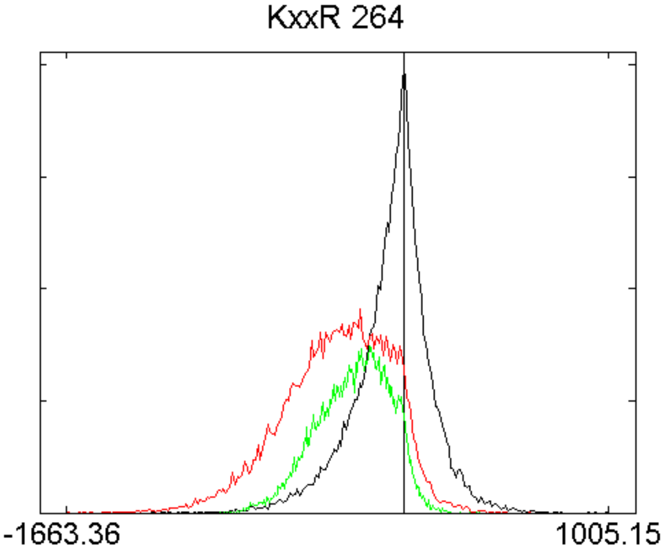


Figure 9.4: A histogram of the K-R interaction parameter at the i to i+3 distance.

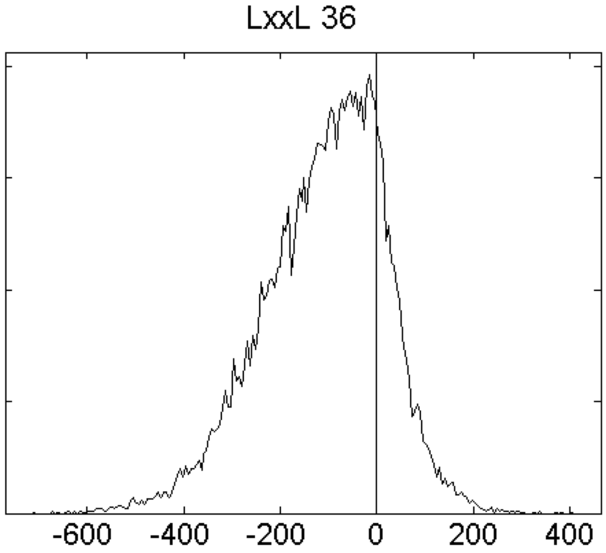
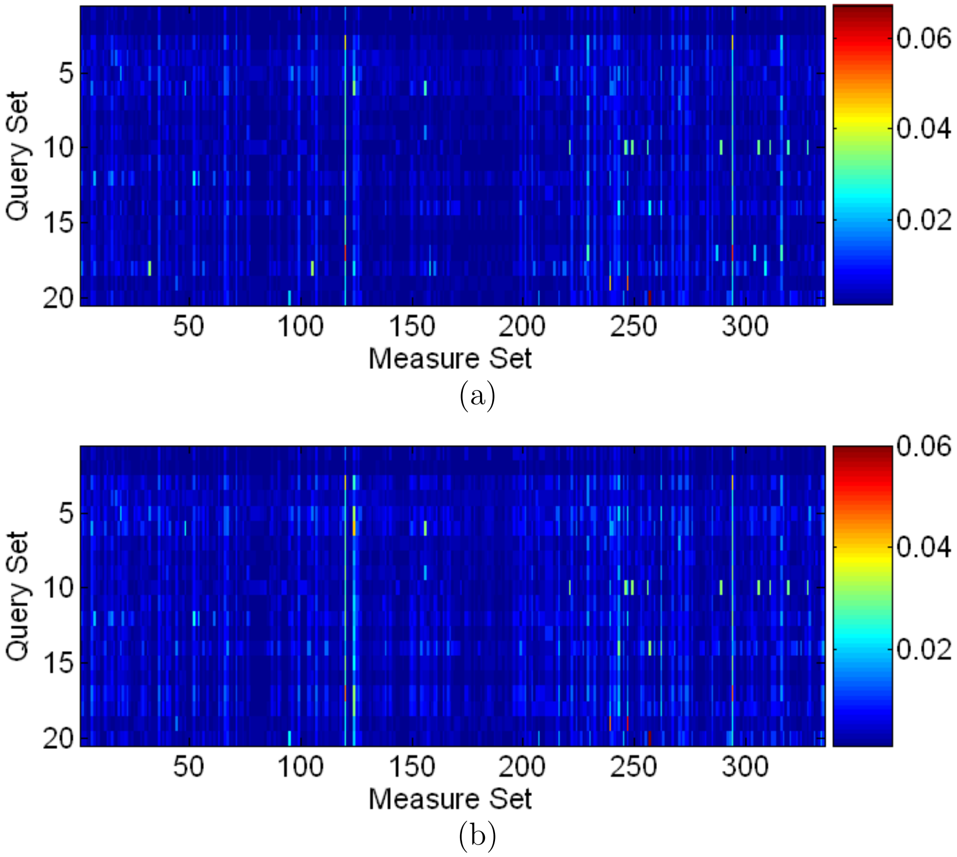


Figure 9.5: A histogram of the L-L interaction parameter at the i to i+3 distance.



(a)

(b)

Figure 9.6: Heatmaps of the expected utility for the polypeptides in the measure set from studying the polypeptides in the query set. Figure (a) shows the results from querying at pH 4 and (b) shows the results from querying at pH 10. There is little difference between figure (a) and the corresponding figure for pH 7, so the pH 7 figure was omitted.

Figure 9.6 shows the expected utility for the polypeptides in the measure set from studying the polypeptides in the query set. It is apparent that at positions 120 and 298 there are two polypeptides in the measure set that are highly related to the query polypeptides. In fact, the query polypeptide was generated from one of the peptides in the measure set, 'APAELKAAEAAFKRHGPY', and this polypeptide is located at position 120 of the measure set. This particular point mutation is at position 17 in the query set, thus the bright red position $(120, 17)$

is not surprising.

The polypeptide at position 298 of the measure set is ‘YGKFRFEQQKKEKEARKK’. There are a number of parameter overlaps between these two polypeptides, including a number of ΔS_r parameters, i to $i+3$ parameters, and i to $i+4$ parameters.

The polypeptide at position 124 of the measure set becomes highly prominent when the query set is studied at pH 10, but is much less so at pH’s 4 and 7. This polypeptide is ‘YGGKAVAAKAVAAKAVAAK’. It contains numerous repeats of the amino acid Lys (K), which has a pK of just over 10. Also, the pK of Tyr (Y), is between 7 and 10. Thus the parameters that become relevant for this polypeptide at higher pH are $\Delta\Delta S_Y$, $\Delta\Delta S_K$, ΔpK_Y and i to $i+3$ for the Y - K interaction. It is difficult to determine which of these parameters are most important for this observed change in utility with pH, but three of these four parameters are also shared by the query polypeptides.

Another interesting feature of these heatmaps is the fact that there is generally more information to be gained by performing this point mutation experiment at higher pH. Of the 20 different point mutations, three quarters provide the most information at pH 10. The root cause of this may be that the pK of Lys is around 10. This means that at a pH of 10, we are effectively querying both $\Delta\Delta S_K$ and ΔS_K parameters. Not surprisingly, the ability to effectively update more parameters will lead to a higher expected utility.

The two that offer the lowest expected gain in utility are ‘G’ and ‘P’. These are at positions 1 and 2 of the query set (see Figure 9.6). Interestingly, these are the two least helical amino acids. Proline in particular is known as a helix breaker. Our measure set was chosen to target lower helicity amino acids, however, placement of either ‘P’ or ‘G’ directly in the center may well produce a polypeptide with a helicity indistinguishable from zero. In this case, it is not surprising that studying these two polypeptides offers such low expected utility.

The two point mutations that offer the largest expected gain in utility are the point mutations ‘I’ and ‘L’, at positions 5 and 6 in the query set. Both of these are hydrophobic and have the potential to interact with ‘F’ in the i to $i+3$ position as well as ‘L’ in the $i-4$ to i position. Figure 9.7 lists the number of polypeptides in the measure set with hydrophobic interactions in these positions relative to ‘F’ and ‘L’. The rows represent all of the five amino acids in the hydrophobic interaction group. Within the measure set, there is a relative abundance of ‘I’ and ‘L’ compared to the other three hydrophobic amino acids. Because of this relative overabundance in the measure set, we expect that these two point mutations will be deemed more important than ‘V’, ‘M’, or ‘F’. Because both of the top two point mutations originate from the set of five hydrophobic amino acids, there is the suggestion that understanding these hydrophobic interaction parameters has the potential to improve our understanding of lower helicity polypeptides.

Interaction	count	interaction	count
$I\cdot\cdot F$	3	$L\cdot\cdot\cdot L$	7
$L\cdot\cdot F$	5	$L\cdot\cdot\cdot I$	8
$V\cdot\cdot F$	1	$L\cdot\cdot\cdot V$	5
$M\cdot\cdot F$	0	$L\cdot\cdot\cdot M$	0
$F\cdot\cdot F$	2	$L\cdot\cdot\cdot F$	2

Figure 9.7: Interaction counts from the measure set. There are substantially more interactions involving ‘I’ and ‘L’ than any of the other hydrophobic amino acids.

This feature of our expected utility measurements demonstrates that the choice of the measure set may be of critical importance in the determination of which polypeptides to single out for future study.

9.3 Summary

We have described a technique for using the results from our model for polypeptide helicity to inform future studies. This will allow a directed approach to choosing polypeptides of interest for biologists, and allow a faster convergence to the model parameters that are best suited for accurate helicity prediction.

The approach to experimental design put forth in this chapter is general, and will work, without modification, for any model with a Gaussian error. Because errors associated with taking measurements of many kinds are assumed to be Gaussian, this approach is appropriate in many applied problems.

Chapter 10

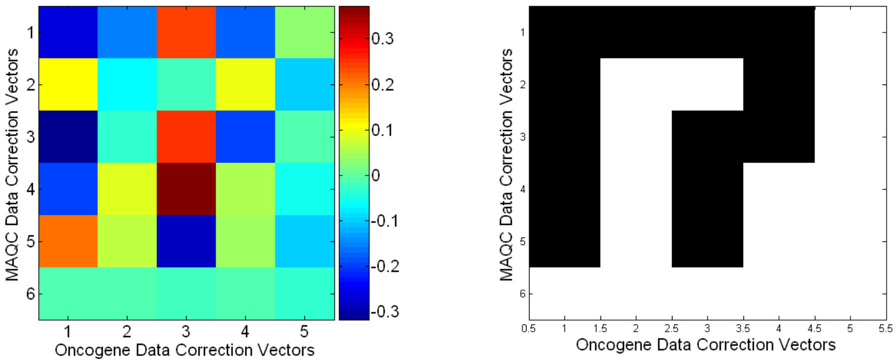
Conclusions and Future Work

We have discussed a number of aspects of variable selection priors as they apply to high dimensional systems. In the general field of hierarchical modeling, we have demonstrated improvements in prior structure that lead to a decrease in false discovery as well as a technique for including sparsity in variance modeling. In the field of gene microarrays, we have demonstrated a method for correcting widespread systematic bias, and in the field of highly multivariate non-linear modeling, we have demonstrated a technique for fitting parameters with point mass priors.

While these are significant advances, there are still many issues to be explored.

10.1 Latent Factors for Microarrays

In our treatment of error correction for microarrays, we have touched on latent factors. However, these factors have dimension equal to the number of observations, and run along the short edge of the data matrix. One of the great hopes for gene microarrays is to bring them into the clinical setting. If this is to ever be achieved, some technique for error correction and signature detection must be developed for a single array.



(a) (b)

Figure 10.1: The dot product of the 2000-dimensional correction vectors from the oncogene and MAQC data sets. Figure (a) shows the actual values of the dot product and Figure (b) is an indicator of whether the dot product is greater than .1. The probability that two randomly generated 2000-dimensional vectors will be greater than .1 is zero to machine precision. The vectors are ordered from left to right (x-axis) and top to bottom (y-axis) according to their importance as correction factors for their respective data sets.

During the processing of the oncogene upregulation data and the MAQC data (see Chapter 4.8 for details of these data sets), we generate correction vectors from the principal components of the housekeeping genes (5 and 6 of them respectively). These are then added/subtracted to the genes under study to correct systematic bias, leading to a separation of the raw data into a true signal and a lab bias signal. Let us consider this lab bias signal. Specifically, let us consider the principal components of this correction signal in the gene (long) direction. We know, for example, that the oncogene corrections were generated from 5 vectors, therefore there are exactly 5 non-zero principal components (each of dimension 10,777 in the case of the oncogene data).

We have supposed that these correction vectors are indeed accounting for changes in a few environmental conditions such as pH and temperature. If this is truly the case, then the hyperplane generated by the five oncogene correction principal components and the hyperplane generated by the six MAQC correction principal components should be close in some sense.

Figure 10.1 shows the pairwise dot products two sets of correction vectors, calculated on a subset of 2000 genes shared in the analysis of each. Recall that the two sets of vectors were generated completely independently of each other. Also, they are ordered according to their importance as correction vectors for their respective data sets. It is straightforward to calculate that the probability of two randomly generated 2000-dimensional unit vectors having an inner product greater than .1 in absolute value is $< 10^{-15}$. Clearly, then, there is strong evidence that these correction vectors are being generated from a much lower dimensional space.

In future work, we will explore latent factors with strong priors (generated from some seed data sets such as these) as a potential mechanism for correcting out systematic error in just a single gene array. Similar latent factors might be designed for important cellular pathways such as the oncogene pathways or the lactic acidosis and hypoxia pathways discussed in earlier chapters.

10.2 Covariance

As discussed in Section 8.5, one of the features of variable selection techniques discussed thus far is the independence of the prior distributions. This structure precludes the use of prior knowledge of dependence structure, and can lead to overestimation of the significance of a particular variable. Consider, for example, Figure 10.2. This shows the posterior for two parameters that are clearly correlated. Not only that, but the density cloud encompasses zero for both variables. The standard variable selection priors will place independent prior masses, p_x and p_y, at zero for each of these variables. The prior mass for both variables being zero is clearly the product $p_x \cdot p_y$. However, when there is prior knowledge that these two variables may correlate, this type of prior is somewhat unsatisfying.

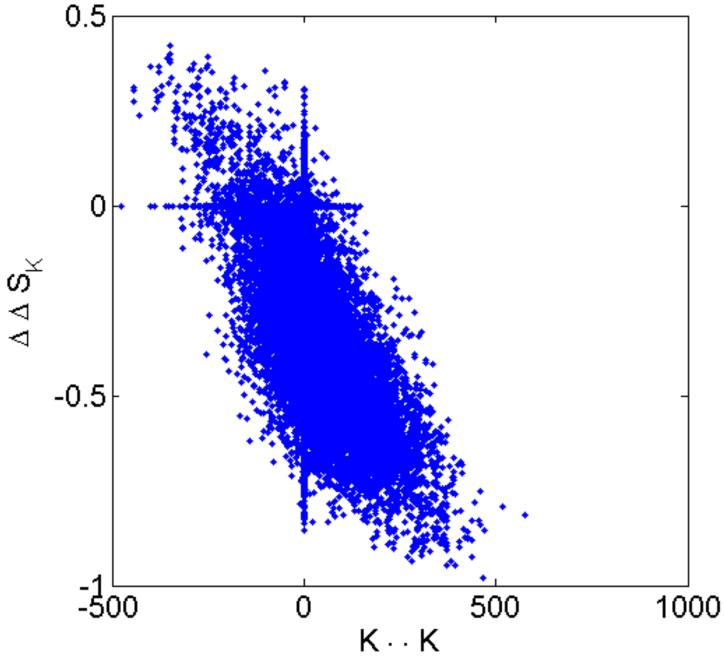


Figure 10.2: The posterior marginal density of two variables, each of which is given a variable selection prior. As the names would suggest, there is prior knowledge that these variables may describe the same phenomenon, thus the independent prior structure is insufficient.

A second type of interaction is shown in Figure 10.3. Of the two, only $\Delta\Delta S_y$ is given a variable selection prior. Again, as suggested by the naming, there is prior knowledge that the two may be highly correlated. Because the point cloud covers the point $\Delta\Delta S_y = 0$, and because of the high correlation, there is stronger evidence for a posterior value of $\Delta\Delta S_y = 0$ than the prior distribution will allow.

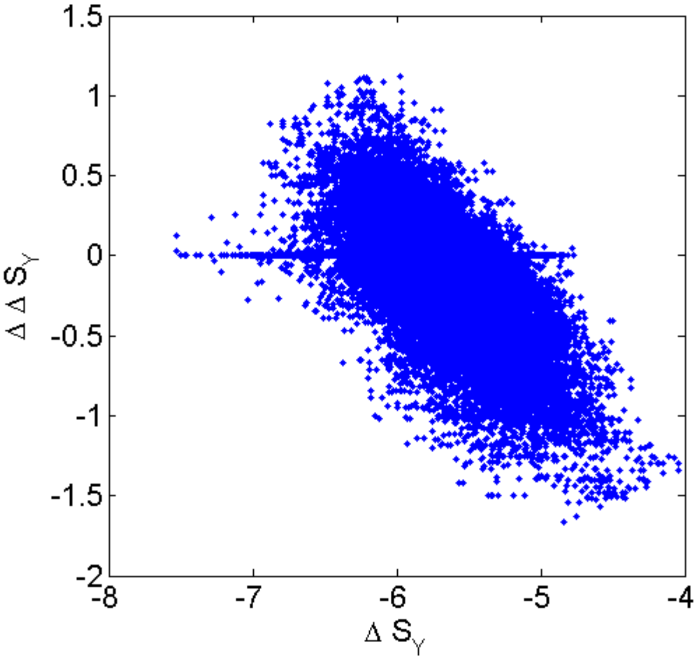


Figure 10.3: The posterior marginal density of two variables from a certain Gibbs energy model. Only $\Delta\Delta S_y$ is given a variable selection prior, but there is prior knowledge that the two may be correlated. The high posterior correlation, and the fact that the cloud covers the point $\Delta\Delta S_y = 0$ lends greater evidence for a zero parameter than our independent prior will allow.

In these examples, it may be that we want to create a hierarchy that allows the inclusion of one variable only when another, more important variable gets far from its believable range. For example, in both Figures 10.2 and 10.3, we know that there is a good chance of correlation between the two variables, and that one should be chosen over the other whenever possible.

In the case of our structural biology problem, we have an informative prior distribution for ΔS_y. We reiterate the prior structure introduced in Section 8.5, for the parameter shown in Figure 10.3.

$$\begin{aligned}\Delta\Delta S_y &\sim \frac{(1-r)N(\Delta S_y|\mu,\sigma)}{(1-r)N(\Delta S_y|\mu,\sigma)+r}\delta_0 + \frac{r}{(1-r)N(\Delta S_y|\mu,\sigma)+r}N(m,v) \\ \Delta S_y &\sim N(\mu,\sigma)\end{aligned}$$

That is to say, the farther the parameter ΔS_y ranges from its prior distribution, the more likely we are to allow the inclusion of the less important variable, $\Delta\Delta S_y$.

In fact, the coupling of these two variables need not be in the form of inserting the prior for one into the prior for the other. In general, any function might be used

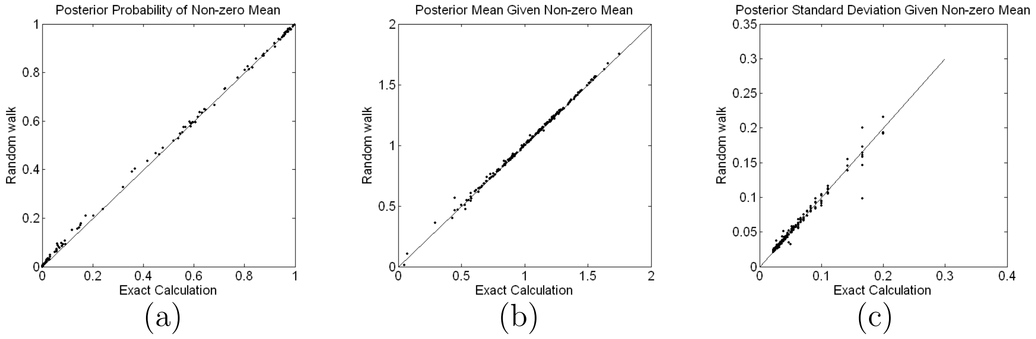
$$\Delta\Delta S_y \sim \frac{(1-r)g(\Delta S_y,\Delta\Delta S_y)}{(1-r)g(\Delta S_y,\Delta\Delta S_y)+r}\delta_0 + \frac{r}{(1-r)g(\Delta S_y,\Delta\Delta S_y)+r}N(m,v)$$

One special case of this type of prior has already been explored by Yuan and Lin (2005) for the purpose of excluding one of two highly colinear explanatory variables in a standard generalized linear models setup. Another use for this type of prior is for the exclusion of the product of two explanatory variables whenever one or the other has been excluded singly. Such a use has the potential for better elucidating posteriors in the case of log-linear (and other) models.

10.3 Random Walks with a Point Mass

The burnin algorithm discussed in Section 7.3 is not a Markov chain. However, after some testing, it is clear that it does well at fitting the posterior probability of inclusion of a variable with a point mass prior. Additionally, when a distribution is poorly approximated with a normal distribution, the ability to use a random walk rather than a blanket proposal distribution produces faster convergence. With this in mind, it is natural to ask whether the algorithm converges at all, and if so, what it converges to.

Figures 10.4a, b, and c show the behavior of the random walk burnin algorithm taken by itself, without any subsequent updating by sampling from a blanket distribution. The toy example of Section 7.4 was repeated 200 times with uniformly generated observation numbers between 5 and 45 and with observation means generated from a N(1,.4) distribution. There is very high agreement between the probabilities, means, and variances generated by the random walk algorithm and those calculated directly.



(a) (b) (c)

Figure 10.4: The behavior of the random walk burnin algorithm from Section 7.3 by itself on 200 randomly generated toy examples (comparable to that described in Section 7.4). Figure (a) compares posterior probability of a change in mean and Figures (b) and (c) compare the poster mean and variance given a non-zero mean.

In some sense, the property of Markov chains that requires restriction of memory to a finite number of previous steps is counterintuitive. There is little or no computational benefit to doing so, and one is potentially throwing away significant amounts of information in the process. The use of this information to produce better and faster estimates of posterior distributions is intriguing and worth exploration.

Bibliography

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle.* Budapest: Akademia Kiado.

Albert, J. H. (1995). Bayesian selection of log-linear models. *Working Paper, Duke University, Institute of Statistics and Decision Sciences* **95-15**.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society* **36**, 99–102.

Atkinson, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *Journal of the Royal Statistical Society A* **165**, 349–373.

Bazarov, A. V., Adachi, S., Li, S. F., Mateyak, M. K., Wei, S., and Sedivy, J. M. (2001). A modest reduction in c-Myc expression has minimal effects on cell growth and apoptosis but dramatically reduces susceptibility to ras and raf transformation. *Cancer Research* **61**, 1178–1186.

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics* **19**, 1, 43–56.

Bernardo, J. (1979). Expected information as expected utility. *Annals of Statistics* **7**, 686–690.

Bild, A., Yao, G., Chang, J., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J., Berchuck, A., Olson, J., Marks, J., Dressman, H., West, M., and Nevins, J. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357.

Bishop, C. M. (1999). Variational principal components. *Artificial Neural Networks* **470**, 509–514.

Box, G. E. P. and Meyer, R. D. (1986). An analysis of unreplicated fractional factorials. *Technometrics* **28**, 1, 11–18.

Broet, P., Richardson, S., and Radvanyi, F. (2002). Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *Journal of Computational Biology* **9**, 671–683.

Carrero, P., Okamoto, K., Coumailleau, P., O'Brien, S., Tanaka, H., and Poellinger, L. (2004). Redox-regulated recruitment of the transcriptional coactivators creb-binding protein and src-1 to hypoxia-inducible factor 1α. *Molecular and Cellular Biology* **20**, 1, 402–415.

Chakrabartty, A., Kortemme, T., and Baldwin, R. L. (1994). Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing sidechain interactions. *Protein Science* **3**, 843–852.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10**, 3, 273–304.

Chen, M. and Dey, D. K. (2003). Variable selection for multivariate logistic regression models. *Journal of Statistical Planning and Inference* **111**, 37–55.

Clyde, M. and DeSimone-Sasinowska, H. (1997). Accounting for model uncertainty in Poisson regression models: Particulate matter and mortality in Birmingham, Alabama (1997). *Technical Report, Duke University, Institute of Statistics and Decision Sciences* .

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

Coffey, T., Gennings, C., Simmons, J. E., and Herr, D. W. (2005). D-optimal experimental designs to test for departure from additivity in a fixed-ration mixture ray. *Toxicological Sciences* **88**, 2, 467–476.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties.* W. H. Freeman and Company, 2^{nd} edn.

Cui, W. and George, E. I. (2004). Empirical Bayes vs. fully Bayes variable selection. *Tech. rep., The Wharton School, University of Pennsylvania* .

Delmar, P., Robin, S., Roux, D. T. L., and Daudin, J. J. (2005). Mixture model on the variance for the differential analysis of gene expression data. *Applied Statistics* , 1, 31–50.

Dixon, J. W. and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association* **43**, 109–126.

Do, K. A., P., M., and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society, Ser. C (Applied Statistics)* **54**, 627–644.

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics* **in press**.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403–417.

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

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Econometrics* **100**, 381–427.

Foster, D. P. and George, E. L. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**, 1947–1975.

Freeman, P. R. (1970). Optimal Bayesian sequential estimation of the median effective dose. *Biometrika* **59**, 1, 79–89.

Froimowitz, M. and Fasman, G. D. (1974). Prediction of the secondary structure of proteins using the helix-coil transition theory. *Macromolecules* **7**, 5, 583.

Fujiwara, M., Nagy, Z. K., Chew, J. W., and Braatz, R. D. (2005). First-principles and direct design approaches for the control of pharmaceutical cyrstallization. *Journal of Process Control* **15**, 493–504.

George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* 1304–1308.

George, E. I. and Foster, D. P. (2000). Calibration and Empirical Bayes variable selection. *Biometrika* **87**, 731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection and Gibbs sampling. *Journal of the American Statistical Association* **88**, 423, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. *Bayesian Statistics 5* 609–620.

H., R. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22**, 400–407.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **to appear**.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: data mining, inference and prediction.* Springer.

Hu, Y. C., Lam, K. Y., Law, S., Wong, J., and Srivastava, G. (2001). Identification of differentially expressed genes in esophageal squamous cell carcinoma (ESCC) by cDNA expression array. *Clinical Cancer Research* **7**, 2213–2221.

Ibrahim, J. G., Chen, M., and Ryan, L. M. (2000). Bayesian variable selection for time series count data. *Statistica Sinica* **10**, 971–987.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix genechip probe level data. *Nucleic Acids Research* **31**, 4e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 2, 249–264.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* **98**, 438–455.

Ishwaran, H. and Rao, J. S. (2005). Spike and slabe variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730–773.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32**, 4, 1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics* **33**, 4, 1700–1752.

Jung, M. S., Yun, J., Chae, H. D., Kim, J. M., Kim, S. C., Choi, T. S., and Shin, D. Y. (2001). p53 and its homologues, p63 and p73, induce a replicative senescence through inactivation of nf-y transcription factor. *Oncogene* **20**, 5818–5825.

Karra, R., Wemullapalli, S., Dong, C., Herderick, E. E., Song, X., Slosek, K., Nevins, J. R., West, M., Goldschmidt-Clermont, P. J., and Seo, D. (2005). Molecular evidence for arterial repair in atheroslcerosis. *Proceedings of the National Academy of Sciences* **102.46**, 16789–16794.

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to Schwarz criterion. *Journal of the American Statistical Association* **90**, 928–934.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kohn, R., Shively, T. S., and S., W. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the Ameriacan Statistical Association* **94**.

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, J., Brunet, J. P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S., and Golub, T. R. (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 5795, 1929–1935.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2005). Mixtures of g-priors for Bayesian variable selection .

Linja, M. J., Porkka, K. P., Kang, Z., Savinainen, K. J., Janne, O. A., Tammela, T. L. J., Vessella, R. L., Palvimo, J. J., and Visakorpi, T. (2004). Expression of androgen receptor coregulators in prostate cancer. *Clinical Cancer Research* **10**, 1032–1040.

Lucas, J. E., Carvalho, C., Wang, Q., Bild, A., Nevins, J., and West, M. (2005). Sparse statistical modelling in gene expression. *Discussion papers, Duke University Institute of Statistics and Decision Sciences* .

Miller, L., Smeds, J., George, J., Bega, V., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E., and Bergh, J. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciencies* **102**, 38, 13550–13555.

Misra, G. P. and Wong, C. F. (1998). Predicting helical segments in proteins by a helix-coil transition theory with parameters derived from a structural database of proteins. *Proteins: Structure, Function, and Genetics* **28**, 3, 344–359.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 404, 1023–1032.

Morrish, F., Giedt, C., and Hockenbery, D. (2003). c-myc apoptotic function is mediated by nrf-1 target genes. *Genes and Development* **17**, 2, 240–255.

Munoz, V. and Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nature Structural and Molecular Biology* **1**, 399–409.

Munoz, V. and Serrano, L. (1995a). Elucidating the folding problem of helical peptides using empirical parameters. ii. helix macrodipole effects and rational modification of the helical content of natural peptides. *Journal of Molecular Biology* **245**, 275–296.

Munoz, V. and Serrano, L. (1995b). Elucidating the folding problem of helical peptides using empirical parameters. iii. temperature and pH dependence. *Journal of Molecular Biology* **245**, 297–308.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31**, 103–115.

Qian, H. and Chan, S. I. (1996). Interactions between a helical residue and tertiary structures: helix propensities in small peptides and in native proteins. *Journal of Molecular Biology* **261**, 2, 279–288.

Schmaltz, C., Hardenbergh, P., Wells, A., and Fisher, D. (1998). Regulation of proliferation-survival decisions during tumor cell hypoxia. *Molecular and Cellular Biology* **18**, 5, 2845–2854.

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

Scott, J. and Berger, J. O. (2005). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference* .

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., Goodsaid, F. M., Herman, D., Jensen, R. V., Johnson, C. D., Lobenhofer, E. K., Puri, R. K., Scherf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P. K., Zhang, L., Amur, S., Bao, W., Barbacioru, C. C., Lucas, A. B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X. M., Cebula, T. A., Chen, J. J., Cheng, J., Chu, T. M., Chudin, E., Corson, J., Corton, J. C., Croner, L. J., Davies, C., Davison, T. S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A. C., Fan, X. H., Fang, H., Fulmer-Smentek, S., Fuscoe, J. C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P. K., Han, J., Han, T., Harbottle, H. C., Harris, S. C., Hatchwell, E., Hauser, C. A., Hester, S., Hong, H., Hurban, P., Jackson, S. A., Ji, H., Knight, C. R., Kuo, W. P., LeClerc, J. E., Levy, S., Li, Q. Z., Liu, C., Liu, Y., Lombardi, M. J., Ma, Y., Magnuson, S. R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M. S., Osborn, T. W., Papallo, A., Patterson, T. A., Perkins, R. G., Peters, E. H., Peterson, R., Philips, K. L., Pine, P. S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B. A., Samaha, R. R., Schena, M., Schroth, G. P., Shchegrova, S., Smith, D. D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K. L., Tikhonova, I., Turpaz, Y., Vallanat, B., Van, C., Walker, S. J., Wang, S. J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., and Slikker, W. J. (2006). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 9, 1151–1161.

Shippy, R., Fulmer-Smentek, S., Jensen, R. V., Jones, W. D., Wolber, P. K., Johnson, C. D., Pine, P. S., Boysen, C., Guo, X., Chudin, E., Sun, Y. A., Willey, J. C., Thierry-Mieg, J., Thierry-Mieg, D., Setterquist, R. A., Wilson, M., Lucas, A. B., Novoradovskaya, N., Papallo, A., Turpaz, Y., Baker, S. C., Warrington, J. A., Shi, L., and Herman, D. (2006). Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nature Biotechnology* **24**, 9, 1123–1131.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**, 1, 267–288.

Tung, W. S., Lee, J. K., and W., T. R. (2001). Simultaneous analysis of 1176 gene products in normal human aorta and abdominal aortic aneurysms using a membrane-based complementary dna expression array. *Journal of Vascular Surgery* **34**, 1, 143–150.

Verotta, D. (1990). An application of the d-optimal criterion to define the experimental design for a particular class of semi-parametric models. *Computer Methods and Programs in Biomedicine* **33**, 3, 181–187.

Viallefont, V., Raftery, A. E., and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* **20**, 3215–3230.

Walenta, S. and Mueller-Klieser, W. F. (2004). Lactate: Mirror and motor of tumor malignancy. *Seminars in Radiation Oncology* **14**, 3, 267–274.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics 7* .

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A. J., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences, USA* **98**, 11462–11467.

Wolfe, P. J., Godsill, S. J., and Ng, W. J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *Journal of the Royal Statistical Society, Series B* **66**, 3, 575–589.

Wynn, H. (1970). The sequential generation of d-optimum experimental designs. *Annals of Mathematical Statistics* **41**, 1655–1664.

Yuan, M. and Lin, Y. (2005). Efficient Empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association* **100**, 472, 1215–1225.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia, Spain* 585–603.

Zhu, L., van den Heuvel, S., Fattney, A., Ewen, M., Livingston, D., Dyson, N., and Harlow, E. (1993). Inhibition of cell proliferation by p107, a relative of the retinoblastoma protein. *Genes and Development* **7**, 1111–1125.

Biography

Joseph Edward Lucas was born in Indianapolis, Indiana on August 18, 1972. He graduated from Brown University with a Bachelor of Science in Applied Math/Biology in May of 1994. He met his wife two years later, and was married in October of 1998. Following some work in physiology and then computing, he received a Master of Science degree in Computer Science in May of 2002 and a Master of Science in Mathematics in August of 2002, both from the University of Pennsylvania. It was during this time that his first child, Kiran Ela Lucas, was born. His second, Tal Singh Lucas was born in 2005, just after he received his third Master of Science degree, this time in Statistics from Duke University.