BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Bayesian Factor Regression Models in the "Large p, Small n" Paradigm

MIKE WEST
ISDS, Duke University, Durham, NC 27708, USA
mw@isds.duke.edu

SUMMARY

I discuss Bayesian factor regression models with many explanatory variables. These models are of particular interest and applicability in problems of prediction, but also for elucidating underlying structure in predictor variables. One key motivating application here is in studies of gene expression in functional genomics. I first discuss empirical factor (principal components) regression, and the use of general classes of shrinkage priors, with an example. These models raise foundational questions for Bayesians, and related practical issues, due to the use of design-dependent priors and the need to recover inferences on the effects of the original, high-dimensional predictors. I then discuss latent factor models for high-dimensional variables, and regression approaches in which low-dimensional latent factors are the predictor variables. These models generalise empirical factor regression, provide for more incisive evaluation of factor structure underlying high-dimensional predictors, and resolve the modelling and practical issues in empirical factor models by casting the latter as limiting special cases. Finally, I turn to questions of prior specification in these models, and introduce *sparse latent factor models* to induce sparsity in factor loadings matrices. Embedding such sparse latent factor models in factor regressions provides a novel approach to variable selection with very many predictors. The paper concludes with an example of sparse factor analysis of gene expression data and comments about further research.

Keywords: DIMENSION REDUCTION, GENE EXPRESSION ANALYSIS, HIGH-DIMENSIONAL COVARIATES, LATENT FACTOR MODELS, SHRINKAGE PRIORS.

1. EMPIRICAL FACTOR REGRESSION MODELS

1.1 *SVD Regression*

Begin with the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{y}$ is the $n-$vector of responses, $\boldsymbol{X}$ is the $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is the $p-$vector regression parameter, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{\epsilon} \,|\, 0, \sigma^2 \boldsymbol{I})$ is the $n-$vector error term. Of key interest are cases when $p >> n$, when $\boldsymbol{X}$ is "long and skinny." The standard empirical factor (principal component) regression is best represented using the reduced singular-value decomposition (SVD) of $\boldsymbol{X}$, namely $\boldsymbol{X} = \boldsymbol{F}\boldsymbol{A}$ where $\boldsymbol{F}$ is the $n \times k$ factor matrix (columns are factors, rows are samples) and $\boldsymbol{A}$ is the $k \times p$ SVD "loadings" matrix, subject to $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$ and $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{D}^2$ where $\boldsymbol{D}$ is the diagonal matrix of k positive singular values, d_i, arranged in decreasing order. This reduced form assumes factors with zero singular values have been ignored without loss; $k \leq n$ with equality only if all singular values are positive. Now the regression transforms via $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{F}\boldsymbol{\theta}$ where $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$ is the $k-$vector of regression parameters for the factor variables, representing a possibly massive dimension reduction from p to k parameters.

Inherently, observing $\boldsymbol{y}$ provides information only on the underlying factor regression parameters $\boldsymbol{\theta}$, so prior specification directly in factor space has become common. Generalised shrinkage (or ridge regression) priors on $\boldsymbol{\theta}$ have become popular as MCMC methods now permit their routine use. A particularly flexible class of such priors assumes independent T distributions for the elements $\theta_1, \ldots, \theta_k$ of $\boldsymbol{\theta}$, so allowing for varying degrees of shrinkage in each of the orthogonal factor dimensions. A particular example has $\theta_i \sim \mathrm{N}(\theta_i \,|\, 0, c_i/\phi_i)$ where $\phi_i \sim \mathrm{Ga}(\phi_i \,|\, r/2, r/2)$ independently, for some $r > 0$; the tuning parameter r is the degree-of-freedom parameter for the implied T distribution for θ_i that follows on marginalisation over the random precision ϕ_i. Here the c_i are weights that may be used, for example, to indicate the prior view that higher-order factors are expected to play lesser roles in the regression – often, though not necessarily, the case. In the first example below this is the case, and the model uses $c_i = \rho i^{-2}$, with scale factor ρ to be estimated, for example. This also allows for removal of factors in prior specification, by setting a c_i to zero. Analyses typically also adopt an inverse gamma prior for the error variance σ^2.

These models are easily implemented using MCMC, with complete conditional posteriors of generally standard forms. The conditional posterior for $\boldsymbol{\theta}$ given the ϕ_i is multivariate normal, and the ϕ_i are conditionally independent gamma variates given $\boldsymbol{\theta}$. The example below utilises this to generate sequences of posterior samples for $\boldsymbol{\theta}$ and the ϕ_i.

1.2 *Coherence and Inference on Original Regression Parameters*

A basic modelling issue arises from the explicit design-, and sample size-, dependence of the empirical factor model. The key $\boldsymbol{\theta}$ parameter is directly defined as a function of $\boldsymbol{X}$ and $\boldsymbol{\beta}$, so parameter definition changes as the sample size and design changes; the specification of priors over these design-dependent parameters must be coherent with respect to changing n and $\boldsymbol{X}$, and the question is raised of whether this can be assured. This question is answered in Section 2.

A related practical issue is that of inference on the original regression parameters $\boldsymbol{\beta}$. The framework has priors and posteriors for $\boldsymbol{\theta}$, but leaves open questions of "inverting" the dimension-reducing map to make inferences on $\boldsymbol{\beta}$. The many-one map $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$ has multiple generalised inverses $\boldsymbol{\beta} = \boldsymbol{A}'\boldsymbol{\theta} + \boldsymbol{b}$ for all $p-$vectors $\boldsymbol{b}$ such that $\boldsymbol{A}\boldsymbol{b} = 0$. Again, this issue is fully resolved in Section 2. Here, simply note that, for predictive purposes, the choice of $\boldsymbol{b}$ is irrelevant. A canonical choice of generalised inverse is the standard "least-norm" inverse based on $\boldsymbol{b} = 0$, i.e., $\boldsymbol{\beta}^* = \boldsymbol{A}'\boldsymbol{\theta}$. The analysis in the example below uses $\boldsymbol{\beta}^*$; again, Section 2 explains and justifies this properly. Posterior samples for $\boldsymbol{\theta}$ trivially imply samples for $\boldsymbol{\beta}^*$ which may be summarised for inference.

One interesting connection to make is that the prior on $\boldsymbol{\beta}^*$ implied by the prior on $\boldsymbol{\theta}$ of Section 1.1 above is a generalisation of the $g-$prior of Zellner (1986). Given conditional $\mathrm{N}(\theta_i \,|\, 0, c_i/\phi_i)$ priors, the implied prior for $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ is singular normal with density proportional to $\exp\{-\boldsymbol{\beta}'\boldsymbol{A}'\boldsymbol{G}\boldsymbol{A}\boldsymbol{\beta}/2\}$, where $\boldsymbol{G}$ is diagonal with elements ϕ_i/c_i. This makes explicit the design-dependency of the prior and also the individual scaling in each of the singular factor dimensions. The Zellner $g-$prior corresponds to taking $\boldsymbol{G} = g\boldsymbol{D}^2$ for some positive scalar $g > 0$. I therefore refer to this approach as defining a class of *generalised singular $g-$priors* (or $gsg-$priors).

1.3 *Prediction*

Prediction is practically straightforward, though raises foundational questions related to the design-dependency issue. Technically, response values to be predicted are simply treated as missing values to be imputed, and the MCMC analysis is trivially extended to sample these values at each iteration. That is, $\boldsymbol{y}$ is partitioned into a vector of training samples, $\boldsymbol{y}_t$, and a vector of validation cases $\boldsymbol{y}_v$ to be predicted; the design matrix is conformably partitioned as $[\boldsymbol{X}_t; \boldsymbol{X}_v]$. The MCMC imputes $\boldsymbol{y}_v$ from the implied (normal) conditional posterior. This requires that the model is specified and analysed conditional on all predictor values, including $\boldsymbol{X}_v$, and the empirical factor regression model is based on decomposition of the full $\boldsymbol{X}$ matrix. As a result, the factors $\boldsymbol{F}$ are evaluated based on predictors $\boldsymbol{X}_v$ as well as $\boldsymbol{X}_t$; thus $\boldsymbol{X}_v$ forms part of the model and prior structure even though the corresponding responses are missing. One message is that required predictor values must be contemplated prior to analysis, or analysis fully repeated if new predictions are required.

Again, this apparent issue is interpreted and resolved in Section 2 where the empirical model is understood to arise from a more elaborate latent factor regression model. In the example now, this approach is simply adopted for prediction of validation samples.

1.4 *Example: Analysis of Biscuit Dough Data*

This example concerns biscuit dough data analysed in Brown *et al* (1999), and originally in Osborne *et al* (1984). The study aims to predict biscuit dough constituents based on spectral characteristics of dough measured using near infrared (NIR) spectroscopy. Hence the predictor for each dough sample is a reflectance spectrum on a grid of wavelengths. The analysis here uses the same data as Brown *et al* (although their framework is multivariate); these authors utilise a decision-theoretic variable selection method, rather than shrinkage priors or factor models. The response chosen here is fat content of dough samples, the predictors are $p = 300$ NIR reflectance measures at equally spaced wavelengths over $1202 - 2400$ nanometres (nm), with 39 training samples and 39 validation cases to be predicted. The fat content response is standardised, and the predictors are centred; the centred spectra are graphed in Figure 1.

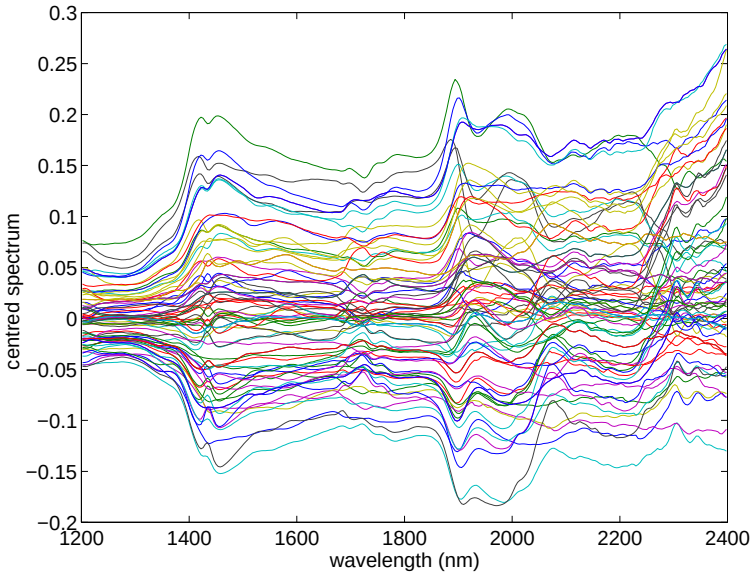

Figure 1. *Centred spectral predictors of 78 biscuit dough samples*

Several analyses have been studied, varying the tuning parameters r and c_i. A summary of analysis with $r = 5$ and $c_i = \rho i^{-2}$ is given here; analyses with r between 1 and 10 give substantially similar results, and $r = 5$ is marginally optimal as measured simply by the mean square prediction error computed in the validation sample. The prior for σ^{-2} is unit mean exponential, reflecting the known range of the standardised response data but otherwise representing a relatively diffuse prior, and ρ has a diffuse prior. The singular values of the 300×78 matrix $\boldsymbol{X}$ decay rapidly and are truncated to zero, with the corresponding factors dropped, past the point where 99.995% of the total variation, as measured by the cumulative sum of squares of the d_i, is accounted for; this leads to $k = 16$ factors in the model, with $c_i = 0$ for $i > 16$. Adding more factors, at the expense of increased computation, does not in this example improve predictions. Further, a reduced number is consistent with the inherent smoothness of the predictor variable, and will generally be experienced when predictors are curves. The MCMC analysis summarised has 20,000 samples selected from a run of 100,000 by choosing every fifth sample, and following a burn-in of 1,000 samples. Convergence is swift and clean, consistent with experiences with a range of other examples and MCMC experiments.

Figures 2 and 3 graph the approximate posterior means of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}^* = \boldsymbol{A}'\boldsymbol{\theta}$, respectively, the former with equal-tails 90% posterior intervals. Figure 3 uses asterisks to mark the ten β_i^* values with largest absolute posterior means. Of these, there is a small cluster at just over 1700nm, at values 1718, 1722, 1726, 1730 and 1734. These are noteworthy since, as remarked by Brown *et al*, this is a region where fat is known to have a characteristic absorbance; Brown *et al* identify the point 1718nm too. In Figure 1 it is possible to discern small wiggles in the spectra in this wavelength region.

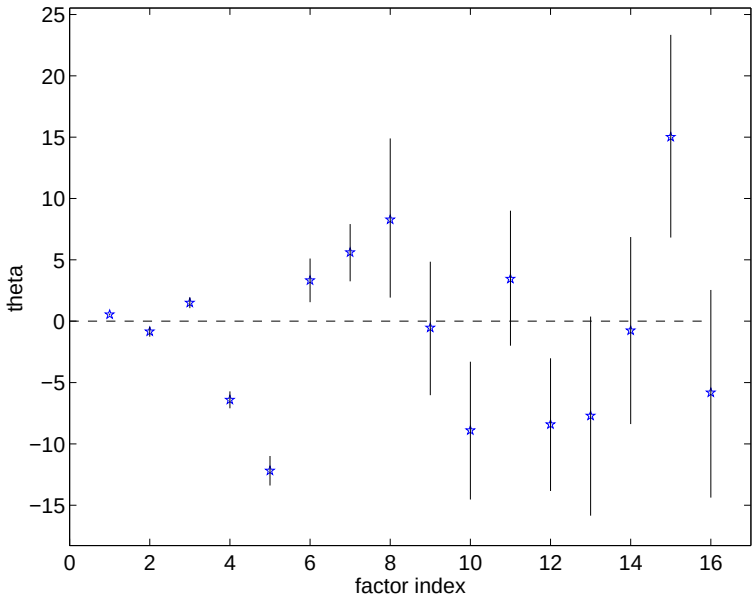

Figure 2. *Biscuit dough analysis: Estimates of empirical factor regression coefficients* $\boldsymbol{\theta}$

Figure 4 addresses prediction and model assessment via display of data plotted against fitted and predicted values; training data are indicated by asterisks, and validation data by circles (recall the original fat content values are standardised to define the response variable). The close concordance between observed and fitted/predicted values (as well as several other exploratory residual analyses, not reported) give no reason to question model fit. Of particular interest is the fact that the out-of-sample

predictions are very accurate indeed. In fact, on the basis of simple mean square prediction errors, this analysis improves on results of Brown *et al*, albeit only marginally. There is sensitivity to the tuning parameters r and the c_i, and some experimentation is needed to investigate this; the results reported are based on such experimentation and rough optimisation over these values with respect to the predictive assessment.

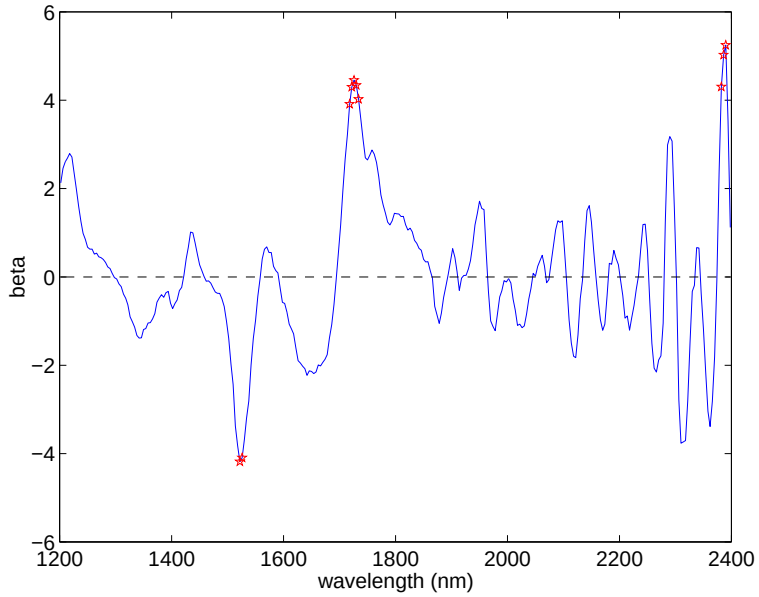

Figure 3. *Biscuit dough analysis: Estimates of wavelength regression coefficients* $\boldsymbol{\beta}^*$

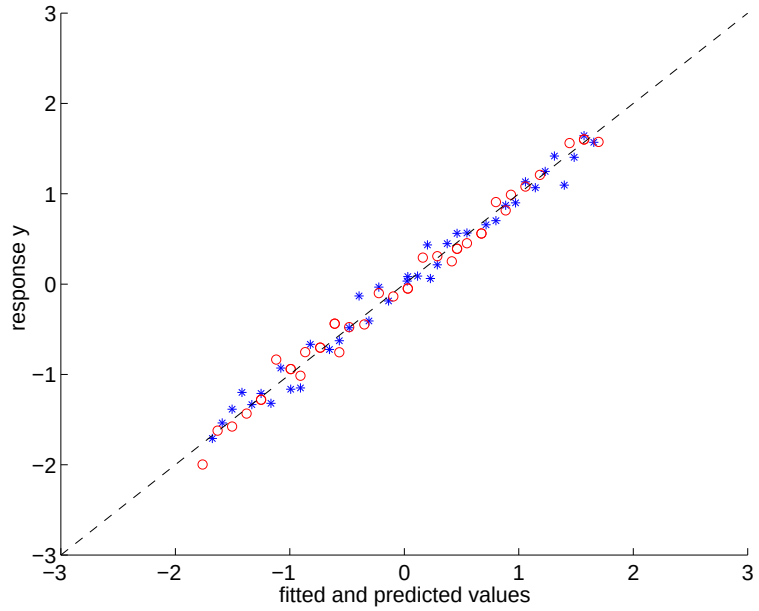

Figure 4. *Biscuit dough analysis: Response versus fitted and predicted values*

2. LATENT FACTOR REGRESSION MODELS

Formal latent factor models aim to partition variation in the predictor variables into multiple components that reflect common patterns, and separate these from variation that is idiosyncratic to each variable, or "noise." Here I note the theoretical structure of standard linear, latent factor models and define a class of *factor regression models* that naturally relate underlying latent structure in high-dimensional predictors to responses. I show that the empirical model of Section 1 arises as a limiting case, and how this clarifies the design-dependency issues discussed in Section 1.

Write $\boldsymbol{x}_i'$ for the i^{th} row of $\boldsymbol{X}$ and consider the latent factor model (e.g., Aguilar and West, 2000; Lopes and West, 1999)

$$\boldsymbol{x}_i = \boldsymbol{B}\boldsymbol{\lambda}_i + \boldsymbol{\nu}_i \tag{1}$$

where

$$\boldsymbol{\lambda}_i \sim \mathrm{N}(\boldsymbol{\lambda}_i \,|\, 0, \Delta^2) \quad \text{and} \quad \boldsymbol{\nu}_i \sim \mathrm{N}(\boldsymbol{\nu}_i \,|\, 0, \Psi^2). \tag{2}$$

Here $\boldsymbol{\lambda}_i$ is a $k-$vector of uncertain latent factors for case i, $\boldsymbol{B}$ is a $p \times k$ factor loadings matrix parameter, and $\boldsymbol{\nu}_i$ is a vector of idiosyncratic noise terms; both Δ and Ψ are diagonal. The number of factors is fixed and $k << p$. With appropriate identifying constraints on $\boldsymbol{B}$, this is an estimable model that attributes common structure in $\boldsymbol{X}$ to underlying $k-$dimensional factors, and isolates variation that is purely idiosyncratic in the $\boldsymbol{\nu}_i$ terms. MCMC based Bayesian analysis, and aspects of identification and prior specification, appear in Lopes and West (1999) and, in more elaborate factor models in time series, in Aguilar and West (2000). These two papers provide copious references to a large literature on Bayesian factor models. Mackay and Miskin (2001) is a recent, independent contribution that is also partly motivated by gene expression studies.

Assume that the responses $\boldsymbol{y} = (y_1, \ldots, y_n)'$ relate directly to the k latent factors; for each i,

$$y_i = \boldsymbol{\lambda}_i'\boldsymbol{\theta} + \epsilon_i \quad \text{where} \quad \epsilon_i \sim \mathrm{N}(\epsilon_i \,|\, 0, \sigma^2). \tag{3}$$

The original design variables $\boldsymbol{x}_i$ provide information on the latent variables through (1), but do not enter the regression; y_i is conditionally independent of $\boldsymbol{x}_i$ given $\boldsymbol{\lambda}_i$. One implication is that idiosyncratic variation in $\boldsymbol{X}$ now has no influence on the regression.

In Section 3 I discuss and exemplify analysis of the latent factor model alone; space here precludes full development of analysis and examples of the linked factor regression models, but it is of key interest to consider a theoretical limiting special case under the natural prior specification $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\theta} \,|\, 0, \boldsymbol{G}^{-1})$ with diagonal precision matrix $\boldsymbol{G}$.

Equations (1-3) imply a joint normal distribution for $(y_i, \boldsymbol{x}_i, \boldsymbol{\lambda}_i)$, and hence an implied conditional distribution for y_i given only $\boldsymbol{x}_i$ and the model parameters. After some algebra, this is

$$(y_i \,|\, \boldsymbol{x}_i) \sim \mathrm{N}(y_i \,|\, \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2 + \boldsymbol{\theta}'\boldsymbol{C}\boldsymbol{\theta}) \tag{4}$$

where $\boldsymbol{\beta} = \Psi^{-2}\boldsymbol{B}\boldsymbol{C}\boldsymbol{\theta}$ with $\boldsymbol{C}^{-1} = \Delta^{-2} + \boldsymbol{B}'\Psi^{-2}\boldsymbol{B}$. Clearly the implied regression of y_i on $\boldsymbol{x}_i$ is linear, with a theoretically implied and unique extension of the (low-dimensional) factor regression parameter $\boldsymbol{\theta}$ to the (high-dimensional) predictor regression parameter $\boldsymbol{\beta}$. Further, under the specific prior for $\boldsymbol{\theta}$, there is a unique (singular normal) prior implied for $\boldsymbol{\beta}$.

Consider now the special case in which $\Psi = s\boldsymbol{I}$, and in which identification is enforced by assuming $\boldsymbol{B}$ to be orthogonal. Then $\boldsymbol{C}$ is diagonal with elements $s^2d_i^2/(s^2 + d_i^2)$ where the d_i are the elements of the diagonal matrix Δ. Now take the limit as $s \to 0$,

so that the latent factors explain essentially all the variation in the predictors. Then (1) reduces to $\boldsymbol{x}_i = \boldsymbol{B}\boldsymbol{\lambda}_i$ or, in matrix form, $\boldsymbol{X} = \Lambda\boldsymbol{B}'$ where Λ has rows $\boldsymbol{\lambda}_i'$. Assuming $n \geq k$, this recovers the SVD decomposition of $\boldsymbol{X}$ with $\boldsymbol{B} = \boldsymbol{A}'$, and this limiting special case of the latent factor model defines the empirical factor model. In this limit, it also easily follows that $\boldsymbol{\beta} = \boldsymbol{A}'\boldsymbol{\theta}$. Under the chosen prior $\mathrm{N}(\boldsymbol{\theta} \,|\, 0, \boldsymbol{G}^{-1})$ with $\boldsymbol{G}$ diagonal, it follows that $\boldsymbol{\beta}$ has precisely the $gsg-$prior of Section 1.2.

Hence, this special limiting case of a formal latent factor model leads to the SVD regression and the $gsg-$prior. The tie-up is exact, and explains away the issues of design-, and sample size-, dependence of the parameter and prior, and of recovering inferences on the regression in the original predictor variables. Inference on $\boldsymbol{\beta}$ flows directly from that on $\boldsymbol{\theta}$. All predictor values for validation cases must be included in the analysis of training data as they inform, under (1), on parameters of the latent factor model and therefore, indirectly, on values of the latent factors underlying the training data.

3. SPARSE FACTOR MODELS

3.1 *Motivating Applications in Gene Expression Profiling*

Original motivation for this work comes from gene expression analysis in which predictors are genes and p may range up to 30,000. Some of our initial studies (Spang *et al* 2001; West *et al* 2000, 2001) involved binary regression, with a probit model constructed by treating the y_i as latent and observing only indicators of $y_i > 0$. Logistic and other variants are also standard extensions (Albert and Johnson, 1999, ch. 3). This development of generalised shrinkage priors with singular factors enabled the use of high-dimensional predictors in gene expression analysis and, in part, underlies the interest in more formal and flexible factor models.

In gene expression profiling, the predictor variables are recorded expression levels of individual genes, and the responses are clinical or physiological outcomes. Inherently, multiple biological factors underlie patterns of gene expression variation, so latent factor approaches are natural – we imagine that latent factors reflect individual biological functions (gene networks or pathways). This is also a motivating context for *sparse* models. Each biological factor involves a number of genes, perhaps a few to a few hundred, but not all genes; so each column of $\boldsymbol{B}$ will have many zeros. Similarly, a given gene may play roles in one or a small number of biological pathways, but will not be involved in all; so each row of $\boldsymbol{B}$ will have many zeros. It is therefore substantively appropriate to use priors that induce sparsity in $\boldsymbol{B}$.

3.2 *Bayesian Specification of Sparse Factor Models*

A Bayesian approach to defining sparse factor structure uses priors on the elements B_{ij} (gene i, factor j) of $\boldsymbol{B}$ that induce zeros with high probability. Within column (factor) j, take the B_{ij} to be independent with priors

$$\pi_j \delta_0(B_{ij}) + (1 - \pi_j)\mathrm{N}(B_{ij} \,|\, 0, 1)$$

where $\delta_0(\cdot)$ is the unit point mass at zero, and where π_j has a prior heavily concentrated near 1. Note that the unit scale of the normal component is convenient; the arbitrary scale of factor j is already accommodated in the variance parameter d_j^2. Assuming specified priors on all model parameters, MCMC analysis extends existing approaches that utilise normal priors on $\boldsymbol{B}$ (Aguilar and West, 2000; Lopes and West, 1999) to

incorporate these new mixture priors, and include sampling of the π_j. The analysis is inherently parallelisable; MCMC sequences through columns of $\boldsymbol{B}$ (i.e., through factors), and within each column the (many) elements B_{ij} are, *a posteriori*, conditionally independent given values of the prior and model parameters, latent factors and data. Hence, for fixed j, the set of p values B_{ij} $(i = 1, \ldots, p)$, may be sampled efficiently, in parallel. Some consideration of identification constraints is needed; we may use the popular lower triangular method (Aguilar and West, 2000; Lopes and West, 1999, and references therein) that simply fixes $k(k+1)/2$ selected elements of $\boldsymbol{B}$ at 0 or 1, and then use the mixture prior for the remaining (many) elements.

3.3 *Example: Factors in Breast Cancer Gene Expression Data*

Some illustration comes from analysis of expression levels of $p = 6128$ genes measured (using Affymetrix DNA microarrays) on $n = 49$ breast cancer tumour samples. The data comes from the study reported in West *et al* (2000, 2001) and Spang *et al* (2001), where full details of the data and context may be found. Here I discuss some aspects of a new analysis using the sparse latent factor model, with $k = 25$ factors. Additional prior specifications include priors on the variances d_j^2 of the latent factors and the idiosyncratic variances ψ_i^2 (the elements of the diagonal matrix Ψ); these are each specified via independent $\mathrm{Ga}(\cdot \,|\, 0.01, 0.01)$ priors on reciprocal variances. Finally, the use of a prior on π_j that heavily favours very high values is critical in enforcing very many zeros in the loadings matrix; here, with $p = 6128$ and $k = 25$, analysis utilises $\pi_j \sim \mathrm{Be}(\pi_j \,|\, 999, 1)$.

Figure 5 displays posterior means of the values of four of the factors, chosen as those four with largest values of the posterior means of the d_j^2 and plotted from the top down. The values of the factors (vertical axis) are plotted against sample number (horizontal axis). The first factor is essentially zero apart from on samples (tumours) 7,8,11 and 46, and the second essentially zero but for cases 7 and 8. These four cases had been much explored in earlier analyses, and, relative to most of the data, have quite apparent differences in large numbers of genes. The second factor shows that cases 7 and 8 share a common pattern of covariation not exhibited by 11 and 46. These four cases, and particularly 7 and 8, are in fact questionable due to concerns about the quality of the DNA microarray hybridisation; in analyses in West *et al* (2001) cases 7 and 8 had been held out due to these data quality concerns. It is of interest here to note that the full data analysis using the sparse factor model is itself capable of identifying these questionable cases, and protecting inferences on other factors from their effects.

The third factor plotted is of key interest in connection with comparison of oestrogen receptor (ER) status of tumours. We had earlier developed binary factor regression models to predictively discriminate ER status based on a selected set of about 50-100 genes (West *et al* 2000, 2001; Spang *et al* 2001). That analysis format is effective but involves the pre-selection of smaller numbers of discriminatory genes, and one motivating interest in sparse factor models is the potential to automate variable selection at the factor loading level. This potential is realised and evident here. The third factor is color coded: red indicates ER positive tumours, blue ER negative. Factor 3 evidently separates the two groups quite well, with four cases (16,31,40,43) in the mid-ground. Earlier analysis using gene screening had identified these four cases, and follow-on investigations reversed the ER status determination of number 31, so the analysis is in fact discriminating cases very well based on this factor alone.

For comparison, Figure 6 displays the four dominant empirical SVD factors (princi-

pal components) from the 6128 genes. Factors 1 and 2 bear resemblance to those from the model-based analysis, though the identification of questionable samples via two "outlier factors" is significantly obscured by the confounding of idiosyncratic noise in gene expression levels - a key inherent and limiting feature of empirical factor analyses with many variables. The third empirical factor certainly relates to the ER discrimination, but again the signal is obscured and much less clearly defined than it is in the model-based analysis.

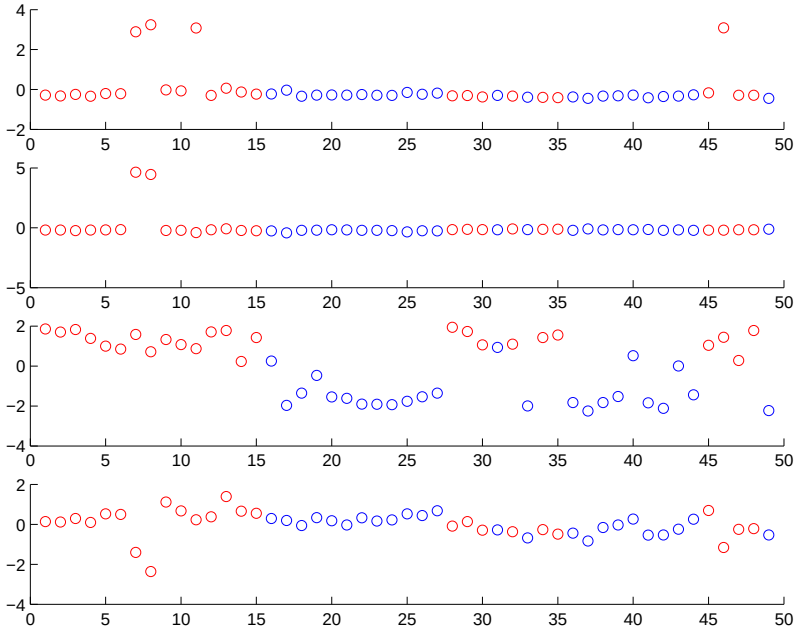

Figure 5. *Four factors in sparse factor analysis of gene expression data*

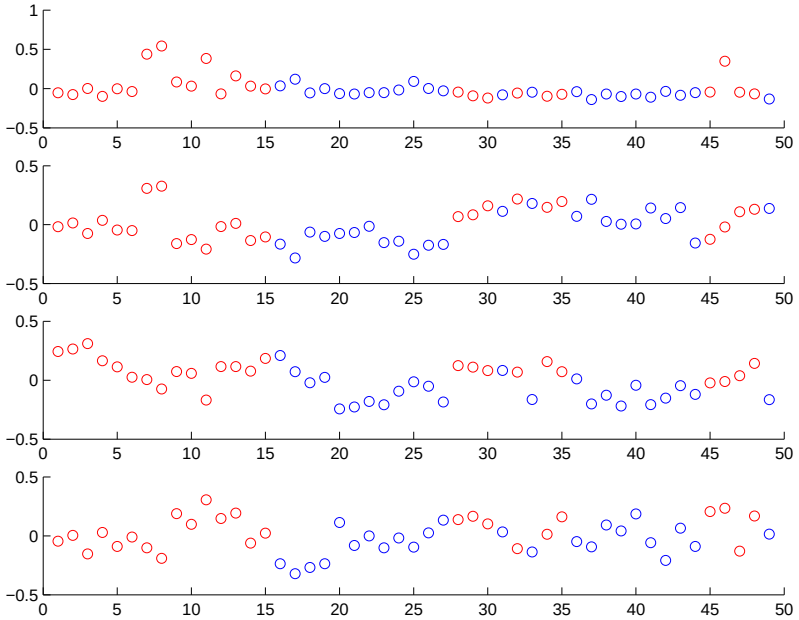

Figure 6. *Four factors in SVD of gene expression data*

4. ADDITIONAL COMMENTS

Sparse factor regression models offer a promising framework for dimension reduction in predictor space and for regression variable selection with many predictors. If a small number of latent factors associate with the response (as in the breast cancer ER study, where a single factor is primarily implicated) then a sparse factor model implies that only those genes with non-zero loadings on those factors are relevant; variable selection is then induced, automatically. In the ER study, only 60 genes have posterior probability of non-zero values exceeding 0.5; most of these genes show up in our prior studies and those of other groups exploring ER pathways in breast cancer.

Full analysis of the sparse factor model combined with binary regression has been explored in cross-validation studies of both ER and lymph node (LN) status with this breast cancer data, comparing with results in West *et al* (2001). This prior work uses gene selection/screening and SVD factor regression. The cross-validation predictions are very similar, perhaps even slightly better with the sparse factor model. I stress that this model uses all 6128 genes whereas our prior published analysis selects 100 based on correlation with ER or LN status, so removing noise via ad-hoc preliminary variable selection. The comparability of predictions is very strong evidence for the efficacy of the sparse modelling approach in dealing formally – and automatically – with that most critical and challenging variable selection problem. Further development and experience with this approach is needed; a key need is the development of efficient software for distributed processing to address the very challenging computational demands of model fitting.

Additional questions concern the incorporation of substantive, informative prior information into these large-scale models. This raises questions of both how flexible the current models are in terms of the scope for customising them to incorporate specific prior information, and of how they might be generalised. A further key question is the identification, or estimation, of the number of factors. In some studies, simply increasing k and exploring posterior estimates of factors and their variances suffices (Aguilar and West, 2000), though the problem remains an open research area and utilising formal approaches is a challenge (Lopes and West, 1999). This question is discussed also by Mackay and Miskin (2001). In the new sparse factor model introduced here, using too few factors confounds higher-order structure in the factors being estimated and, in particular, induces likelihood functions that very strongly suggest non-zero factor loadings for many gene-factor combinations than might be expected on scientific grounds; this seems to be simply an artifact of the use of too few factors. Though simple to diagnose, this problem is far from simple to resolve as fitting large numbers of factors is significantly challenging in terms of computation.

ACKNOWLEDGEMENTS

This research was partially supported by the NSF (grants DMS-0102227 and DMS-0112340). Some aspects of the work relate to collaborations with Ming Liao and Hedibert Lopes. I am grateful to Marina Vanucci for provision of the biscuit dough data, for useful conversations with Jim Berger, Merlise Clyde and Rainer Spang, and for comments from the editors and two referees.

REFERENCES

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

Aguilar, O. and West, M. (2000) Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics* **18**, 338-357.

Brown, P.J., Fearn, T. and Vannucci, M. (1999) The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika*, **86**, 635-648.

Lopes, H., and West, M. (1999) Model uncertainty in factor analysis. *ISDS Discussion Paper #98-38.* Submitted for publication.

Mackay, D.J.C., and Miskin, J. (2001) Latent variable models for gene expression data. Technical report: `www.inference.phy.ca.ac.uk/mackay`

Osborne, B.G., Fearn, T., Miller, A.R. and Douglas, S. (1984) Applications of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.,* **35**, 99-105.

Spang, R., Zuzan, H., West. M., Nevins, J.R., Blanchette, C. and Marks, J.R. (2001) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biology,* **2** (on-line publication: `http://www.bioinfo.de/isb/gcb01/talks/spang/index.html`).

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

West, M., Nevins, J.R., Marks, J.R, Spang, R. and Zuzan, H. (2000) DNA microarray data analysis and regression modeling for genetic expression profiling. ISDS Discussion Paper: `www.isds.duke.edu`

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), pp233-243. Amsterdam: North-Holland.