DNA Microarray Data Analysis and Regression Modeling for Genetic Expression Profiling

Mike West,
Joseph R Nevins, Jeffrey R Marks, Rainer Spang, Carrie Blanchette & Harry Zuzan
Duke University, Durham NC 27708-0251
[1]

[1] Mike West is the Arts and Sciences Professor of Statistics and Decision Sciences, and Director of the Institute of Statistics and Decision Sciences, at Duke University. The corresponding address is ISDS, Duke University, Durham, NC 27708-0251, USA; the web site is http://www.stat.duke.edu. Joe Nevins is James B Duke Professor and Chair of the Department of Genetics, in the School of Medicine at Duke University, and an investigator of the Howard Hughes Medical Institute. Jeff Marks is Associate Professor of Surgery, in the School of Medicine at Duke University. Carrie Blanchette is PhD candidate in Cell and Molecular Biology at Duke University. Rainer Spang and Harry Zuzan are Visiting Assistant Professors at Duke University, and NISS Research Fellows, and were partially supported by NSF grant DMS-9711365. The research reported here is part of a collaboration of the Center for Bioinformatics and Computational Biology, the Center for Genome Technology and the Comprehensive Cancer Center at Duke University. We are grateful to Sandrine Dudoit of MSRI for providing precise details of the algorithms used to pre-process the leukemia array expression data.

Abstract

We report on our studies in large-scale gene expression profiling using DNA microarray data. The problem of molecular phenotyping – linking observed genetic expression profiles to identified physiological or clinical states and outcomes – is one of simply critical importance for improved understanding of disease progression and for improved therapies. Our main application here is in breast cancer, where interest lies in identifying characteristics of genetic expression, involving possibly very many genes, that are useful in predictive discrimination between cancer states. In this applied setting, we frame the problem as one of *predictive discrimination*, and approach its solution in a binary regression modeling framework. Breast cancers are clinically or histologically identified into one of two outcome groups, and analysis aims to produce binary regression models for outcomes based on observed genetic expression data measured via high-density DNA microarrays using RNA extracted from the tumors. The formal statistical problem is ill-posed, since tumor sample sizes are substantially smaller than the number of available and potentially interesting explanatory variables – i.e., we are in the *Large p, Small n* paradigm. We address this in a Bayesian framework using singular-value regression ideas and novel classes of informative and structured prior distributions for the very high-dimensional regression parameters. The singular-value decomposition analysis of design matrices of expression measures is also valuable in exploratory analysis of large-scale expression array data sets. In the context of our breast cancer study we discuss the methodology, implementation of our Bayesian analysis, aspects of model specification, aspects of posterior and predictive analysis, model assessment and validation issues. We develop detailed studies of the breast data in connection with estrogen receptor status as the defined outcome of interest, and highlight significant scientific findings as well as aspects of predictive discrimination performance using the regression approach. We also discuss similar issues in a benchmark leukemia study.

Keywords: Bayesian bioinformatics, Bayesian regression analysis, binary regression, breast cancer, estrogen receptor status, gene expression profiles, DNA microarrays, molecular phenotyping, singular value decompositions

1 Introduction

The emergence of DNA microarray technologies for measuring large-scale gene expression levels has generated substantial interest in characterizing and classifying biological samples based on their genetic expression profiles as measured via microarray methods. The scale and complexity of the gene expression profile offers the opportunity to develop extremely precise and subtle molecular characterizations of identifiable biological states. Our work reported here is focussed in breast cancer studies, where there is massive potential for improved understanding of disease progression and for improved therapies based on linkage of such genomic data sources to defined outcomes. The primary issues that drive any investigation into breast cancer – and any other cancer – also provide the motivation for our current studies. These primary scientific questions are simple yet are still largely unanswered: What are the genetic changes associated with the onset of breast cancer? Why do some cancers metastasize while other similar looking malignancies remain confined to the breast? What parameters of the tumor define the response to therapy (hormonal, radiation, chemotherapy, or biological therapies)? Categorization of breast tumors with respect to their biological properties, morphologic appearance, and expression of certain genes provides the current standard for classification to guide the clinician in prescribing treatment and judging prognosis. The current study aims to extend the available resources to include the genomic component – linking in the parallel expression of multiple genes, possibly thousands, as part of the basis for classification. Our program targets sets of breast cancers that represent some of the major molecular and biological categories of the disease to develop methods for analysis and determine the utility of gene expression arrays. Among the outcomes under study are hormone receptor status of cancers, including the central estrogen and progesterone receptor status (ER/PR), metastatic status of auxiliary lymph nodes at the time of diagnosis, primary versus metastatic disease status, and response of cancers to adjuvant chemotherapy. Our report here concerns the ER/PR status study. We note that expression array data generation produces data resources that may be brought to bear on a range of analyses – the expression data represents a snap-shot of the genetic profile of the tumor whose characteristics may be linked to several or many different defined clinical outcomes or physiological states in repeat analyses.

The estrogen receptor is a genetic transcription factor known to control the expression of a set of genes, and that, when evident at increased expression levels, is implicated in breast cancer conditions. Estrogen receptor status of breast cancer tumors is routinely assessed by immunohistochemical methods which actually measure the estrogen and progesterone receptor proteins in the tumor. Therapeutic regime decisions rest critically on a determination of ER/PR status, a binary classification into ER+, i.e., tumors identified as having non-negligible levels of the estrogen protein and progesterone proteins, or ER−, i.e., tumors with undetectable levels of the proteins. In connecting large-scale expression data to ER status, we have two complementary goals. The first is predictive diagnosis/discrimination of status, i.e., to use gene expression profiles to predict the status. The second broad goal is to generate information about genes – individually and in possibly large subsets – whose expression patterns relates to ER status, in order to guide related and further studies of potential oncogenes and tumor suppressor genes. Following identification of tumor tissue, RNA is extracted to generate cDNA and then RNA probes for hybridization to DNA microarrays. The analysis uses the Affymetrix Human GeneFL arrays and system, so providing profiles of sequences corresponding to 6800 human genes (the arrays actually have 7129 sequences, some of which represent control sequences.) Following hybridization, the microarrays are scanned and processed using the Affymetrix GeneChip$^{\mathrm{TM}}$ system software, producing a range of outputs that include measures of expression levels of these genes in the RNA from the tumor sample. The expression summary we take as our predictor here is the Affymetrix "average log ratio" measure, a dimensionless quantity that estimates, for each of the gene sequences, the level of abundance of the gene in the tumor tissue sample, where abundance is measured relative to a benchmark level of random cross-hybridization in the RNA preparation. In other work, we are exploring a range of questions related to data extraction and data quality when working with this array technology, including definitions of alternative measures of expression. Follow-on studies aim to explore how the results discussed in our analyses of the ER+/− problem, and others, might be modified using alternative measures. For the purposes of this paper, we use the single measure. The result of the experimental analysis is therefore a set of $p = 7129$ measured expression levels, on a standardized and dimensionless scale, from each microarray, that represents the measured genetic expression profile for the corresponding tumor. As it turns out, data quality issues mean that 3 of the microarrays are removed from the analysis, leaving a total of $n = 27$ arrays split as 15 ER+ and 12 ER−. The statistical issues are to explore, understand and characterize these profiles, and to incorporate them as covariates in binary regression models for ER status as outcomes.

We note that most of the recent literature dealing with statistical, or other bioinformatics methods for exploring microarray expression data have focussed on exploratory methods for clustering and pattern recognition. In studies where a clinical or physiological condition is of interest, such studies typically apply one or more clustering methods to organize genes into subgroups, and then consider how the clusters match up with outcomes. In contrast, we address the inherent predictive discrimination problem head-on, and are, as far as we know, the first group to take this formal formal statistical perspective. One of the major, and substantial benefits of taking this view is that we are in no way concerned with the daunting prospect of modeling the very high-dimensional distributions of expression levels. Further, by adopting a formal probability model we aim to produce full probabilistic inferences on regression parameters and predictive, out-of-sample classification probabilities for future tumors based on their expression profiles. By the way, our analysis strategy does generate useful and informative exploratory tools and insights about relationships among, and "clusters" and subgroups of, genes that may guide further analysis.

Our formal statistical problem falls into the *Large p, Small n* paradigm – we have a regression problem with a small number n of observations and an extremely large number p of available explanatory variables. This context is not specific to the expression profiling application – similar issues are faced in dealing with regression models with large sets of higher-order interactions between predictor variables, and in problems in which the predictors are discretized versions of continuous functions or values of processes, such as spectral profiles or times series, for example. In the Bayesian framework, proper prior distributions for regression parameters provide, in principle, a path to solution of the ill-posed statistical problem, though it is trite to simply state that proper priors solve the problem. The dangers of simply proceeding to perform Bayesian analysis with "imprecise" proper priors in even moderate dimensional parameter spaces can be quite dramatic, and so the development of structured, informative priors is critically needed. Our analysis is based on a class of priors introduced in West (2000). These address the modeling and analysis challenges in approaches that:

- Utilize singular-value decompositions of matrices of measured values of large numbers of predictors across samples, generating factor representations and possibly massive dimension reduction to summary *factors* of use in exploratory analyses. In the expression profiling context, the factors are referred to as *supergenes* for obvious reasons.

- Introduce classes of novel prior distributions for large regression parameters to reflect the dependence and singularity structure evident in likelihood functions based on large numbers of predictors, and that utilize the singular-value structure of the design matrices to induce a potentially massive reduction in the parameter space relevant to posterior computation.

- Develop easily implemented and standard MCMC methods for binary regression models to produce posterior inferences on the high-dimensional regression parameter, and consequent evaluation of out-of-sample predictive utility in probabilistic classification of new cases.

Section 2 describes the formal binary regression framework, and briefly summarizes the form and relevance of our new structured prior distributions, aspects of model fitting and computation, and posterior and predictive distributions. More mathematical detail and development of the prior and posterior distributions is given in full in West (2000). The section goes on to discuss practically important issues of prior specification that respect a desire for a prior predictive distributions to represent appropriately uninformative initial views about ER status outcomes. Section 3 then provides a detailed discussion of analysis of the ER+/− data set, highlighting significant scientific findings as well as aspects of predictive discrimination performance using the regression approach. Section 4 provides a short summary of analysis of a benchmark expression data set from a published leukemia discrimination study, to make comparisons and further highlight questions of screening genes to select subsets for use in predictive discrimination in validation samples.

2 Bayesian binary regression when p >> n

2.1 Model form, latent structure and SVD

Write $z_1, \dots, z_n$ for the binary outcomes, so that $z_i = 1(0)$ if tumor sample i is ER+(−). We use probit regression in which the probability that $z_i = 1$ is modeled as $\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ where: $\mathbf{x}_i$ is the vector of p predictor values for case i; $\boldsymbol{\beta}$ is the $p-$vector regression parameter; and $\Phi(\cdot)$ is the standard normal distribution function. In the cases of interest here, p is very large so that $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are high-dimensional vectors. Our application has $n = 27$ and $p = 7129$.

We exploit the standard latent variable construction of this probit model in which $z_i = 1$ if, and only if, $y_i \geq 0$, where the latent quantity y_i is defined by $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$ with independent standard normal variates ϵ_i, $\epsilon_i \sim N(\epsilon_i|0, 1)$. One reason for using this representation is that Bayesian analysis of the binary regression model may be routinely implemented using MCMC methods that incorporate the inherent latent variables y_i as missing data to be imputed and inferred along with the parameter vector $\boldsymbol{\beta}$ (Albert and Chib 1993; Johnson and Albert 1999, chapter 3). The latent variable structure further leads to a theoretical representation in terms of an underlying linear model that we use in developing singular-value regression analysis, as follows.

In vector-matrix notation, we have the linear model

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{I})$$

where: $\mathbf{I}$ is the $n \times n$ identity matrix; $\mathbf{y}$ is the vector of n latent variable $y_i, (i = 1, \dots, n)$; $\mathbf{X}$ is the $p \times n$ matrix whose columns are the $\mathbf{x}_i$, so that the rows represent variables and the columns represent samples; and $\boldsymbol{\epsilon}$ is the $n-$vector of independent standard normal deviates $\epsilon_i, (i = 1, \dots, n)$.

Note that $\mathbf{X}$ is *tall and skinny* in the cases of interest; in the ER status context $\mathbf{X}$ is 7129×27. The linear predictor vector $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\beta}$ lies at the heart of the regression model. Standard singular-value decomposition (SVD) analysis evidences the implicit singular-value regression form of the linear predictor. Begin with the standard SVD of $\mathbf{X}$, namely

$$\mathbf{X} = \mathbf{ADF}$$

where: $\mathbf{A}$ is the $p \times n$ *singular value factor loadings* matrix, and has orthonormal columns; $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$, the diagonal matrix of positive *singular values,* ordered as $d_1 \geq d_2 \geq \cdots \geq d_n \geq 0$; and $\mathbf{F}$ is the $n \times n$ orthogonal *SVD factor* matrix. Write $\mathbf{f}_i'$ for row i of $\mathbf{F}$; this represent the values of the i^{th} SVD factor across all n microarrays, and is of course a linear combination of the original p predictor variables. The linear combinations are simply defined (up to scale factors in $\mathbf{D}$) by the columns of $\mathbf{A}$, i.e., $\mathbf{F} = \mathbf{D}^{-1}\mathbf{A}'\mathbf{X}$. Assume that the singular values are all positive, i.e., $d_n > 0$, so that $\mathbf{A}$ (and $\mathbf{X}$) are of full rank n. This loses no generality since, otherwise, we would simply reduce from n to whatever the rank implied by the number of non-zero singular values.

The SVD reduces the linear predictor $\boldsymbol{\mu}$ to the standard singular value regression form $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\beta} = \mathbf{F}'\mathbf{D}\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$ is the implied $n-$vector of regression parameters for the factor variables weighted by the singular values in $\mathbf{D}$. This represents a possibly massive dimension reduction from p to n parameters. We stress that there is no information loss or approximation here; the dimension reduction is inherent in the large p, small n context; the orthogonal factor variables describe and sufficiently summarize the patterns of variation among the original variables as far as the latent linear regression is concerned. In the expression profiling framework, the factors are linear combinations of gene specific expression levels, and are referred to as *supergene* factors, or simply *supergenes*, for that reason. The regression model on expression levels of p genes reduces to that on the $n << p$ supergenes derived from the SVD analysis of the expression matrix.

2.2 Likelihood and priors

Based on the binary data $\mathbf{z}$, the resulting likelihood function for $\boldsymbol{\gamma}$ is simply $\prod \pi_i^{z_i}(1-\pi_i)^{1-z_i}$ where

$$\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}) = \Phi(\mathbf{f}_i'\mathbf{D}\boldsymbol{\gamma}).$$

With n observations and n effective predictors, we have an ill-posed problem with no standard frequentist or likelihood solution. Take a simple example with $\mathbf{F} = \mathbf{I}$, so that the linear predictors are simply scaled values of the individual γ_i. As a result, each π_i is a one-one transform of a single γ_i; resulting maximum likelihood values are at $\pm\infty$, and the likelihood functions are flat apart from in a region close to the origin. As a result, purely likelihood-based inference is untenable. The same issues arise in the general model, and point to the imperative for Bayesian approaches with informative priors. Further, priors for $\boldsymbol{\gamma}$ must be derived from underlying priors for $\boldsymbol{\beta}$ in order that inferences be available for the original $p >> n$ predictor variables.

Our approach utilizes a new class of structured prior distributions for $\boldsymbol{\beta}$ detailed in West (2000). The invention of this class of priors was motivated wholly by this genetic expression profiling problem, though they will obviously have application in any regression problem in which $p >> n$. A prior from this class is termed a *generalized $g-$prior,* or $gg-$prior, as the prior family extends and generalizes the class of standard $g-$priors (Zellner 1986). In current, parallel work, the theoretical underpinnings of these priors is being further developed. For analysis here, a prior in this class has the form

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{T}) \propto \exp\{-\boldsymbol{\beta}'\mathbf{A}\mathbf{D}\mathbf{T}^{-1}\mathbf{D}\mathbf{A}'\boldsymbol{\beta})/2\},$$

i.e., a singular normal prior $N(\boldsymbol{\beta}|\mathbf{0},\mathbf{B})$ with prior mean $\mathbf{0}$ and singular precision matrix $\mathbf{B}^- = \mathbf{A}\mathbf{D}\mathbf{T}^{-1}\mathbf{D}\mathbf{A}'$ for some $p \times p$ matrix of variances $\mathbf{T} = \text{diag}(\tau_1^2,\ldots,\tau_n^2)$. The standard generalized inverse gives the posterior variance matrix $\mathbf{B} = \mathbf{A}\mathbf{D}^{-1}\mathbf{T}\mathbf{D}^{-1}\mathbf{A}'$.

To motivate this prior form, note that a standard $g-$prior arises when $\tau_i = g$, a constant for each i, whereupon $\mathbf{B} = g\mathbf{X}\mathbf{X}'$. Then the prior shares the "shape" and collinearity structure of the likelihood function, but allows for greater dispersion through values of the precision g. The major interest in these priors is due to the fact that we can rather easily specify diffuse forms ($g << 1$) but – at least in expectation – avoid problems of significant prior:likelihood conflict by having the prior "oriented" appropriately through the rotations and scalings implied by the $\mathbf{X}\mathbf{X}'$ matrix as prior precision matrix. The new $gg-$prior extends this idea by replacing $\mathbf{X}\mathbf{X}' = (\mathbf{A}\mathbf{D})(\mathbf{A}\mathbf{D})'$ with $(\mathbf{A}\mathbf{D})\mathbf{T}^{-1}(\mathbf{A}\mathbf{D})'$. That is, we "insert" a diagonal matrix of p scale factors τ_i "inside" the precision matrix, so permitting different scales in each of the orthogonal factor directions. As discussed below, we treat the τ_i as hyperparameters to be inferred in the analysis.

Note that the prior can be generalized to admit non-zero prior means for the $\boldsymbol{\beta}$. For our expression analysis application we require prior means zero in order that the prior predictions be unbiased with respect to binary outcomes. Note that prior mean zero implies that the priors for the classification probabilities π_i are centered at, and symmetric about, 0.5, so naturally incorporating the notion of an unbiased initial position. It is important to note that, due to the singularity structure of the prior, this would also be true if we used a prior for $\boldsymbol{\beta}$ with any non-zero prior mean, say $\mathbf{b}$, such that $\mathbf{A}'\mathbf{b} = \mathbf{0}$. Any such prior embodies the notion that one or more of the predictors has an expected non-zero effect, however, and so is precluded by specifying $\mathbf{b} = \mathbf{0}$. This is particularly relevant in the scientific study of expression profiles, where we aim to elucidate the relationships between genes and outcomes without anticipating any directional relationship.

Under the above prior for $\boldsymbol{\beta}$, the implied prior for $\boldsymbol{\gamma} = \mathbf{A}'\boldsymbol{\beta}$ is simply

$$p(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} N(\gamma_i|0, \tau_i^2/d_i^2).$$

That is, the form of the $gg-$prior coupled with the inherent orthogonality of the SVD factor regression leads to independent, conjugate priors for the γ_i. The precisions d_j^2 are modified by conditional variances τ_i^2, to be determined. Note that it is critically important that we allow for major differences among the γ_i to allow for possibly quite substantial variations in the effects of the factors, so the τ_i^2 must vary. Below we address this by specifying appropriate hyperpriors on these variances.

2.3 Posterior distributions

In Bayesian analysis of the binary regression using MCMC methods, posterior simulations of the regression parameters $\boldsymbol{\gamma}$ flow from posteriors under the linear model based on iteratively updated imputations of the latent variates y_i. Hence posterior distributions for $\boldsymbol{\beta}$ and/or $\boldsymbol{\gamma}$ are those conditional on any imputed value of $\mathbf{y}$. We have two equivalent sampling models:

$$(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \sim N(\mathbf{y}|\mathbf{X}'\boldsymbol{\beta}) \quad \text{and} \quad (\mathbf{y}|\mathbf{X}, \gamma) \sim N(\mathbf{y}|\mathbf{F}'\mathbf{D}\boldsymbol{\gamma}).$$

We deal first with posteriors for $\boldsymbol{\gamma}$, since operational methodology is implemented in terms of $\boldsymbol{\gamma}$.

Working in terms of $\boldsymbol{\gamma}$, the likelihood function reduces to

$$p(\mathbf{y}|\mathbf{X}, \gamma) \propto \prod_{i=1}^{n} \exp\{-(\gamma_i - \hat{\gamma}_i)^2 d_i^2/2\}$$

where $\hat{\gamma}_i$ is the i^{th} element of the least-squares vector $\hat{\boldsymbol{\gamma}} = \mathbf{D}^{-1}\mathbf{F}\mathbf{y}$. The singular values d_i enter here through the precisions d_i^2 of the estimates $\hat{\gamma}_i$. The prior for $\boldsymbol{\gamma}$ is therefore conjugate and leads to independent posteriors

$$p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}, \mathbf{T}) = \prod_{i=1}^{n} N(\gamma_i|g_i^*, G_i^2/d_i^2)$$

where $g_i^* = G_i^2\hat{\gamma}_i$ and $G_i^2 = \tau_i^2/(1+\tau_i^2)$.

Working in terms of $\boldsymbol{\beta}$, the prior is also conditionally conjugate to the likelihood $p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta})$ and the posterior reduces to

$$p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\mathbf{T}) \propto \exp\{-(\boldsymbol{\beta}-\mathbf{b}^*)'\mathbf{ADG}^{-1}\mathbf{DA}'(\boldsymbol{\beta}-\mathbf{b}^*)/2\}$$

where $\mathbf{b}^* = \mathbf{Ag}^*$, $\mathbf{g}^* = (g_1^*,\ldots,g_n^*)'$ and $\mathbf{G} = \mathrm{diag}(G_1^2,\ldots,G_n^2)$. This is a singular normal posterior with precision matrix $\mathbf{ADG}^{-1}\mathbf{DA}'$; the standard generalized inverse gives the posterior variance matrix $\mathbf{AD}^{-1}\mathbf{GD}^{-1}\mathbf{A}'$.

Note that, due to the singularity structure, the posterior for $\boldsymbol{\beta}$ is not uniquely defined (though that for $\boldsymbol{\gamma}$ is). The above posterior density is unchanged under location shifts of $\mathbf{b}^*$ to $\mathbf{b}^*+\mathbf{b}$ for any $p-$vector $\mathbf{b}$ such that $\mathbf{A}'\mathbf{b} = \mathbf{0}$. As shown in West (2000), however, this apparent identification problem is resolved, and the above posterior unique, under the specification of a fixed prior mean, here of $\mathbf{0}$ as earlier discussed. The identification of $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X},\mathbf{T})$ above, with posterior mean $\mathbf{b}^*$ as specified, corresponds precisely to the use of *minimum norm* generalized inverses in deriving least-squares estimates in under-determined linear systems. The constraint to a prior mean of zero for $\boldsymbol{\beta}$ in fact determines $\mathbf{b}^*$ as a corresponding Bayesian *minimum norm* posterior mean.

Finally note the operational and computational implications. The posterior parameters for $\boldsymbol{\beta}$ are trivially computed from those $\boldsymbol{\gamma}$ simply by reversing the map from $\boldsymbol{\beta}$ to $\boldsymbol{\gamma}$, i.e., mapping back to $p >> n$ dimensions via $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$. Practically, the implication is that simulation-based posterior computations may be performed in the low, $n-$dimensional $\boldsymbol{\gamma}$ space, and that resulting posterior simulations for $\boldsymbol{\beta}$ are trivially derived by applying the transformation.

2.4 Further aspects of prior specification

To complete the prior specification requires that we specify a prior distribution for the elements of the prior variance matrix $\mathbf{T}$. It should be clear that, in this binary regression context, the data $\mathbf{z}$ provides only limited information about the values. For example, the zero-mean prior for $\boldsymbol{\beta}$ induces a prior for each of the latent variates y_i that is a zero-mean normal with variance depending on $\mathbf{T}$. Hence, integrating over $\boldsymbol{\beta}$ implies $Pr(z_i = 1|\mathbf{X},\mathbf{T}) = 0.5$ whatever $\mathbf{T}$ may be; each of the z_i marginally will provide no information about $\mathbf{T}$. Their values as a sample are of course informative, though the information may be very limited. We do need to allow for variations in the τ_i values across factors i, consistent with the expectation that a small number of the γ_i may be large and others small. Hence careful specification of informative proper priors for the τ_i^2 is called for.

Consider again the marginal prior distributions for each of the n classification probabilities $\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$ as a result of the prior $p(\boldsymbol{\beta}|\mathbf{X},\mathbf{T})$. The linear predictor $\mathbf{x}_i'\boldsymbol{\beta} = \mathbf{f}_i'\mathbf{D}\boldsymbol{\gamma}$ has a zero-mean normal prior with a variance that reduces to $\mathbf{f}_i'\mathbf{T}\mathbf{f}_i$ where $\mathbf{f}_i'$ is the i^{th} row of the factor matrix $\mathbf{F}$, i.e., the vector of n sample values of the i^{th} factor variable. It is a trivial observation that the choice $\mathbf{T} = \mathbf{I}$ implies, as a result of the orthogonality of $\mathbf{F}$, that $\mathbf{f}_i'\mathbf{T}\mathbf{f}_i = 1$. This in turn trivially implies that the marginal prior distributions of each of the classification probabilities π_i are standard uniform. Importantly, that this applies across all cases, whatever the $\mathbf{x}_i$ are, so providing an initial and appealing neutral prior specification. However, it is critical to allow for possibly quite significant variability among the τ_i^2 hyperparameters in order to provide flexibility in adapting to differences in the factor effects γ_i. Hence the need for priors on the τ_i^2. We treat the τ_i^2 exchangeably, adopting a common prior for each. Any such prior represents uncertainty about their values and, as a result, introduces more uncertainty into the implied common prior for each of the π_i. Following the above discussion, we anchor the choice of prior for τ_i^2 around the value 1; i.e., we specify independent priors on each of the τ_i^2 centered at 1 but allowing deviation from that benchmark value. Conditional conjugacy suggests inverse gamma priors; treating the τ_i^{-2} as a random sample from a gamma prior with a moderate degrees of freedom and mean 1 induces priors on the π_i that are still symmetric about 0.5, but that now have heavier tails than uniform, being more U-shaped with increased mass near the extremes of 0 and 1. Priors of this form arise, in fact, as standard reference or uninformative priors in Bayesian analysis; the canonical reference prior has the U-shaped density function proportional to $\pi_i^{-1/2}(1-\pi_i)^{-1/2}$, i.e., the $Beta(\pi_i|1/2, 1/2)$ prior. Via our construction we induce closely similar priors on π_i by taking the prior $\tau_i^{-2} \sim Gamma(\tau_i^{-2}|k/2, k/2)$ for low degree of freedom k, say $k = 2$ or 3. Repeat analyses with ranges of such priors indicate that, on the data analyzed here, posterior inferences are generally insensitive to changes in the prior so long as it is not too diffuse. In analysis of the binary regression data, it is clear that the data provides limited information on the τ_i^2, so in studying posterior distributions a key focus lies in assessment of how the U-shaped priors for each of the π_i are altered to posteriors.

2.5 Model fitting via MCMC

A key practical implication of the above posterior structure is that MCMC based computation may be carried out in the $n-$dimensional parameter space for $\boldsymbol{\gamma}$, rather than the possibly very much larger $p-$dimensional space for $\boldsymbol{\beta}$. An algorithm designed to produce posterior sample values for $\boldsymbol{\gamma}$ leads immediately to posterior samples for $\boldsymbol{\beta}$ by transformation $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$. The framework is a variant of the standard MCMC analysis of a binary regression model (Albert and Chib 1993; Johnson and Albert 1999, chapter 3) and the sequence of conditional distributions to iteratively simulate is detailed below. Implementation of the simulation is straightforward, and MCMC convergence is generally clean and routine. The iterations are initialized at arbitrary values for the latent normal variates $\mathbf{y}$ and variances $\mathbf{T}$, and then run to eventually generate values that, at each iteration, represent approximate samples from the full joint posterior $p(\mathbf{y}, \boldsymbol{\gamma}, \mathbf{T}|\mathbf{z})$ hence $p(\mathbf{y}, \boldsymbol{\beta}, \tau|\mathbf{z})$ on transforming the $\boldsymbol{\gamma}$ samples to $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$. The components of each iteration are as follows.

- Conditional on previous values of $\mathbf{y}, \mathbf{T}$ draw a new vector $\boldsymbol{\gamma}$ from the conditional posterior $p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}, \mathbf{T})$. Directly compute the corresponding $\boldsymbol{\beta} = \mathbf{A}\boldsymbol{\gamma}$ to provide a draw from the conditional posterior $p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \mathbf{T})$.
- Given the current $\boldsymbol{\gamma}$ vector, draw a new diagonal $\mathbf{T}$ matrix of values from the n independent gamma posteriors $Ga(\tau_i^{-2}|(k+1)/2, (k+d_i^2\gamma_i^2)/2)$.

- Given this $\boldsymbol{\gamma}$ vector, compute the linear predictor vector $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\beta}$ via the computational efficient formula $\boldsymbol{\mu} = \mathbf{F}'\mathbf{D}\boldsymbol{\gamma}$; the i^{th} element of this $n-$vector is then just $\mu_i = \mathbf{x}_i'\boldsymbol{\beta} = \mathbf{f}_i'\mathbf{D}\boldsymbol{\gamma}$ at the current $\boldsymbol{\beta}$ value. For $i = 1, \ldots, n$, sample new values of the latent y_i variates independently from their implied truncated normal posteriors. In detail, set

  $$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \Phi^{-1}[z_i P_i + u_i(z_i + (1 - 2z_i)P_i)]$$

  where the u_i are independent $U(0,1)$ variates and $P_i = \Phi(-\mu_i)$.

2.6 Kernel regression and machine learning

In the last few years there has been a major growth of research in optimization algorithms for regression and classification at the interfaces of statistics and the machine learning communities in computer science. Ranges of developments in kernel methods, including support vector machine (SVM) methods (Schölkopf *et al* 1999), are receiving publicity in applications in gene expression profiling (e.g., Brown *et al* 1999) so it is germane and important to make the connections here. Many kernel-based algorithms, including varieties of SVMs, have approximate interpretation as optimizers (such as Bayes' classifiers in discrimination problems, or posterior modes as point-wise estimates of regression functions) in statistical models. It is far beyond the scope here to develop details, but it is worth highlighting the inherent kernel structure and connections in the standard linear/binary regression context here.

In our notation, $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\beta}$ is the $n-$vector of means of the latent normal variates $\mathbf{y}$, and the latent linear model states that $(\mathbf{y}|\mathbf{X}, \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \mathbf{I})$. Standard Bayesian non-parametric regression approaches based on Gaussian process priors (e.g., the recent overview in Neal 1999) relax the assumed linearity of the regression by treating $\boldsymbol{\mu}$ as a realization of n values on an uncertain regression function, and describe uncertainty about the function using a Gaussian process prior. Maintaining the zero mean assumption, such a model is specified as follows:

- Write $\mu(\mathbf{x})$ for the univariate function value at any point $\mathbf{x}$ in the covariate space, so that $\mu_i = \mu(\mathbf{x}_i)$ are the values corresponding to the observed predictors;
- Specify a marginal variance function $v(\mathbf{x})$ and suppose that, for any $\mathbf{x}$, the marginal prior for $\mu(\mathbf{x})$ is $N(\mu(\mathbf{x})|0, v(\mathbf{x}))$;
- Assume that any set of values $\mu(\mathbf{x}), \mu(\mathbf{x}^*), \ldots,$ have a joint normal prior with defining covariances $r(\mathbf{x}, \mathbf{x}^*)$ between $\mu(\mathbf{x})$ and $\mu(\mathbf{x}^*)$, and such that $v(\mathbf{x}) = k(\mathbf{x}, \mathbf{x})$.

Under such a specification, the $n-$vector $\boldsymbol{\mu}$ has a normal prior $(\boldsymbol{\mu}|\mathbf{X}) \sim N(\boldsymbol{\mu}|\mathbf{0}, \mathbf{K})$ where $\mathbf{K}$ is the $n \times n$ variance matrix with elements $k(\mathbf{x}_i, \mathbf{x}_j)$. We will not develop analysis details further; suffice to say that the normal prior is conjugate to the conditional normal likelihood arising from $(\mathbf{y}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \mathbf{I})$, so that the MCMC analysis in the binary regression model is trivially modified. Predictive distributions for new cases are derived by utilizing the $(n+1)-$dimensional normal prior distributions for $\boldsymbol{\mu}$ and $\mu(\mathbf{x}_{n+1})$ jointly, for any future predictor vector $\mathbf{x}_{n+1}$. Applying such a model requires the initial specification of the *covariance kernel function* $k(\mathbf{x}, \mathbf{x}^*)$, representing expected dependence between function values at two different predictor values, and also interpretable as a measure of distance between the two predictor values. This provides the connection with kernel-SVM machine learning algorithms, that are also based on such a kernel function. The specification is essentially arbitrary, modulo the requirement that the implied matrix $\mathbf{K}$ be a valid variance matrix for all n and $\mathbf{X}$.

The key connection to be made here is with the linear regression $\boldsymbol{\mu} = \mathbf{X}'\boldsymbol{\beta}$ and assuming a normal prior $N(\boldsymbol{\beta}|\mathbf{0}, \mathbf{B})$ on the regression vector, where $\mathbf{B}$ may be singular or non-singular. On integrating over $\boldsymbol{\beta}$ this implies the normal prior $N(\boldsymbol{\mu}|\mathbf{0}, \mathbf{X}'\mathbf{B}\mathbf{X})$, i.e., of the above kernel form with $\mathbf{K} = \mathbf{X}'\mathbf{B}\mathbf{X}$. Under the stylized $gg-$prior for $\boldsymbol{\beta}$ we have $\mathbf{B} = \mathbf{A}\mathbf{D}^{-1}\mathbf{T}\mathbf{D}^{-1}\mathbf{A}'$ and $\mathbf{K}$ reduces to $\mathbf{K} = \mathbf{F}'\mathbf{T}\mathbf{F}$. Thus the kernel functions $k(\mathbf{x}_i, \mathbf{x}_j)$ are weighted inner-product measures of distance between covariate vectors represented in the factor space, and with weights defined by variances in $\mathbf{T}$. Developing these connections with kernel methods further is obviously of interest. For the purposes of this paper, however, we simply explore the structure of the implied kernel matrix $\mathbf{K} = \mathbf{F}'\mathbf{T}\mathbf{F}$ below in the main example. Since $\mathbf{K}$ depends on the hyperparameters $\mathbf{T}$, it is the posterior distribution and derived point estimates of elements of $\mathbf{K}$ that are of interest.

3 Gene expression profiling of breast cancers

3.1 Context and data summaries

As introduced earlier, our interest lies in expression profiling of from tissues from individual breast tumors. Our small data set comprises RNA samples from tissues of 27 tumors, which was used to generate RNA probes for hybridization to DNA oligonucleotide microarrays (Lockhart *et al* 1996). Tissue from tumor specimens comes from the frozen tissue bank of the Duke Breast SPORE tissue resource. The data set here originated as a set of 30 cancers selected with 15 ER+ and 15 ER−, and with a careful match of tumor size and other relevant physical characteristics. Receptor status was determined by immuno-histochemical staining and image analysis, performed by the Image Cytometry Laboratory, in the Pathology Department at Duke. Following tissue selection, RNA was extracted from tumor samples to generate cDNA and then RNA probes for hybridization to DNA microarrays. The analysis uses Affymetrix Human GeneFL arrays, currently standard high-density oligonucleotide microarrays that contain sequences corresponding to 6800 human genes (see `http://www.affymetrix.com/`). The arrays actually have 7129 sequences, some of which represent control sequences. There are significant technological and experimental issues raised in the use of these arrays, including questions of data extraction, quality, summarization and the choice of form and definition of expression measures used for further analysis (Schadt *et al* 2000). For our purposes here, we take the Affymetrix *average log ratio* measure of gene expression as comprising the predictor information $\mathbf{X}$. This is a standardized, dimensionless measure of expression of each gene that is directly comparable across arrays and that, in our view, provides a relatively robust indicator of differential expression. For each array, the Affymetrix GeneChip$^{\text{TM}}$ system software computes this summary measure for each gene. In follow-on studies, we are exploring the effects on the binary regression analysis results of using different summary measures, and expect to report on this in the near future.

The expression data matrix $\mathbf{X}$ is 7129×27, providing expression measures for each of the $p = 7129$ identified nucleotide sequences on each of the microarrays. Arrays $\#1 - 15$ inclusive were determined as ER+, and arrays $\#16 - 27$ as ER$-$. As noted above, the original experiment produced 30 arrays, 15 of each status, Of these, two either failed to adequately hybridize and so were discarded. The determination of ER status for the third is subject to further investigation, and, since a single case has a marked influence on the analysis with such small sample sizes, this array has also been withdrawn from the data set for the purposes of this illustration.

Figures 1 and 2 provides some examples of expression measures by gene, displaying the measured expression levels of 16 chosen genes across all 27 microarrays. The genes chosen here are those that are eventually identified as among a much larger number with practically significant regression parameters, though here they serve only to exemplify the data. Figure 1 displays expression measures for genes that eventually turn out to be positively related to ER+ status, i.e., whose expression levels tend to be elevated in the ER+ cases. the data. Figure 2 displays genes that eventually turn out to be negatively related to ER+ status, i.e., whose expression levels tend to be elevated in the ER$-$ cases. Each figure displays bar graphs colored red for ER+ cases and blue for ER$-$. The set includes the estrogen receptor gene (#5524), obviously expected to be of interest, and the ps2 protein gene (#5782) that is known to be estrogen regulated; that is, increased levels of expression of the estrogen gene are generally expected to lead to increased levels of expression of ps2.

SVD analysis of the expression array $\mathbf{X}$ leads to $n = 27$ supergene factors and corresponding singular values d_i such that d_i^2 is a measure of "importance" of the i^{th} supergene in describing the linear correlation structure among the genes. Figure 3 provides display of the elements d_j^2 as a percentage of the total $\sum_{i=1}^{27} d_i^2$. The upper frame indicates the dominance of the first factor, and the lower frame – which omits d_1^2 – shows that several factors contribute to similar levels in describing the linear dependence structure. Figure 4 plots the first four supergene factors by array, clearly showing that the first (blue line) represents average expression levels by array, and that the second (green line) discriminates between ER+ and ER$-$ arrays with the exception of array #16, a case to be further discussed. The relevance of factors 2 and 4 together is evident in the lower right frame of Figure 5 where the 27 arrays are plotted on selected pairs of factors, with arrays identified by number and colour coded to indicate ER status.

3.2 Regression model analysis and summary

We first summarize the results from analysis of all 27 arrays. The only determining prior specification is that for the τ_i^{-2}, and here we take $k = 2$ in the gamma priors as a default specification. The MCMC analysis was run for 50,000 iteration after a burn-in of 200, and the summaries displayed are based on subsamples every tenth iteration to produce posterior samples of size 5,000. Figure 6 plots posterior means for the regression parameter vector $\boldsymbol{\gamma}$ by factor number, and corresponding posterior means for the gene regression parameter vector $\boldsymbol{\beta}$ by gene number. Factor 2 clearly has the largest supergene coefficient (in absolute value). In Figure 7 we display, in the right-hand columns, histogram representations of marginal posterior densities for γ_1, γ_2 and γ_4, to give an indication of the uncertainty associated with the point estimates in Figure 6. Figure 8 provides similar histograms for the three most significant elements of $\boldsymbol{\beta}$, i.e., the coefficients of the top 3 of the many genes implicated in the predictive discrimination of ER status. These marginal densities are each accompanied, in the left hand columns, by MCMC trace plots that are consistent with the view that convergence is generally clean and satisfactory. Of the 7129 elements β_j, in roughly 100 cases does zero lie below the 0.0005% point of the posterior, or above the 99.95% point; these can be therefore viewed as genes of relevance in the predictive discrimination (one-sided significance level of 0.05%). Moving to 0.5% and 99.5% points (0.5% level) we capture a total of about 380 interesting genes, and moving up to the 1% and 99% points we capture a total of about 580 genes. Figure 9 provides an image plot of 750 (of the 7129) rows of the factor loadings matrix $\mathbf{A}$, selecting those rows corresponding to the 750 genes with largest absolute values of the estimates of β_j. The rows (genes) are displayed from the top down in order of decreasing values of the β_j. Note the second column, corresponding to loading of genes on the second supergene (factor). The first 400+ rows score negatively on this supergene (light yellow), the remaining 300+ score positively (darker orange). This indicates a broad classification of these genes into two groups according to this supergene factor. Note also the fourth column, where the last 100 or so rows are clearly distinguished in terms of lighter image intensity, corresponding to a subsidiary categorization by the fourth supergene factor.

Table 1 lists the identifiers of the top 25 genes, chosen by ranking the absolute values of posterior means for the β_j. The first group have positive estimated coefficients, the second negative. Hence increased levels of expression of genes in the first group favor ER+ over ER−, while this is reversed for the second group. Those favoring ER+ are generally dominant, this group including the absolute top gene identified, not surprisingly, as the estrogen receptor gene; others are the liv-1 gene and the ps2 protein gene, both also known to be regulated by the estrogen receptor.

We turn now to classification based on the regression model. The posterior simulation samples for $\boldsymbol{\gamma}$ provide, by direct calculation, corresponding posterior samples for the linear predictors $\mu_i = \mathbf{x}_i'\boldsymbol{\beta} = \mathbf{f}_i'\mathbf{D}\boldsymbol{\gamma}$, and hence for the classification probabilities $\pi_i = \Phi(\mu_i)$, for each of the arrays. These samples may be summarized to provide inferences on the linear predictors and *fitted classification probabilities* based on this analysis of the full data set. Figure 10 displays some simple summaries. In the upper frame, the posterior means of the linear predictors μ_i are plotted by array; the lower frame plots the posterior means of the π_i against posterior means of the μ_i. The arrays are labeled by array number and colour coded – red for ER+ and blue for ER−. Taking $\pi_0 = 0.5$ ($\mu_i = 0$) as a hard decision boundary, it is evident that the arrays are perfectly classified on this basis. However, this is nothing more than a comfort measure, as these are fitted values rather than out-of-sample, or *cross-validation* predictions of ER status. Before we proceed to this evaluation, note that the

ER− array #16, with a fitted probability of ER− of about 0.60, is less surely classified than the rest. This is a case to be discussed further.

As a technical aside, note that an ad-hoc approximate analysis might, in choosing point estimates of classification probabilities, adopt the plug-in estimates obtained by evaluating $\pi_i = \Phi(\mu_i)$ at the posterior mean of μ_i. In contrast, our summaries here use the more appropriate posterior means of the π_i. It is easily shown that the plug-in estimates will always be more extreme than the displayed posterior means, being closer to 0 or 1, and so provide an illusory impression of higher accuracy in predictive fit. The posterior means of the π_i are, by comparison, always closer to the 0.5 boundary. The difference arises from the fact that the plug-in approach completely ignores uncertainty in predictions due to the inherent posterior uncertainty about $\boldsymbol{\beta}$. This uncertainty is formally and appropriately captured in the posterior means of the π_i, leading to the more conservative values.

3.3 Honest prediction: Cross-validation analyses

We now summarize the results of cross-validation, or honest predictive analyses. For each case $\#i = 1, 2, \ldots, 27$, we delete the value of the binary outcome z_i from the data set, treating it as a missing value. The analysis is trivially extended to include z_i as an unknown, simply simulating its value conditional on the currently imputed $\boldsymbol{\gamma}$ at each of the MCMC iterates. This way we deliver simulation samples for this z_i as well as all the other parameters, including the classification probabilities $\pi_i = \Phi(\mu_i)$ for this hold-out case. These samples for μ_i, π_i and z_i represent the *posterior predictive distribution* for case $\#i$ conditional on the rest of the data. These are the summary distributions to use in what honest prediction, or cross-validation, as compared to the fitted values from the analysis based on all the data. This is repeated for $i = 1, \ldots, 27$ producing separate honest predictions for each array. Figure 11 displays plots of the resulting predictive means of the μ_i and π_i in format similar to those of the fitted case in Figure 10.

The first point of note is that the estimated linear predictors, and hence classification probabilities, correctly classify the arrays in all cases but one, array #16, based on a simple threshold at $\pi_0 = 0.5$ ($\mu_i = 0$). A second evident feature is that the values are generally less extreme than in the fitted model; again this is expected and sensible in out-of-sample prediction. For example, the fitted probabilities of ER+ status for arrays #3 (which is ER+) and #25 (which is ER−) are 0.98 and 0.13, respectively; the corresponding cross-validation values are 0.96 and 0.31, respectively – shrunk in somewhat from the more extreme fitted values, so representing more conservative predictions. Array 16 is something of an anomaly. Before investigating that further, however, we illustrate and emphasize second-level uncertainties in the predictive analysis – a most critical aspect that adds measurably to both the classification problem and to interpretation of the analysis. It is also a feature often notably absent from non-statistical classification methods and algorithms.

Each of the estimates of the μ_i and π_i in the figures so far discussed is a posterior mean from a full posterior distribution represented by the computed simulation samples. Here we have posterior samples of 5,000 draws for each of the π_i, in both the full data analysis and in each of the individual cross-validation analyses. Figure 12 presents histogram representations of these posteriors for the cases of arrays #3, #16 and #25 for illustration. The left-hand columns of three histograms displays the fitted posterior distributions based on the full data analysis; the right-hand columns displays the corresponding predictive distributions from the cross-validation analyses for these three cases. In studying these posteriors, recall that the underlying prior in each case is the reference form with a U-shaped density, symmetric about 1/2. The posteriors represent how the data has modified these priors in each case.

Consider first array #3. In the full data analysis, the posterior mean of π_3 (the fitted probability) is 0.98; the posterior (first row, left-hand frame of Figure 12) is massed on very high values, with almost no posterior probability on values lower that 0.8. This ER+ array is clearly correctly classified, and there is almost no uncertainty about the value of π_3. In the cross-validation analysis (right hand frame), the out-of-sample predictive distribution for π_3 is also massed on values very close to 1, and that it does indeed naturally reflect more uncertainty about the value, this is not really noticeable in the graph; the mean is 0.94, slightly lower than the fitted value, again consistent with slightly increased uncertainty and a resulting conservatism about the classification. Nevertheless, this distributions indicates that π_3 is almost surely above 0.8, so correctly predicting the ER status of array #3. This case is quite typical of most of arrays; the ER− arrays have, of course, posterior and predictive distributions massed near 0 rather than near 1.

Consider now case #25. Here the posterior from the full data analysis strongly supports values of π_{25} close to zero, with very little mass above 0.5 and a mean of about 0.13. This ER− array is correctly classified as a result. The out-of-sample predictive distribution for π_{25} is obviously more diffuse, being spread out over a broader range of values and representing a much higher degree of uncertainty about the probability. Evidently, there are features in the $\mathbf{x}_{25}$ expression vector that, compared to $\mathbf{x}_3$, induce this additional uncertainty about array #25. The mass is nevertheless still concentrated on values lower than 0.5 and the mean is 0.31, so the prediction is still correct in terms of hard classification.

Case #16 is quite different, and the most interesting of the set of arrays. Here the posterior from the full data analysis is quite diffuse, spread out almost uniformly across the probability scale, though with a small concentration of mass and mode near zero. This reflects a high degree of uncertainty about π_{16}. The mean is about 0.40, so a simple hard classification correctly identifies this array as ER−, but the graph clearly raises questions about the validity of this hard classification based on this full data analysis. Moving to the out-of-sample predictive distribution for π_{16} we see quite a different picture. The density is massed on high values, with almost no mass below 0.5 and a mean of 0.91. That is, based on the analysis of the data excluding this array, the predictive most strongly indicates ER+ status, contradicting the histochemical classification of ER−.

3.4 Exploring specific genes

Figure 1 provides an initial insight as to why array #16 appears anomalous. Recall that this figure provides bar graphs of the actual expression levels for all arrays for the top 8 genes – those genes with largest, positive values of posterior means of β_j from the full data analysis (reading across rows in order of decreasing value of the β_j). The bars are colour coded, red for ER+ cases and blue for ER−. As mentioned earlier, the top gene is the estrogen receptor gene (#5524), the second is intestinal-trefoil-factor-mRNA (#1734), and the third is the ps2 protein gene (#5782). From the graph we note that the estrogen receptor gene has a negligible level of expression on array #16, perfectly consistent with its classification as ER−. A further gene, encoding the liv-1 protein, ranks ninth in the list of top genes and is known to be estrogen regulated; this genes is also, appropriately, low on array #16. The second ranked intestinal-trefoil-factor is elevated, however, more consistent with ER+ than ER− status. The third ranked ps2 protein gene, also known to be regulated by the estrogen receptor gene, is low in absolute value but notice that it is nevertheless present on array #16 but apparently absent on all other arrays. On each of the remaining group of genes displayed array #16 has elevated values, more consistent with ER+ status than ER−. Cumulatively, these patterns of elevated expression of sets of genes on this array make a real difference in the out-of-sample predictions, where the entire expression profile of this array strongly argues for ER+ status. The fact that this tumor sample was classified as ER− means that the fitted data analysis has difficulty in dealing with the apparent conflict between the observed expression profile $\mathbf{x}_{16}$ and the nominal outcome $z_{16} = 0$, and results in the diffuse posterior distribution for π_{16} based on the full data set analysis as discussed above.

On the biological side, this raises potentially interesting questions about gene regulation. The ps2 protein gene, for example, is known to be regulated by the estrogen receptor gene, leading to the expectation of coincidentally elevated values of expression. In this case of array #16, ps2 is elevated to a degree, being positive compared to the consistently negative values on all other arrays. This anomalous case may therefore suggest the existence of alternative mechanisms of regulation of ps2 and other genes that may be explored in follow-up studies. It also confirms the view that patterns of conjoint expression of possibly large numbers of genes, rather than the variation in specific, individual candidate genes (i.e., the estrogen gene alone) provide the basis for confirmatory discrimination.

As a final point on this theme, note that array #11 has low expression levels on all of the top 8 genes. In this respect this case appears more consistent with ER− than the histochemically inferred ER+ status. Nevertheless, and unlike array #16, the analyses for case #11 are in clear-cut agreement with the original ER status determination. This implies that, beyond these top genes, there are possibly large numbers of interacting genes that strongly indicate ER+ status, over-riding the otherwise negative indications of these top few. Identification of relevant subsets of genes and elucidation of their conjoint action will be facilitated by exploration of the posterior summaries for $\boldsymbol{\beta}$ and the dependence structure among genes evident in the columns of the factor loading matrix $\mathbf{A}$ relating genes to supergene factors. This, and other exploratory studies, remains to be developed.

3.5 Array correlations and kernel structure

For this example, one further graph relates to the discussion of connections between standard regression models and kernel methods outlined in Section 2.6. As discussed there, our model has an implicit Gaussian process prior structure in which the vector of mean responses follows a normal prior, $N(\boldsymbol{\mu}|\mathbf{0}, \mathbf{K})$ with covariance matrix $\mathbf{K} = \mathbf{F}'\mathbf{T}\mathbf{F}$. Plugging-in posterior means of the elements τ_i^2 in the diagonal matrix $\mathbf{T}$ here provides an estimate of the kernel matrix that provides some insight into the nature of correlations across cases, i.e., across microarrays, as estimated by the model. Figure 16 displays the correlations extracted from this matrix. The upper frame displays correlations between microarray i and the remainder, for $i = 1, \ldots, 15$, i.e., for all ER+ arrays; the lower frame is a similar display for all ER− arrays. Note the very strong general patterns of correlation among arrays in the same class, and the weaker relationship between array 6 and the other ER+ cases. The most striking feature is, however, the inversion of the correlation patterns between array #16 and the rest, graphed as the broken black line in the lower frame. This reaffirms the issues of identification of array #16 as being much more strongly related to the ER+ arrays than to the ER−.

3.6 Screening genes and reanalysis using discriminatory subsets

There is much potential for exploring the posterior samples for $\boldsymbol{\beta}$ to generate further insights into the relevance of subsets of genes in terms of their contributions to the regression-based discrimination. For example, this kind of analysis provides a useful starting point for understanding gene interactions and generating direction in studies of genetic regulatory pathways. On the specific and narrower issue of capacity to discriminate new cases, however, it is evident that screening the large number of covariates to reduce to a smaller discriminatory subset can be expected to improve raw predictions. Several recent studies of discrimination using gene expression profiles have adopted simple rules for choosing small subsets of genes based on measures of variation of expression levels of individual genes between outcome groups (Golub *et al* 1999, Dudoit *et al* 2000). We do not subscribe to this as a general principle, since much relevant biological information may be lost in the process. That is, possibly many genes whose expression patterns conjointly relate in important ways to biological characteristics may be screened out using simple summary measures of variation in expression levels of individual genes. We prefer to explore subsets of genes based on understanding the posterior distribution for $\boldsymbol{\beta}$ from the full data analysis. However, to provide evidence of the potential to significantly improve predictive discrimination accuracy, we now explore a further analysis of the breast cancer data following a model-based screen.

As mentioned earlier, the marginal posterior samples for several hundred of the β_j coefficients indicate significant effects in the marginal sense. With this in mind we selected the top 400 genes by simply identifying those 400 genes with the largest estimated effects. We define this as the absolute values of the corresponding posterior means for regression coefficients, i.e., $|E(\beta_j|\mathbf{X}, \mathbf{z})|$. Note that this excludes genes with low estimated effects even though they are judged significant under the posterior, focusing more on practical rather than statistical significance. Using this subset of 400 genes, we then reran the full analysis, now with $n = 27$ and $p = 400$. Figures 13, 14 and 15 display results in the formats of those from the full data analysis in Figures 10, 11 and 12, respectively. We note the following. First, the analysis of the screened top 400 leads to increased accuracy in point estimates of classification probabilities, both in the fitted model analysis and in the one-at-a-time cross-validatory predictions, and for all microarrays. In the cross-validation analysis, all cases are now correctly classified, including the interesting case of array #16. For example arrays #3 and #25, the posterior histograms of predictive distributions in Figure 15 show little change relative to those from the full data analysis in Figure 12. For case #16, however, there are major differences; the posterior predictive distribution for π_{16} now concentrates more on values less than 0.5. The implication is that focusing only on the selected 400 genes has removed from the analysis a subset of genes whose expression levels and patterns are more consistent with ER$-$ status than ER+, and which, in the full data analysis, combine to weigh the posterior predictions significantly toward ER$-$ status. This subset of genes, possibly a very large subset of the initial group, are of obvious potential importance in understanding the biology behind ER status. Reducing to subsets driven by the focus on predictive discrimination is therefore only one component of the broader enterprise, and should not be adopted in a blind manner.

4 Expression profiling and discrimination of leukemias

Golub *et al* (1999) report on studies in molecular discrimination of leukemia types using microarray expression data. The leukemia study concerns expression arrays arising from individuals with one of two types of leukemia, ALL ($z_i = 1$) and AML ($z_i = 0$). The expression data is extracted from Affymetrix expression arrays, as in our breast studies, and provides two sets of arrays: an initial training set of 38 arrays (of which 27 are 1/ALL, and 11 are 0/AML), and a separate validation set of 34 additional arrays (of which 20 are 1/ALL and 14 are 0/AML). Using transformed expression data that has been screened to isolate genes whose observed expression levels are both variable and show evidence of substantial differences between the two outcome groups, the authors develop ad-hoc measures to score arrays for discrimination and classification. The summary analysis correctly identifies the class membership for most arrays, but several are identified as in a indifference zone and are left unclassified. We analyze data from this study here. First, we select the subset of 3,571 genes based on an initial processing adopted by the authors of the leukemia study. This preliminary processing is performed precisely as detailed in the recent work of Dudoit *et al* (2000), who analyzed precisely the same set of 3,571 observations. The expression summaries are the log (base 10) values of the actual expression levels following this initial filtering and transformation. Analysis in Golub *et al* (1999) is based on a much smaller number of genes eventually selected for discriminatory ability. In contrast, we maintain the focus on the large-scale expression patterns, and analyze the full set of 3,571 genes to identify interesting genes from a model-based perspective, and to assess uncertainty in classification both within-sample and in out-of-sample predictions on the full validation data set.

The results of the analysis indicate strong separation of the two leukemia classes based on posterior classification probabilities, for both the training data and the validation data; see Figures 17, 18 and 19. This is of course no surprise for the training cases, where there is clear discrimination. Most significant is the performance in classifying the 34 validation arrays, where the posterior predictive means of classification probabilities for all but 2 of the 34 lie on the correct side of 0.5. The two interesting cases, arrays #66 and #67, are further explored in displays of histograms of posterior predictive distributions for the classification probabilities, in Figure 19. Also displayed, for comparison, are cases #40 and #54, the two arrays least well classified (though correctly classified) in each of the two leukemia groups, respectively. In viewing these histograms, recall that the initial prior for each π_i is a vague, U-shaped density. The posteriors represent the shift from this prior to the posterior in each case. Cases #40 and #54 have posterior mass shifted over to favor the appropriate ranges, values nearer 1 in the case of array #40 and values nearer zero in the case of #54. The posteriors for the two interesting cases #66 and #67 tell rather different stories. They indicate that π_{66} is really rather likely to be nearer 1, with a high probability it exceeds 0.5, contradicting its medical classification as $z_{66} = 0$. Patterns of expression on this array must therefore share more of the characteristics of the arrays in the other class, at least for subsets of genes with meaningful effects as assessed by the model analysis. Case #67, however, has a posterior predictive density that is very similar to the symmetric prior, so indicating that the analysis has added little information about its leukemia status; implicitly, this indicates that the expression profile on array #67 must share characteristics of arrays in each of the leukemia classes as represented by the training data. Figure 20 provides some indication of this. This displays actual expression levels – the analysis used the log (base 10) of these values – for four arrays, including these two interesting cases #66 and #67 along with arrays #48 and #52; these latter two arrays are each well classified by the analysis, one as ALL and one as AML, and appear to typify the two leukemia classes.

The final graphs, Figures 21 and 22, display predictive classification summaries from a further analysis using only the top 50 genes selected based on the ranked values of the estimated effects, on an absolute scale, $|E(\beta_j|\mathbf{X}, \mathbf{z})|$, as in the breast cancer example above. This is summarized only to provide confirmation that the results are relatively unchanged in moving from the full data set of 3,571 genes to a select top 50, and since the Golub *et al* (1999) study is ultimately based on a small, screened subset of genes. From Figure 21 it is clear that the predictive probabilities for all validation arrays are more extreme in this analysis, indicating generally increased classification accuracy. The two interested cases, #66 and #67, still stand out. From Figure 19 it is evident that the major difference in the predictive distributions lies in that for case #67, where the posterior mass is more concentrated around smaller values for π_{67}. This is analogous to the conclusion over the anomalous array in the breast cancer example, arising as screening to a select subset of discriminatory genes removes a great deal of the conflicting information in the expression profile.

As a final comment, note that Golub *et al* (1999) and Dudoit *et al* (2000) use summary measures of expression differences between groups for preliminary screening to select interesting and discriminatory subsets. Some of these methods are purely ad-hoc, some more formally statistically derived. We have explored predictive discrimination using our model with gene subsets selected using the screening measures in the above references. We find that, uniformly, predictive classification in the binary regression framework is more accurate when subsets are selected using our approach, i.e., by selecting subsets based on the initial analysis of the full set of genes. However, there remains much to be explored in studies of subset selection for improved predictive discrimination.

References

1. Albert, J.H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, **88**, 669-679.

2. Albert, J. and Johnson, V.E. (1999), *Ordinal Data Models*, New York: Springer-Verlag.

3. Brown, M.P.S., Grundy, W.N., Lin, D., Christiani, N., Sugnet, C., Ares, M. and Haussler, D. (1999), "Support vector machine classification of microarray gene expression data," Technical report UCSC-CRL-99-09, Department of Computer Science, University of California at Santa Cruz.

4. Dudoit, S., Fridlyand, J. and Speed, T.P. (2000), "Comparison of discrimination methods for the classification of tumors using gene expression data," Unpublished Technical Report, Department of Statistics, University of California at Berkeley.

5. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Dowing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999), "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science,* **286**, 531-537.

6. Lockhart D.J., Dong, H., Byrne, M.C., Folliette, M.T., Gallow, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., and Horton, H. (1996), "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology,* **14**, 1675-1680.

7. Neal, R.M. (1999), "Regression and classification using Gaussian process priors" (with discussion), in *Bayesian Statistics 6,* (eds: J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), Oxford: University Press, 475-502.

8. Schadt, E.E., Li, C., Su. C. and Wong, W. (2000), "Analyzing high-density oligonucleotide gene expression array data," *J. Cell Biochemistry*, (in press).

9. Schölkopf, C., Burges, J. and Smola, A. (1999) *Advances in Kernel Methods: Support Vector Learning*, Cambridge: MIT Press.

10. West, M. (2000), "Bayesian regression analysis in the "Large **p**, Small **n**" paradigm" ISDS Discussion Paper, Duke University (submitted for publication).

11. Zellner, A. (1986), "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (eds: P.K. Goel and A. Zellner), Amsterdam: North-Holland, 233-243.

Table 1

Genes “up” in ER+:

mRNA-for-estrogen-receptor
intestinal-trefoil-factor-mRNA
ps2-protein-gene
nat1-gene-for-arylamine-n-acetyltransferase
hgata3-mRNA-for-trans-acting-t-cell-specific-transcription-factor
hepatocyte-nuclear-factor-3-alpha-(hnf-3-alpha)-mRNA
prolactin-induced-protein
mRNA-for-cardiac-gap-junction-protein
breast-cancer,-estrogen-regulated-liv-1-protein-(liv-1)-mRNA
clone-23948-mRNA-sequence
x-box-binding-protein-1-(xbp-1)-mRNA
gata-3-mRNA
type-1-angiotensin-ii-receptor-[human,-liver,-mRNA,-2268-nt]
mRNA-for-lung-amiloride-sensitive-na+-channel-protein
nonspecific-crossreacting-antigen-mRNA
androgen-receptor-mRNA
neuropeptide-y-receptor-y1-(npyy1)-mRNA

Genes “down” in ER+:

matrilysin-gene
rar-responsive-(tig1)-mRNA
omega-light-chain-protein-14.1-(ig-lambda-chain-related)-gene
guanylate-binding-protein-isoform-i-(gbp-2)-mRNA
cystic-fibrosis-antigen-mRNA
gp-39-cartilage-protein-gene-extracted-from-h.sapiens-gene-encoding-cartilage-gp-39-protein
mRNA-for-antileukoprotease-(alp)-from-cervix-uterus
mesothelial-keratin-k7-(type-ii)-mRNA

Table 1: Descriptors of “Top 25” genes

Captions for Figures

Figure 1. Expression by arrays: Top eight(+) genes

Figure 2. Expression by arrays: Top eight(−) genes

Figure 3. Relative contributions of factors – d_j^2 as % of total

Figure 4. Four of the factors across arrays

Figure 5. Arrays plotted on pairs of three factors

Figure 6. Summary of binary regression fit: Regression parameters

Figure 7. Posterior margins for three $\boldsymbol{\gamma}$ coefficients

Figure 8. Posterior margins for three $\boldsymbol{\beta}$ coefficients

Figure 9. Factor loadings $\mathbf{A}$ for top 750 genes

Figure 10. Summary of binary regression fit: Classification

Figure 11. Cross-validatory predictions

Figure 12. Uncertainty in classification

Figure 13. Summary of binary regression fit on top 400 genes: Classification

Figure 14. Cross-validatory predictions using top 400 genes

Figure 15. Uncertainty in classification using top 400 genes

Figure 16. Estimated kernel correlations among breast cancer microarrays

Figure 17. Summary of binary regression fit for leukemia data: Training sample

Figure 18. Cross-validatory predictions for leukemia data: Validation sample

Figure 19. Uncertainty in predictive classification for some leukemia cases

Figure 20. Expression data for top 50 genes on four leukemia arrays

Figure 21. Predictions for leukemia data using only top 50 genes: Validation sample

Figure 22. Uncertainty in leukemia classification in analysis of top 50 genes

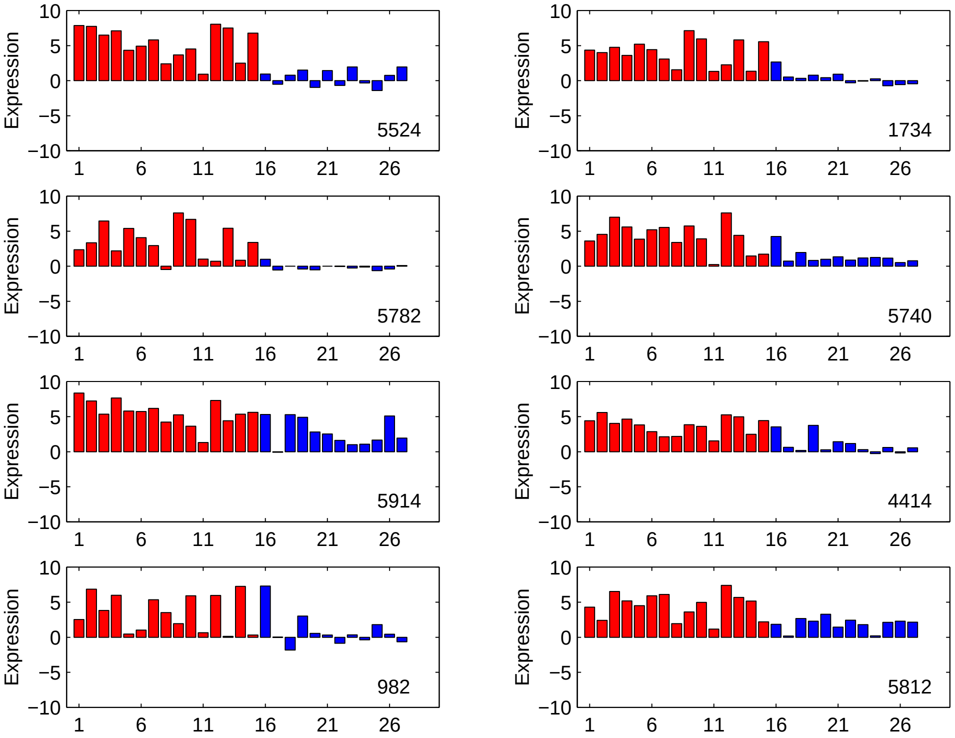

Figure 1: Expression by arrays: Top eight(+) genes

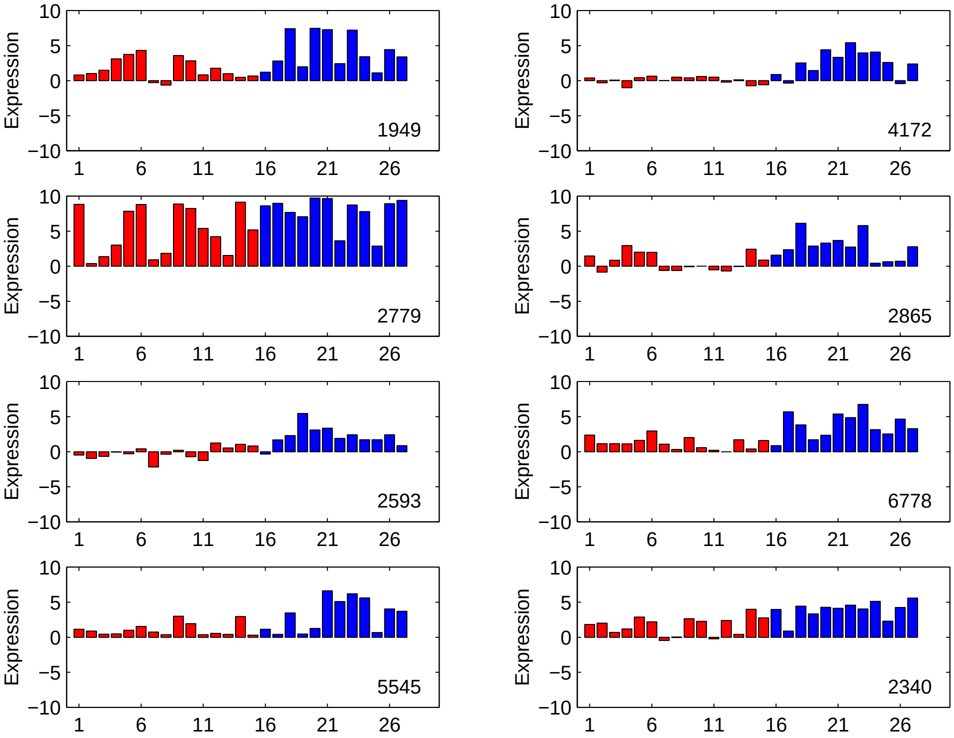

Figure 2: Expression by arrays: Top eight(−) genes

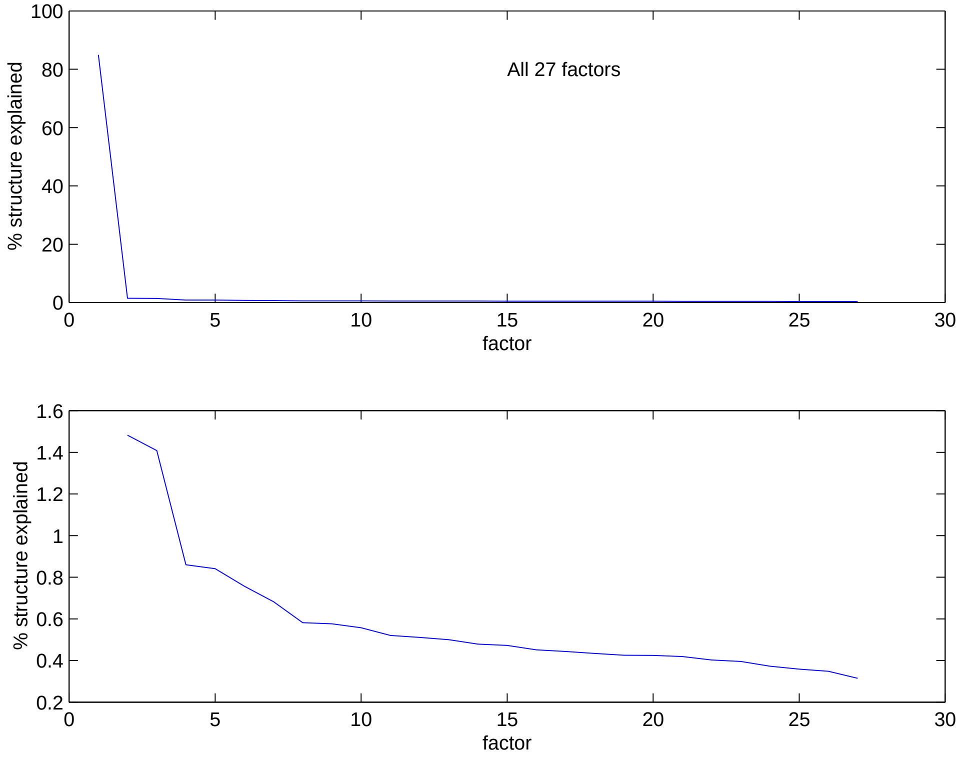

Figure 3: Relative contributions of factors – d_j^2 as % of total

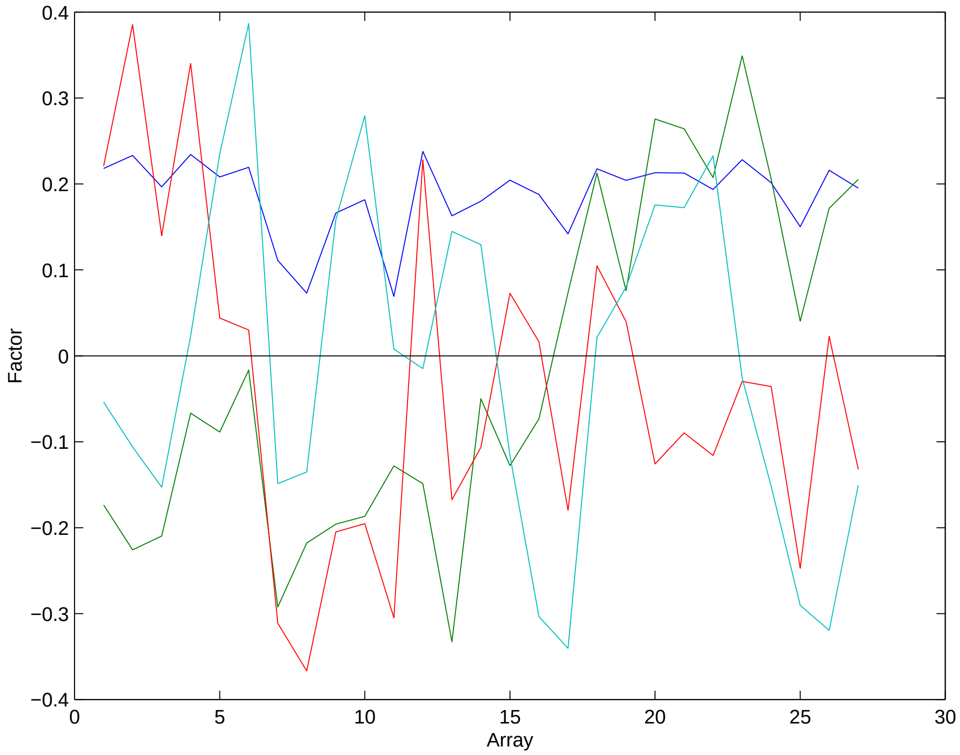

Figure 4: Four of the factors across arrays

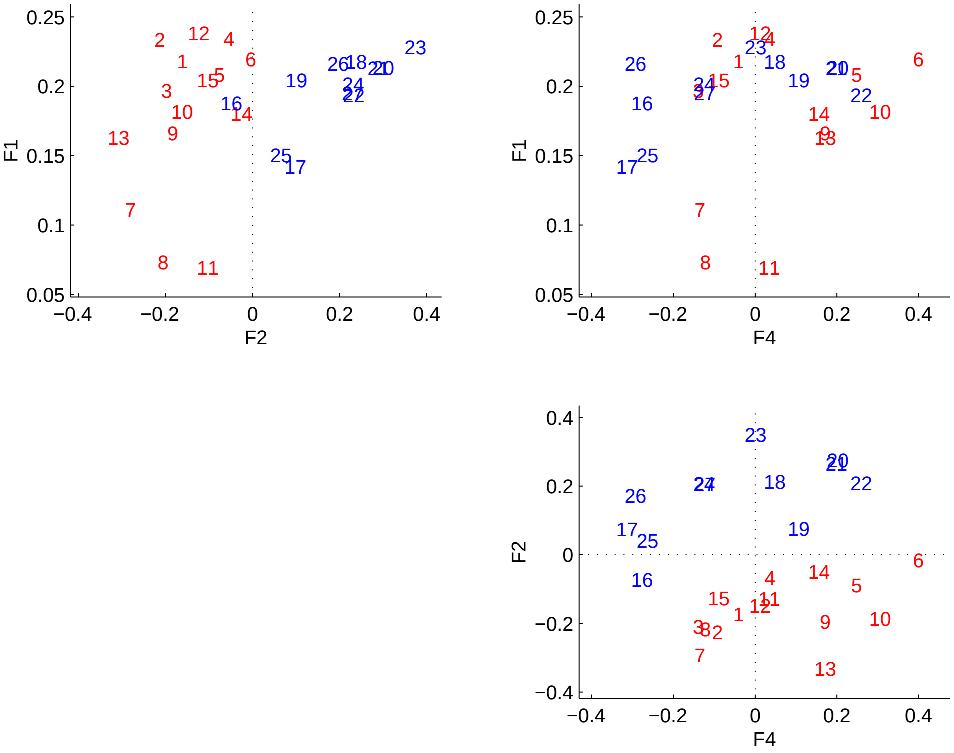

Figure 5: Arrays plotted on pairs of three factors

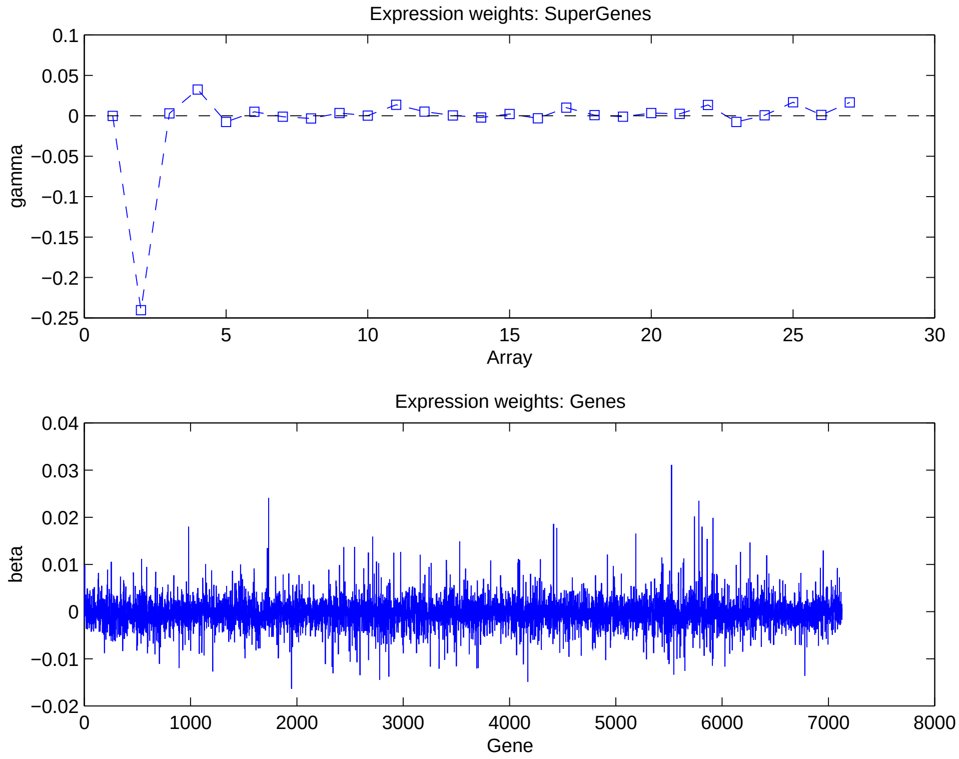

Figure 6: Summary of binary regression fit: Regression parameters

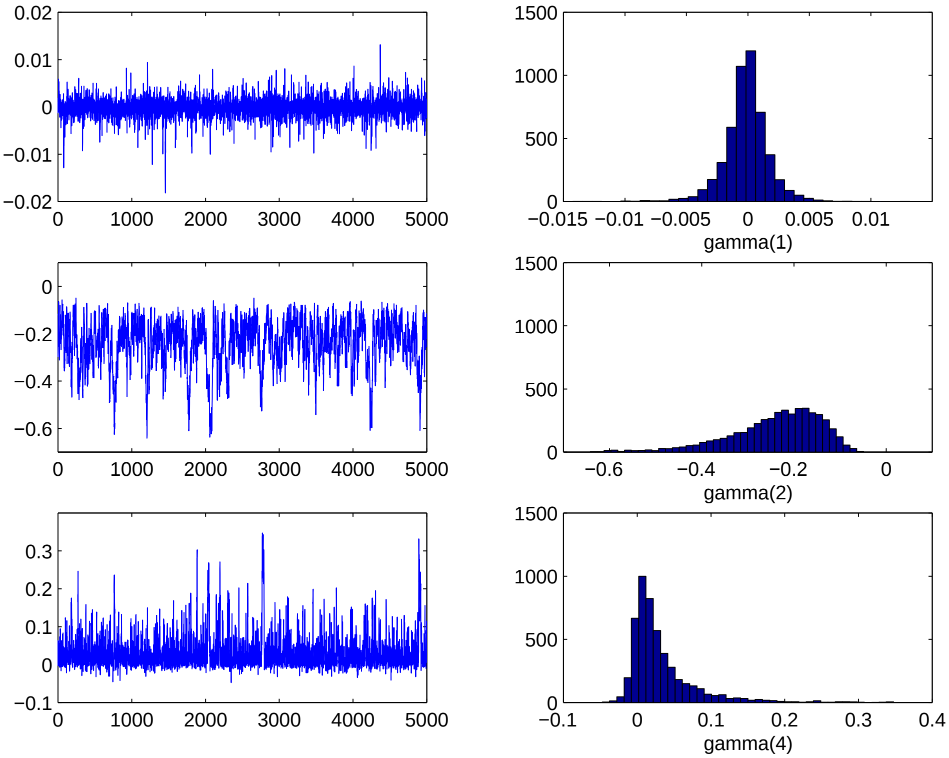

Figure 7: Posterior margins for three γ coefficients

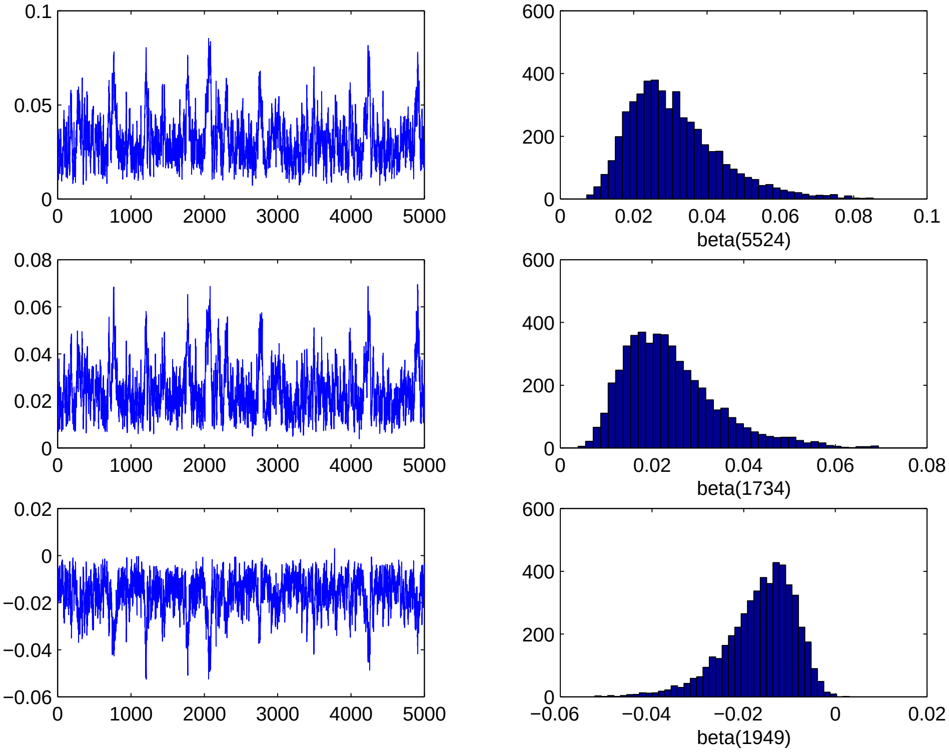

Figure 8: Posterior margins for three β coefficients

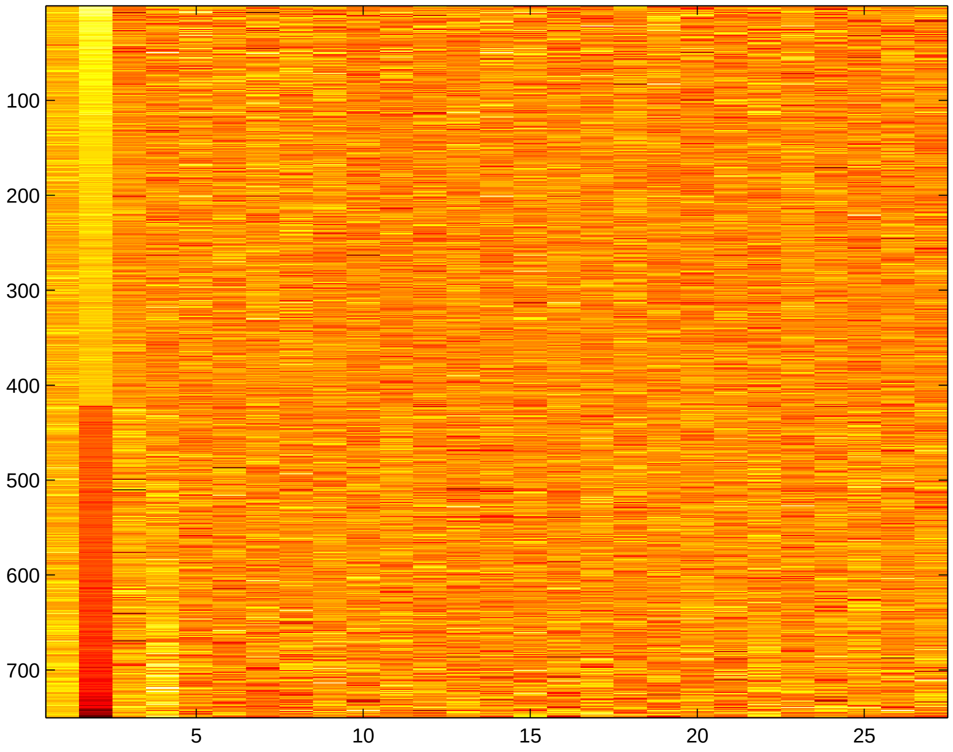

Figure 9: Factor loadings **A** for top 750 genes

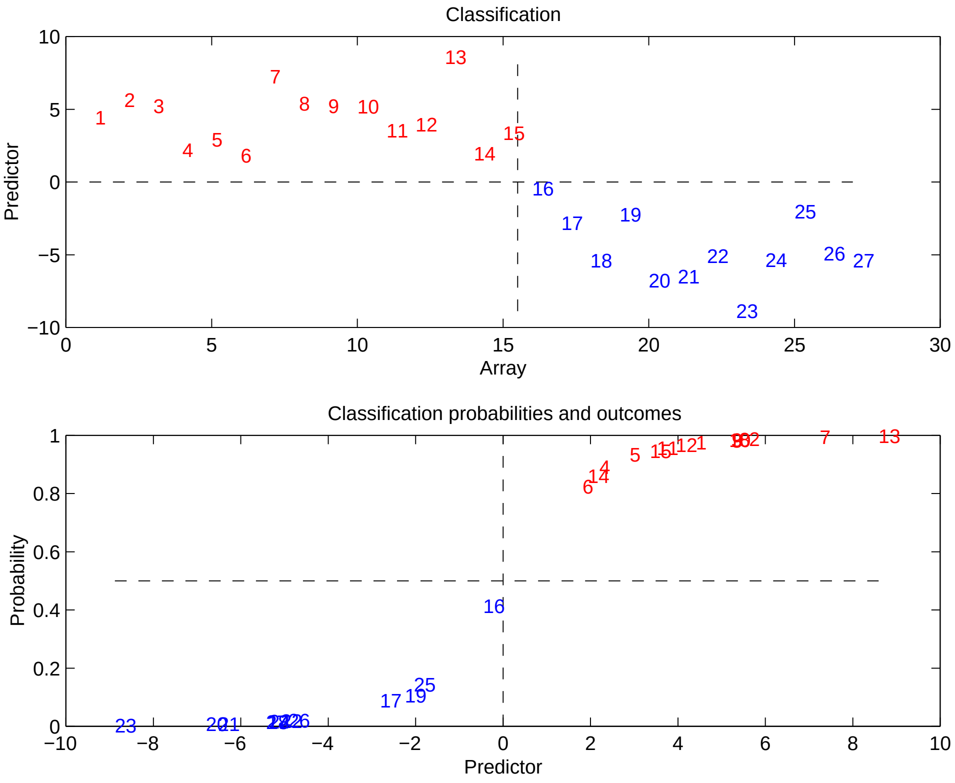

Figure 10: Summary of binary regression fit: Classification

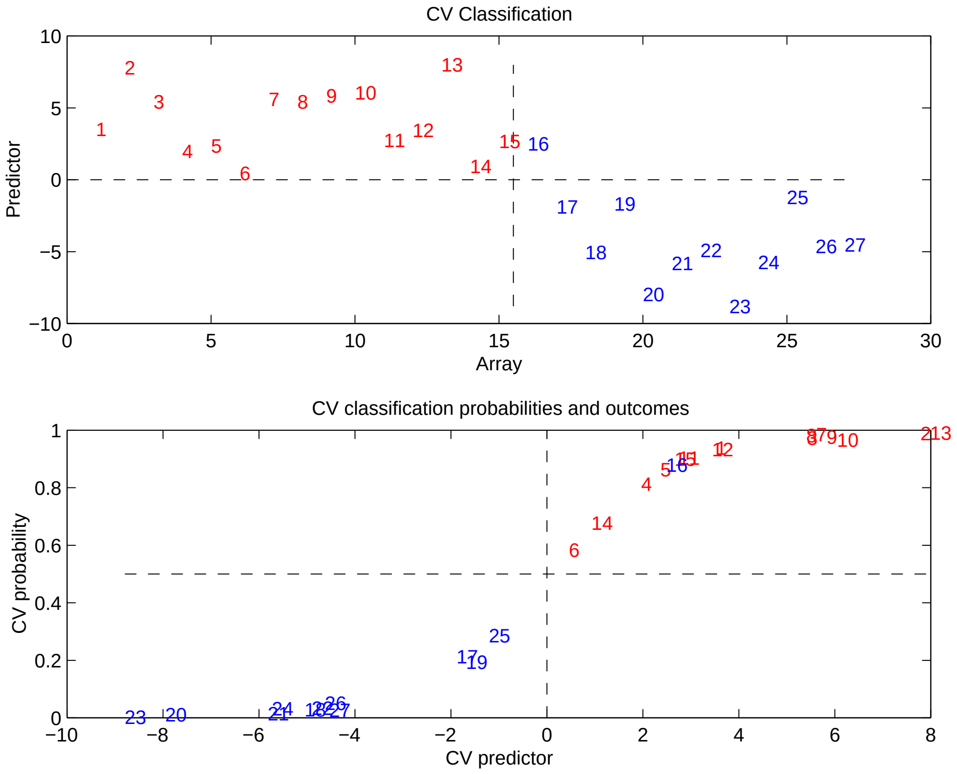

Figure 11: Cross-validatory predictions

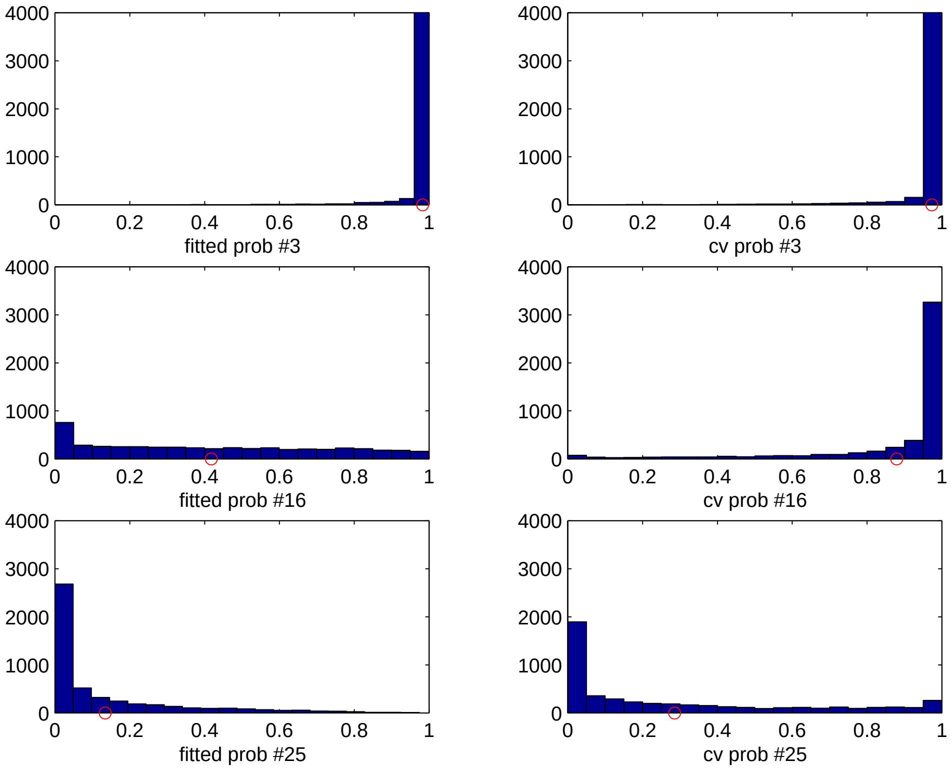

Figure 12: Uncertainty in classification

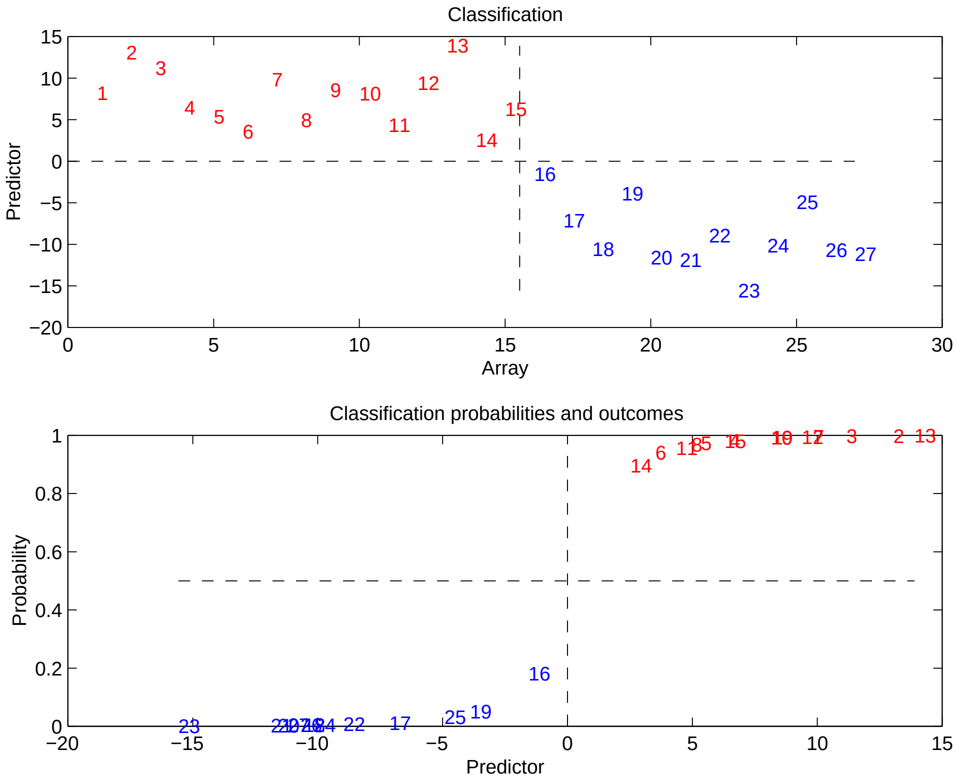

Figure 13: Summary of binary regression fit on top 400 genes: Classification

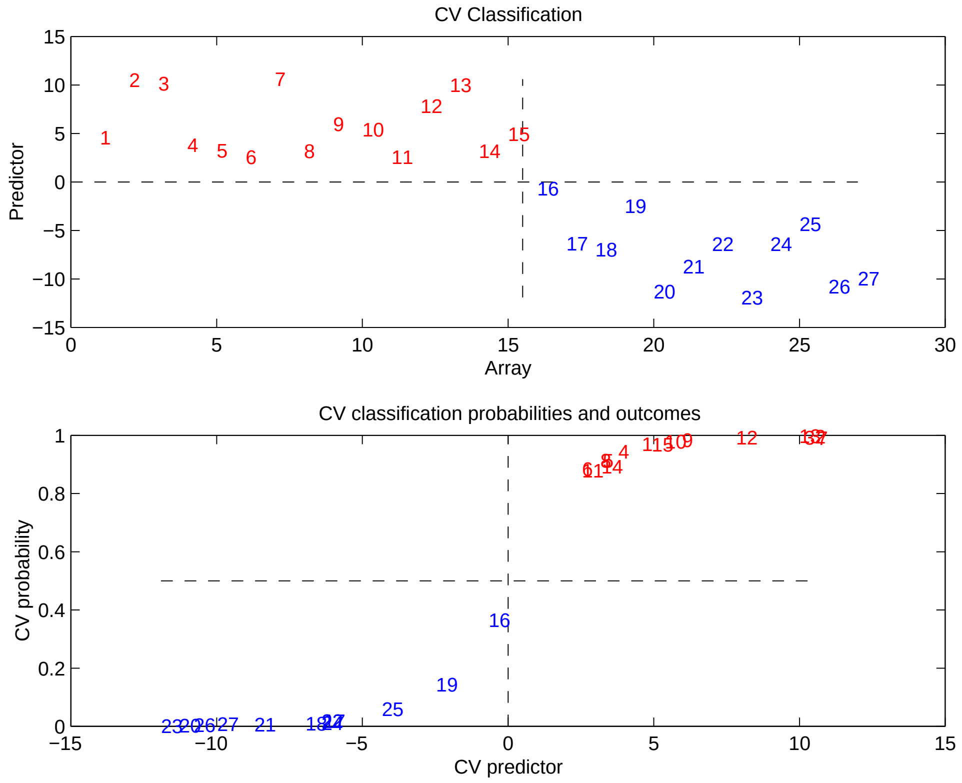

Figure 14: Cross-validatory predictions using top 400 genes

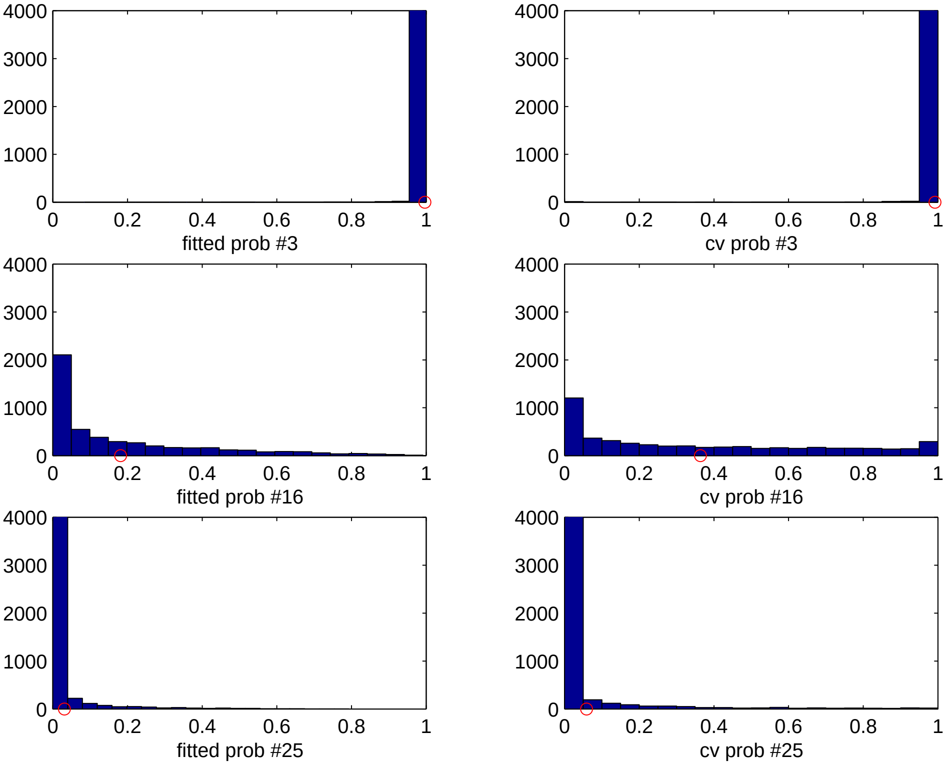

Figure 15: Uncertainty in classification using top 400 genes

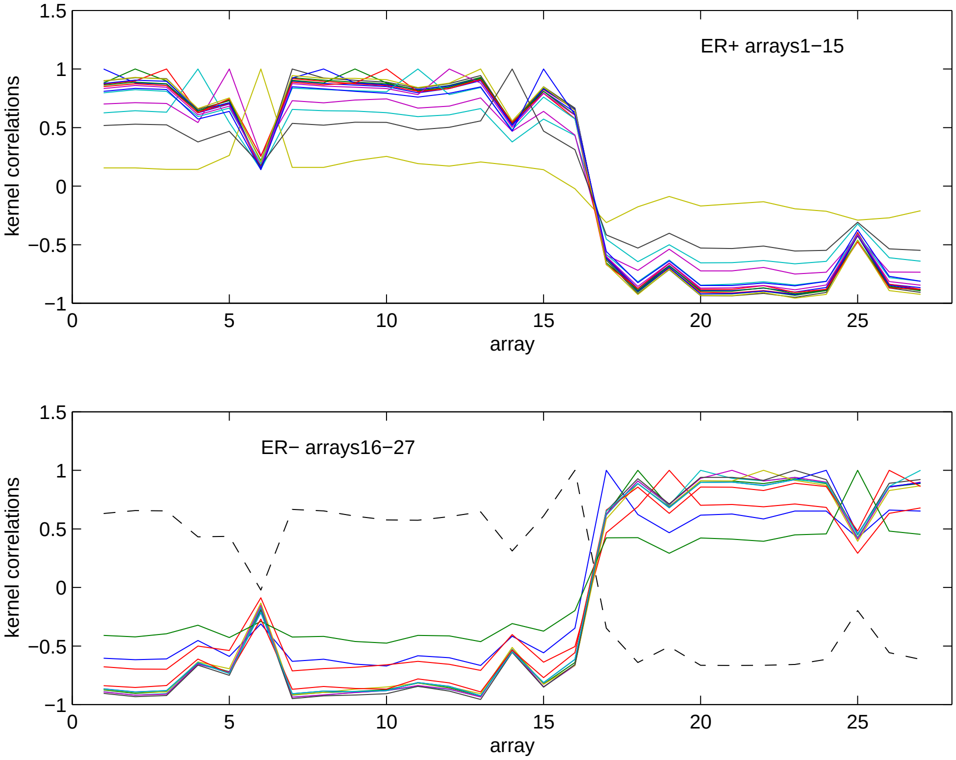

Figure 16: Estimated kernel correlations among breast cancer microarrays

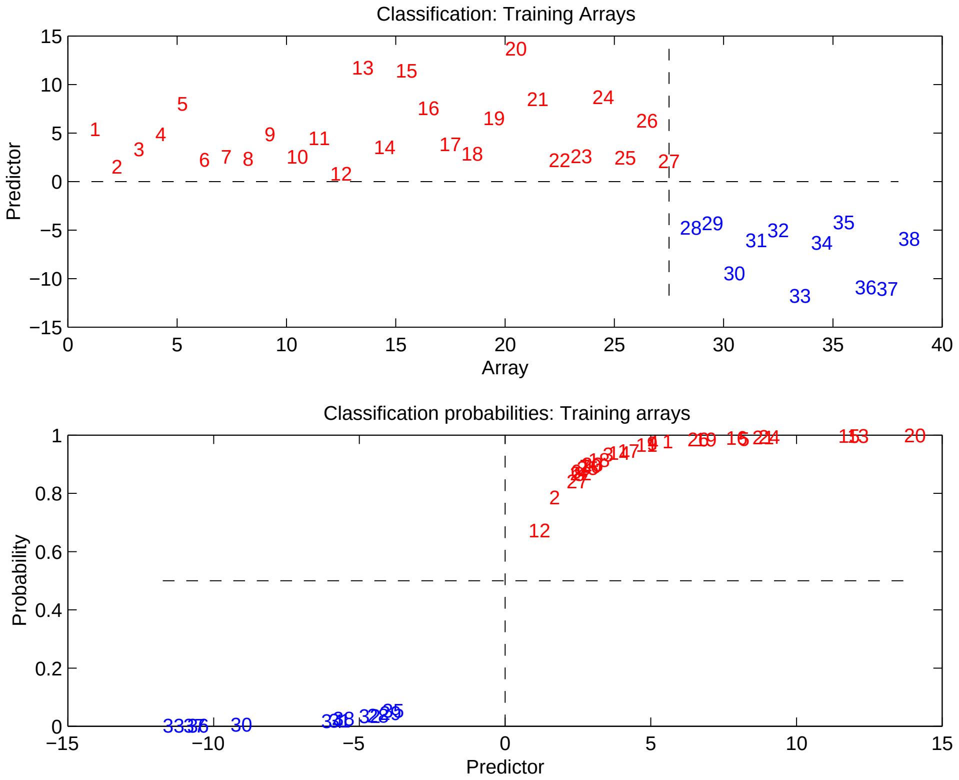

Figure 17: Summary of binary regression fit for leukemia data: Training sample

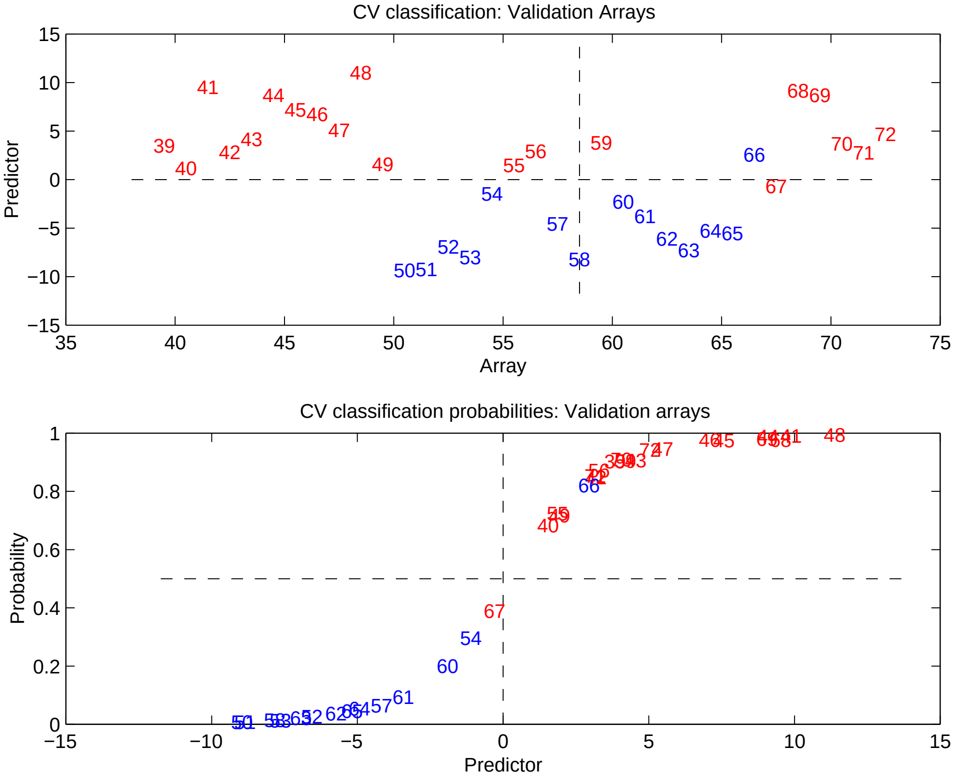

Figure 18: Predictions for leukemia data: Validation sample

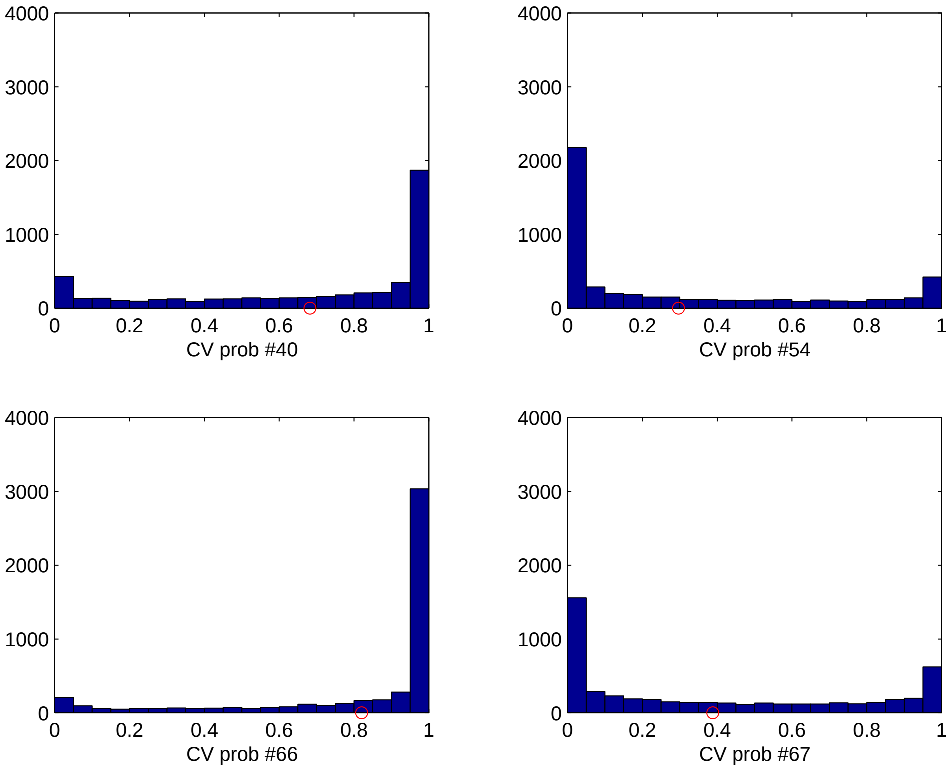

Figure 19: Uncertainty in predictive classification for some leukemia cases

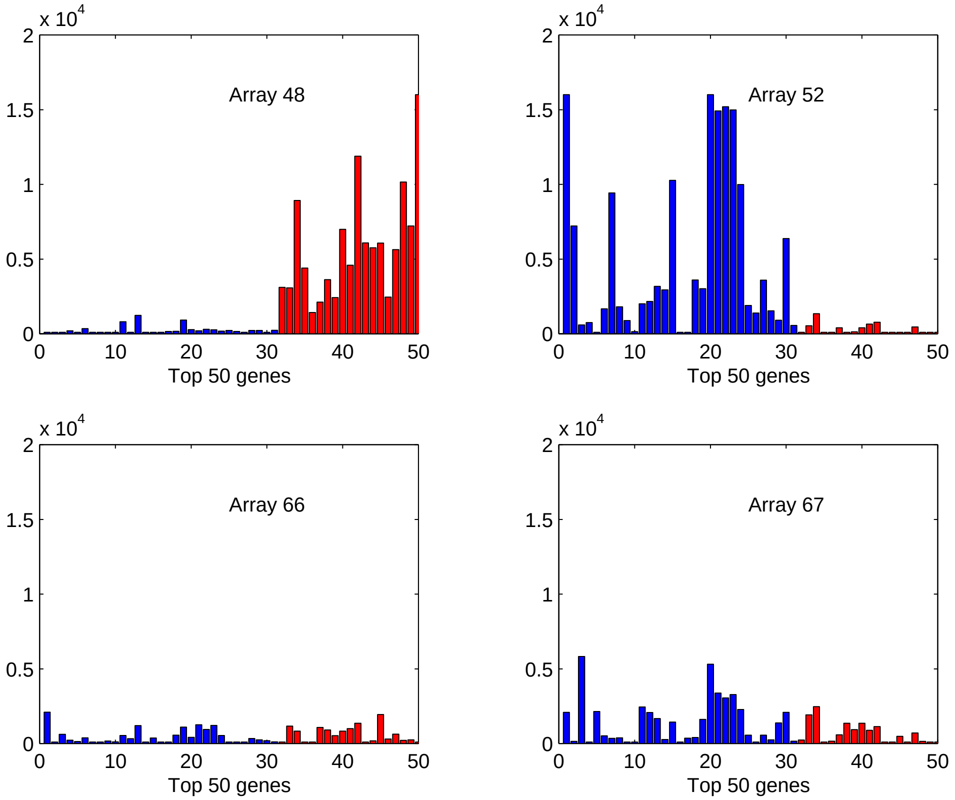

Figure 20: Expression data for top 50 genes on four leukemia arrays

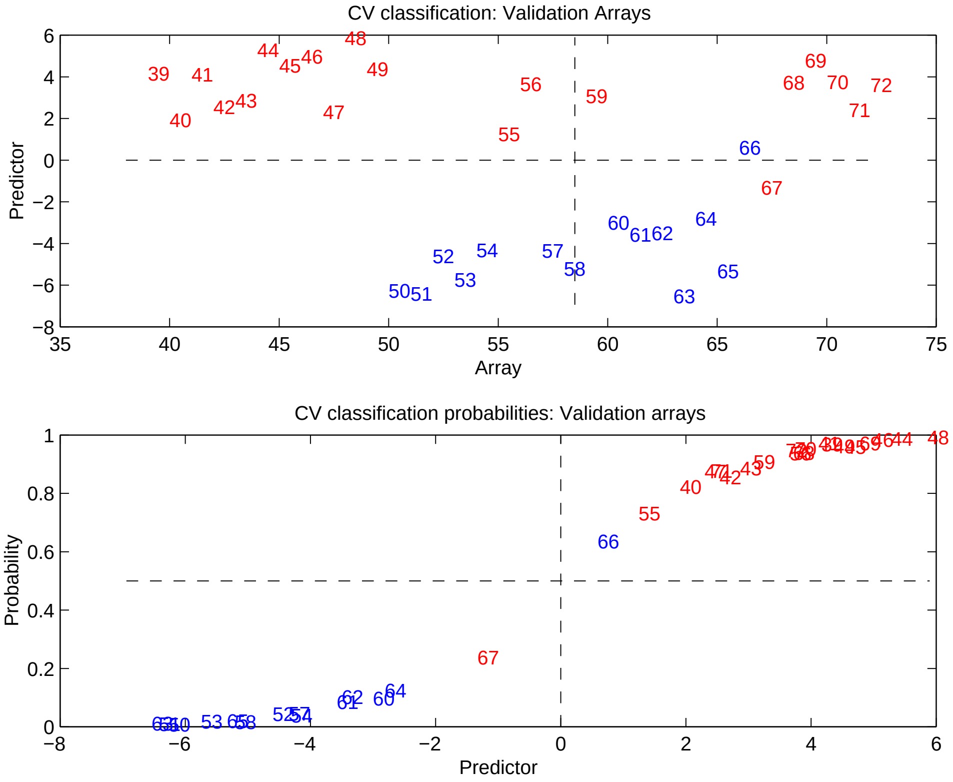

Figure 21: Predictions for leukemia data using only top 50 genes: Validation sample

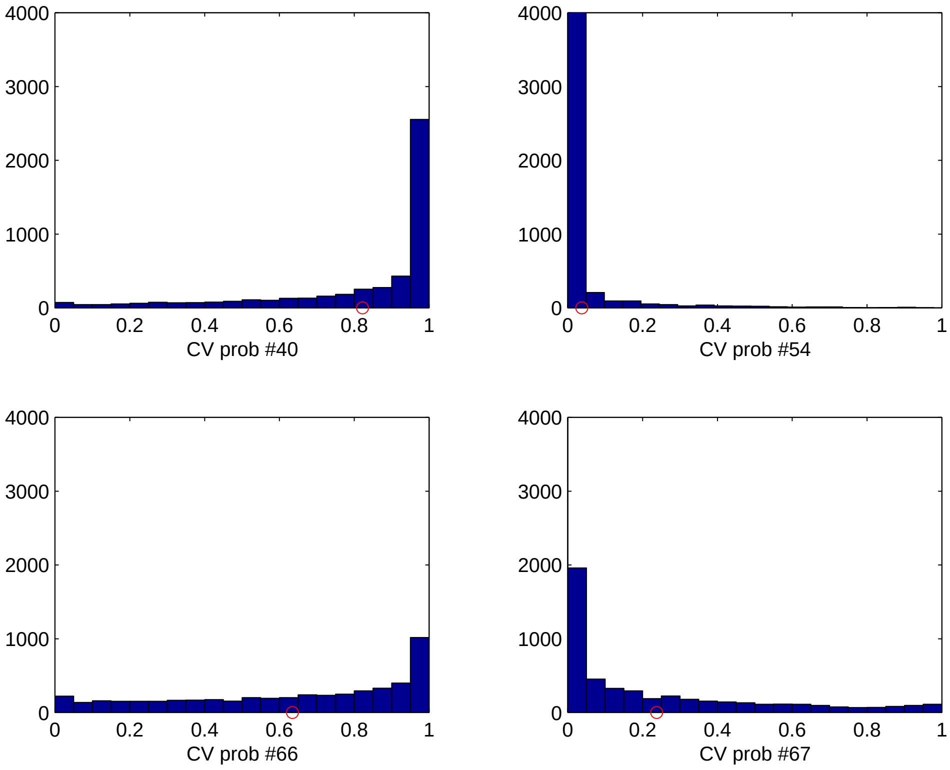

Figure 22: Uncertainty in leukemia classification in analysis of top 50 genes