Model Averaging and Dimension Selection for the Singular Value Decomposition

Peter D Hoff

Model Averaging and Dimension Selection for the Singular Value Decomposition

Peter D. HOFF

Many multivariate data-analysis techniques for an $m \times n$ matrix $\mathbf{Y}$ are related to the model $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{Y}$ is an $m \times n$ matrix of full rank and $\mathbf{M}$ is an unobserved mean matrix of rank $K < (m \wedge n)$. Typically the rank of $\mathbf{M}$ is estimated in a heuristic way and then the least-squares estimate of $\mathbf{M}$ is obtained via the singular value decomposition of $\mathbf{Y}$, yielding an estimate that can have a very high variance. In this article we suggest a model-based alternative to the preceding approach by providing prior distributions and posterior estimation for the rank of $\mathbf{M}$ and the components of its singular value decomposition. In addition to providing more accurate inference, such an approach has the advantage of being extendable to more general data-analysis situations, such as inference in the presence of missing data and estimation in a generalized linear modeling framework.

KEY WORDS: Bayes factor; Factor analysis; Interaction; Multiple hypergeometric function; Multiplicative effects; Relational data; Social network; Steifel manifold.

1. INTRODUCTION

Many inferential and descriptive methods for multivariate and matrix-valued data are variations on the idea of modeling the data matrix $\mathbf{Y}$ as equal to a reduced-rank mean matrix $\mathbf{M}$ plus Gaussian noise, and estimating $\mathbf{M}$ after deciding on its rank. A concept that is central to such models is the singular value decomposition: every $m \times n$ matrix $\mathbf{M}$ has a representation of the form $\mathbf{M} = \mathbf{UDV}'$, where, in the case $m \geq n$, $\mathbf{U}$ is an $m \times n$ matrix with orthonormal columns, $\mathbf{V}$ is an $n \times n$ matrix with orthonormal columns, and $\mathbf{D}$ is an $n \times n$ diagonal matrix, with diagonal elements $\{d_1, \ldots, d_n\}$ typically taken to be a decreasing sequence of nonnegative numbers. The triple $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ is called the singular value decomposition of $\mathbf{M}$. The squared elements of the diagonal of $\mathbf{D}$ are the eigenvalues of $\mathbf{M}'\mathbf{M}$ and the columns of $\mathbf{V}$ are the corresponding eigenvectors. The matrix $\mathbf{U}$ can be obtained from the first n eigenvectors of $\mathbf{MM}'$. The number of nonzero elements of $\mathbf{D}$ is the rank of $\mathbf{M}$.

The appeal of the singular value decomposition is partly due to its interpretation as a multiplicative model based on row and column factors. Given a model of the form $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, the elements of $\mathbf{Y}$ can be written $y_{i,j} = \mathbf{u}_i' \mathbf{D} \mathbf{v}_j + e_{i,j}$, where $\mathbf{u}_i$ and $\mathbf{v}_j$ are the ith and jth rows of $\mathbf{U}$ and $\mathbf{V}$, respectively. This model thus provides a representation of the systematic variation among the entries of $\mathbf{Y}$ by row and column factors. Models of this type play a role in the analysis of relational data (Harshman, Green, Wind, and Lundy 1982), in biplots (Gabriel 1971; Gower and Hand 1996), and in reduced-rank interaction models for factorial designs (Gabriel 1978, 1998). The model is also closely related to factor analysis, in which the row vectors of $\mathbf{Y}$ are modeled as iid samples from the model $\mathbf{y}_i = \mathbf{u}_i \mathbf{DV}' + \mathbf{e}_i$. In this situation, the goal is typically to represent the covariance across the n columns by their relationship to $K < n$ unobserved latent factors. Lawley and Maxwell (1971) provided a comprehensive overview of factor analysis models, as well as methods of maximum-likelihood estimation for model parameters.

Singular value decomposition also plays a role in parameter estimation for the reduced-rank model: Assuming the rank of the mean matrix $\mathbf{M}$ is $K < n$ and letting $(\hat{\mathbf{U}}, \hat{\mathbf{D}}, \hat{\mathbf{V}})$ be the singular value decomposition of the data matrix $\mathbf{Y}$, the least-squares estimate of $\mathbf{M}$ (and maximum-likelihood estimate under Gaussian noise) is given by $\hat{\mathbf{M}}_K = \hat{\mathbf{U}}_{[,1:K]} \hat{\mathbf{D}}_{[1:K,1:K]} \hat{\mathbf{V}}'_{[,1:K]}$, where $\hat{\mathbf{U}}_{[,1:K]}$ denotes the first K columns of $\hat{\mathbf{U}}$, and where $\hat{\mathbf{D}}_{[1:K,1:K]}$ denotes the first K rows and columns of $\hat{\mathbf{D}}$ (Householder and Young 1938; Gabriel 1978). In applications such as signal processing, image analysis, and, more recently, large-scale gene expression data, representing a noisy data matrix $\mathbf{Y}$ by $\hat{\mathbf{M}}_K$ with $K \ll n$ has the effect of capturing the main patterns of $\mathbf{Y}$ while eliminating much of the noise.

Despite its utility and simplicity, several issues limit the use of the singular value decomposition as an estimation procedure. The first is that the rank K of the approximating mean matrix $\hat{\mathbf{M}}_K$ must be specified. Standard practice is to plot the singular values of $\mathbf{Y}$ in decreasing order and then select K to be the index where the last "large gap" or "elbow" occurs. The second issue is that, even if the rank is chosen correctly, one may be concerned with its variance: For high-dimensional parameter spaces, the mean squared error of least-squares estimates is often higher than that of penalized or Bayes estimates. Finally, non-model-based approaches are somewhat limited in their ability to handle missing or nonnormal data.

Philosophical debates aside, Bayesian methods often provide sensible procedures for model selection and high-dimensional parameter estimation. For the model described in the preceding text, a Bayesian procedure would provide a mapping from a prior distribution $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2)$ to a posterior distribution $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2 | \mathbf{Y})$. Of primary interest might be functions of this posterior distribution, such as $E(\mathbf{M}|\mathbf{Y})$ or the marginal posterior distribution of the rank $p(K|\mathbf{Y}) \propto p(K) p(\mathbf{Y}|K)$. Both of these quantities require integration over the complicated, high-dimensional space of $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ for each value of K. In the related factor analysis model, where the elements of $\mathbf{U}$ are modeled as independent normal random variables, the difficulty in calculating marginal probabilities has led to the development of approximate Bayesian procedures: Rajan and Rayner (1997) provided a coarse approximation to the marginal probability $p(\mathbf{Y}|K)$ by plugging in maximum-likelihood estimates. Minka (2000) improved on this by providing a Laplace approximation

to the desired marginal probability. Both of these procedures rely on asymptotic approximations and do not provide Bayesian estimates of $\mathbf{M}$ once the dimension has been selected.

Alternatively, one could turn to Markov chain Monte Carlo (MCMC) methods to obtain approximations to the integrals that are necessary for model selection. Tierney (1994) described the theory of MCMC for very general parameter spaces, and Green (1995) outlined conditions under which the reversibility of the Metropolis–Hastings algorithm is maintained for model-selection problems. However, obtaining efficient proposal distributions for high-dimensional problems is a difficult and delicate art. For the factor analysis problem, Lopes and West (2004) devised a workable reversible-jump algorithm by constructing proposal distributions that approximate the within-model full conditional distributions of the model parameters. In this way, their approach mimics aspects of the Gibbs sampler. However, their approximate full conditional distributions are derived from auxiliary within-model MCMC algorithms, requiring an extra Markov chain for each rank K to be considered. As a result, this approach requires prespecification of the values of K, leaving open the possibility that computational effort is spent on values of K with negligible posterior probability or that values of K with high posterior probability are overlooked altogether.

In the analysis of relational data such as social and biological networks, the row heterogeneity and column heterogeneity of $\mathbf{Y}$ are of equal interest, thus motivating a singular value decomposition approach to the data analysis as opposed to a factor analysis. The purpose of this article is to provide the necessary calculations for model selection, estimation, and inference for statistical models based on the singular value decomposition. The results in the following sections provide a means of Markov chain Monte Carlo approximation to the posterior distribution of the rank and values of the matrix $\mathbf{M}$. This MCMC algorithm is based on Gibbs sampling and, unlike the algorithm of Lopes and West (2004), requires no auxiliary runs or prespecification of the values of K. Additionally, this model-based method allows for estimation in the presence of missing data or replications, and can be incorporated into a generalized linear modeling framework to allow for the analysis of a variety of data types. For example, Section 6 discusses a model extension that allows for the analysis and prediction of binary relational data such as social or biological networks.

In the next section we discuss prior distributions for $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$ given a fixed rank K and show how the uniform distribution for $\mathbf{U}$ (the invariant measure on the Steifel manifold) can be specified in terms of the full conditional distributions of its column vectors. Section 3 presents a Gibbs sampling scheme for parameter estimation when the rank of $\mathbf{M}$ is specified. In the case of unspecified rank, estimation can be achieved via a prior distribution that allows the diagonal elements of $\mathbf{D}$ each to be zero with nonzero probability. In Section 4 we consider posterior inference under such a prior distribution, and develop a Markov chain Monte Carlo algorithm that moves between models with different ranks. This algorithm is constructed via a Gibbs sampling scheme, which samples each singular value d_j from its conditional distribution. This is done marginally over $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$, and requires a complicated but manageable integration. Section 5 presents a small simulation study that examines the sampling properties of the Bayesian procedure. It is

shown that the procedure is able to estimate the true rank of $\mathbf{M}$ reasonably well for a variety of matrix sizes, and the squared error of the Bayes estimate $E(\mathbf{M}|\mathbf{Y})$ is typically much lower than that of the least-squares estimator. In Section 6 the singular value decomposition model is extended to accommodate nonnormal data via a generalized linear modeling framework, which is then used in an example analysis of binary relational data. A discussion follows in Section 7.

2. THE SINGULAR VALUE DECOMPOSITION MODEL AND PRIOR DISTRIBUTIONS

As previously described, our model for an $m \times n$ data matrix is $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{M}$ is a rank-K matrix and $\mathbf{E}$ is a matrix of iid mean-zero normally distributed noise. We induce a prior distribution on the matrix $\mathbf{M}$ by way of a prior distribution on the components of its singular value decomposition $\{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$.

For a given rank K, we can take $\mathbf{U}$ to be an $m \times K$ orthonormal matrix. The set of such matrices is called the Steifel manifold and is denoted $\mathcal{V}_{K,m}$. A natural, noninformative prior distribution for $\mathbf{U}$ is the uniform distribution on $\mathcal{V}_{K,m}$, which is the unique probability measure on $\mathcal{V}_{K,m}$ that is invariant under left- and right-orthogonal transformations. As discussed in Chikuse (2003, sec. 2.5), a sample $\mathbf{U}$ from the uniform distribution on the Steifel manifold $\mathcal{V}_{K,m}$ can be obtained by first sampling an $m \times K$ matrix $\mathbf{X}$ of independent standard normal random variables and then setting $\mathbf{U} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$. Although this construction is straightforward, it does not explicitly specify conditional distributions of the form $p(\mathbf{U}_{[,j]}|\mathbf{U}_{[,-j]})$, which are quantities that will be required for the estimation procedure outlined in Section 3. We now derive these conditional distributions via a new iterative method of generating samples from the uniform distribution on $\mathcal{V}_{K,m}$.

Let $\mathbf{U}_{[,A]}$ denote the columns of $\mathbf{U}$ that correspond to a subset of column labels $A \subset \{1, \ldots, K\}$ and let $\mathbf{N}_A$ be any $m \times (m - |A|)$ matrix whose columns form an orthonormal basis for the null space of $\mathbf{U}_{[,A]}$. A random $\mathbf{U} \in \mathcal{V}_{K,m}$ can be constructed as follows:

1. Sample $\mathbf{u}_1$ uniformly from the unit m sphere and set $\mathbf{U}_{[,1]} = \mathbf{u}_1$.
2. Sample $\mathbf{u}_2$ uniformly from the unit $(m - 1)$ sphere and set $\mathbf{U}_{[,2]} = \mathbf{N}_{\{1\}}\mathbf{u}_2$.

$\vdots$

K. Sample $\mathbf{u}_K$ uniformly from the unit $(m - K + 1)$ sphere and set $\mathbf{U}_{[,K]} = \mathbf{N}_{\{1,\ldots,K-1\}}\mathbf{u}_K$.

By construction this procedure generates an $m \times K$ matrix $\mathbf{U}$ that has orthonormal columns. The following result also holds:

Proposition 1. The probability distribution of $\mathbf{U}$ is the uniform probability measure on $\mathcal{V}_{K,m}$.

A proof is provided in the Appendix. Because this probability distribution is invariant under left- and right-orthogonal transformations of $\mathbf{U}$ (see, e.g., Chikuse 2003), it follows that the rows and columns of $\mathbf{U}$ are exchangeable. As a result, the conditional distribution of $\mathbf{U}_{[,j]}$ given any subset A of columns of $\mathbf{U}$ is equal to the distribution of $\mathbf{N}_A\mathbf{u}_j$, where $\mathbf{u}_j$ is uniformly distributed on the $(m - |A|)$ sphere. This fact facilitates the Gibbs sampling of the columns of $\mathbf{U}$ and $\mathbf{V}$ from their full conditional

$$\text{uniform}(\mathcal{V}_{K,m}) \sim \mathbf{U}$$

$$\text{normal}(\mu, 1/\psi) \sim \mathbf{D} \longrightarrow \{\mathbf{Y} = \mathbf{UDV}' + \mathbf{E}\} \longleftarrow \mathbf{E} \sim \text{normal}\,(0, 1/\phi)$$

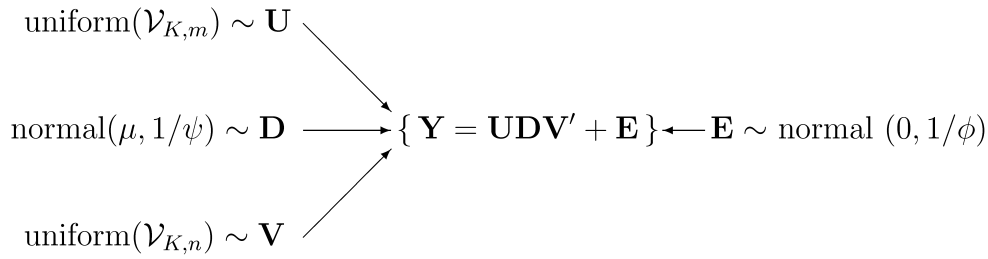$$\text{uniform}(\mathcal{V}_{K,n}) \sim \mathbf{V}$$

Figure 1. A Graphical Representation of the Model.

distributions, as described in Section 3. For simplicity we proceed with posterior calculations using this uniform prior, even though the full conditionals derived in Section 3 indicate that a more general, informative conjugate family could be used with no additional computational difficulty.

For a given rank K, the nonzero singular values $\{d_1, \ldots, d_K\}$ that make up the diagonal of $\mathbf{D}$ determine the magnitude of the mean matrix, in that $\|\mathbf{M}\|^2 = \sum_{k=1}^{K} d_k^2$. We model these nonzero values as being samples from a normal population with mean μ and precision (inverse variance) ψ. For reasons of conciseness, we discuss posterior calculations only for conjugate prior distributions on these parameters, which include a normal distribution with mean μ_0 and variance v_0^2 for μ, and a gamma$(\eta_0/2, \eta_0\tau_0^2/2)$ distribution for ψ, parameterized so that ψ has expectation $1/\tau_0^2$. Other potentially useful prior distributions, along with choices for hyperparameters, are discussed in Section 5. This parameterization of the singular values differs slightly from that of the usual singular value decomposition, in that here the values $\{d_1, \ldots, d_K\}$ are not restricted to be nonnegative. A model that enforces this restriction is possible, but adds a small amount of computational difficulty without any modeling benefit [if $\mathbf{A}$ is a diagonal matrix of ± 1's, then $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}) = p(\mathbf{Y}|\mathbf{UA}, \mathbf{AD}, \mathbf{V})$]. Finally, the elements of $\mathbf{E}$ are modeled as iid normal random variables with mean zero and variance $1/\phi$. The prior distribution for the precision ϕ is taken to be gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$. A graphical representation of the model and parameters is given in Figure 1. Choices for hyperparameters $\{(\mu_0, v_0^2), (\eta_0, \tau_0^2), (\nu_0, \sigma_0^2)\}$ are discussed in Section 5.

3. GIBBS SAMPLING FOR THE FIXED–RANK MODEL

A Markov chain with $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \mu, \psi|\mathbf{Y}, K)$ as its stationary distribution can be constructed via a Gibbs sampling procedure, which iteratively samples ϕ, μ, ψ and the columns of $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ from their full conditional distributions. These samples can be used to approximate the joint posterior distribution and estimate posterior quantities of interest (see, e.g., Tierney 1994).

The full conditional distributions for ϕ, μ, ψ and the elements of $\mathbf{D}$ are standard and are provided subsequently without derivation. Less standard are the full conditional distributions of the columns of $\mathbf{U}$ and $\mathbf{V}$. To derive these, consider the form of $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi)$ as a function of $\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}$, and $d_j \equiv \mathbf{D}_{[j,j]}$. Letting $\mathbf{E}_{-j} = \mathbf{Y} - \mathbf{U}_{[,-j]}\mathbf{D}_{[-j,-j]}\mathbf{V}'_{[,-j]}$, we have

$$\|\mathbf{Y} - \mathbf{UDV}'\|^2$$

$$= \left\|\mathbf{E}_{-j} - d_j\mathbf{U}_{[,j]}\mathbf{V}'_{[,j]}\right\|^2$$

$$= \|\mathbf{E}_{-j}\|^2 - 2d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} + \left\|d_j\mathbf{U}_{[,j]}\mathbf{V}'_{[,j]}\right\|^2$$

$$= \|\mathbf{E}_{-j}\|^2 - 2d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} + d_j^2.$$

It follows that $p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi)$ can be written

$$p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi) = \left(\frac{\phi}{2\pi}\right)^{mn/2} \exp\left\{-\frac{1}{2}\phi\|\mathbf{E}_{-j}\|^2 + \phi d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]} - \frac{1}{2}\phi d_j^2\right\}. \quad (1)$$

Recall that, conditional on $\mathbf{U}_{[,-j]}$, $\mathbf{U}_{[,j]} \overset{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$, where $\mathbf{N}^u_{\{-j\}}$ is a basis for the null space of columns of $\mathbf{U}_{[,-j]}$ and $\mathbf{u}_j$ is uniform on the $m - (K - 1)$ sphere. From (1), we see that the full conditional distribution of $\mathbf{u}_j$ is proportional to $\exp\{\mathbf{u}'_j(\phi d_j\mathbf{N}^{u'}_{\{-j\}}\mathbf{E}_{-j}\mathbf{V}_{[,j]})\}$. This is a von Mises–Fisher distribution on the $m - (K - 1)$ sphere with parameter $\phi d_j\mathbf{N}^{u'}_{\{-j\}}\mathbf{E}_{-j}\mathbf{V}_{[,j]}$. A sample of $\mathbf{U}_{[,j]}$ from its full conditional distribution can therefore be generated by sampling $\mathbf{u}_j$ from the von Mises–Fisher distribution and then setting $\mathbf{U}_{[,j]} = \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$. The full conditional distribution of $\mathbf{V}_{[,j]}$ is derived similarly. In general, the von Mises–Fisher distribution on the p sphere with parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ has density $c_p(\|\boldsymbol{\mu}\|)\exp\{\mathbf{u}'\boldsymbol{\mu}\}$ and is denoted vMF$(\boldsymbol{\mu})$, and the uniform distribution on the sphere is denoted vMF$(\mathbf{0})$. The normalizing constants for these two cases are

$$c_p(\kappa) = (2\pi)^{-p/2}\frac{\kappa^{p/2-1}}{I_{p/2-1}(\kappa)} \quad \text{for } \kappa > 0,$$

$$c_p(0) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \quad \text{for } \kappa = 0,$$

where $I_\nu(x)$ is the modified Bessel function of the first kind (see sec. 9.6 of Abramowitz and Stegun 1972). R code for sampling from this distribution is provided at my website.

Summarizing these results, a Markov chain with the desired stationary distribution can be constructed by iterating the following procedure:

1. For $j \in \{1, \ldots, K\}$,

   - Sample $(\mathbf{U}_{[,j]}|\mathbf{Y}, \mathbf{U}_{[,-j]}, \mathbf{D}, \mathbf{V}, \phi) \overset{d}{=} \mathbf{N}^u_{\{-j\}}\mathbf{u}_j$, where $\mathbf{u}_j \sim \text{vMF}(\phi d_j\mathbf{N}^{u'}_{-j}\mathbf{E}_{-j}\mathbf{V}_{[,j]})$.
   - Sample $(\mathbf{V}_{[,j]}|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}_{[,-j]}, \phi) \overset{d}{=} \mathbf{N}^v_{\{-j\}}\mathbf{v}_j$, where $\mathbf{v}_j \sim \text{vMF}(\phi d_j\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{N}^v_{-j})$.
   - Sample $(d_j|\mathbf{Y}, \mathbf{U}, \mathbf{D}_{[-j,-j]}, \mathbf{V}, \phi, \mu, \psi) \sim$ normal $[(\mathbf{U}'_{[,j]}\mathbf{E}_{-j}\mathbf{V}_{[,j]}\phi + \mu\psi)/(\phi + \psi), 1/(\phi + \psi)]$.

2. Sample $(\phi|\mathbf{Y}, \mathbf{U}, \mathbf{D}, \mathbf{V}) \sim$ gamma$[(\nu_0 + mn)/2, (\nu_0\sigma_0^2 + \|\mathbf{Y} - \mathbf{UDV}'\|^2)/2]$.

3. Sample $(\mu|\mathbf{D}, \psi) \sim \text{normal}[(\psi \sum d_j + \mu_0/v_0^2)/(\psi K + 1/v_0^2), 1/(\psi K + 1/v_0^2)]$.

4. Sample $(\psi|\mathbf{D}, \mu) \sim \text{gamma}[(\eta_0 + K)/2, (\eta_0 \tau_0^2 + \sum (d_j - \mu)^2)/2]$.

4. THE VARIABLE–RANK MODEL

4.1 Prior Distributions

In this section we extend the model of Section 2 to the case where the rank K is to be estimated. This requires comparisons between models with parameter spaces of different dimensions. Two standard ways to view such problems are as follows:

1. Conceptualize a different parameter space for each value of K, that is, conditional on K, the mean matrix is $\mathbf{UDV}'$, where the dimensions of $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ are $m \times K$, $K \times K$, and $n \times K$, respectively.

2. Parameterize $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ to be of dimensions $m \times n$, $n \times n$, and $n \times n$, but allow for columns of these matrices to be identically zero. In this parameterization, $K = \sum_{j=1}^{n} \mathbb{1}(d_j \neq 0)$.

Each of these approaches has its own notational and conceptual hurdles, and which one to present is, to some extent, a matter of style (see Green 2003 for a discussion). Given a prior distribution on K, the first approach can be formulated by using the prior distributions of Section 2 as the conditional distributions of $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ given K. The second approach can be made equivalent to the first as follows:

1. Let $\tilde{\mathbf{U}}, \tilde{\mathbf{D}}$, and $\tilde{\mathbf{V}}$ have the prior distributions described in Section 2 with $\tilde{K} = n$.
2. Let $\{s_1, \ldots, s_n\} \sim p(K = \sum s_j) \times \binom{n}{\sum s_j}^{-1}$, where each $s_j \in \{0, 1\}$.
3. Let $\mathbf{S} = \text{diag}\{s_1, \ldots, s_n\}$. Set $\mathbf{U} = \tilde{\mathbf{U}}\mathbf{S}$, $\mathbf{D} = \tilde{\mathbf{D}}\mathbf{S}$, $\mathbf{V} = \tilde{\mathbf{V}}\mathbf{S}$, and $K = \sum s_j$.

Parameterizing a set of nested models with binary variables has been a useful technique in a variety of contexts, including variable selection in regression models (Mitchell and Beauchamp 1988). We continue with this formulation because it allows for the construction of a relatively straightforward Gibbs sampling scheme to generate samples from the posterior distribution.

The matrices $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ described in the immediateley preceding items 1, 2, and 3 are exchangeable under simultaneous permutation of their columns. It follows from Proposition 1 that, conditional on $s_1, \ldots, s_n$, the nonzero columns of $\mathbf{U}$ and $\mathbf{V}$ are random samples from the uniform distributions on $\mathcal{V}_{\sum s_j, m}$ and $\mathcal{V}_{\sum s_j, n}$ respectively, and that, conditional on $\{s_j = 1, \mathbf{U}_{[, -j]}, \mathbf{V}_{[, -j]}\}$,

$$\mathbf{U}_{[, j]} \stackrel{d}{=} \mathbf{N}_{\{-j\}}^u \mathbf{u}, \qquad \mathbf{V}_{[, j]} \stackrel{d}{=} \mathbf{N}_{\{-j\}}^v \mathbf{v},$$

where $\mathbf{N}_{\{-j\}}^u$ and $\mathbf{N}_{\{-j\}}^v$ are orthonormal bases for the null spaces of $\mathbf{U}_{[, -j]}$ and $\mathbf{V}_{[, -j]}$, and $\mathbf{u}$ and $\mathbf{v}$ are uniformly distributed on the $(m - \sum s_j + 1)$ and $(n - \sum s_j + 1)$ spheres. This property will facilitate posterior sampling of the columns of $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$, as described in the next subsection.

4.2 Posterior Estimation

Let $\mathbf{\Theta} = \{\mathbf{U}, \mathbf{D}, \mathbf{V}\}$, $\mathbf{\Theta}_j = \{\mathbf{U}_{[, j]}, d_j, \mathbf{V}_{[, j]}\}$, and $\mathbf{\Theta}_{-j} = \{\mathbf{\Theta}_k : k \neq j\}$. In this subsection, we derive the full conditional distribution of $\mathbf{\Theta}_j$ given $\{\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi\}$ under the model described in the previous subsection. The prior and full conditional distributions of ϕ, μ, and ψ remain unchanged from Section 2. The full conditional distributions can be used in a Gibbs sampling scheme to generate approximate samples from $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \mu, \psi|\mathbf{Y})$.

Under the model and the parameterization previously described, the components of $\mathbf{\Theta}_j$ are either all zero or have a distribution as described in Section 2. To sample $\mathbf{\Theta}_j$, we first sample whether or not the components are zero, and if not, use sample the nonzero values. More specifically, sampling $\mathbf{\Theta}_j$ from its full conditional distribution can be achieved as follows:

1. Sample from $(\{d_j = 0\}, \{d_j \neq 0\})$ conditional on $\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi$.
2. If $\{d_j = 0\}$ is true, then set $d_j, \mathbf{U}_{[, j]}$, and $\mathbf{V}_{[, j]}$ all equal to zero.
3. If $\{d_j \neq 0\}$ is true,
   (a) sample $d_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi, \{d_j \neq 0\}$;
   (b) sample $\{\mathbf{U}_{[, j]}, \mathbf{V}_{[, j]}\}|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, d_j$.

The foregoing steps 1, 2, and 3 constitute a draw from $p(\mathbf{\Theta}_j|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi)$. The first step requires calculation of the odds:

$$\text{odds}(d_j \neq 0|\mathbf{Y}, \mathbf{\Theta}_{-j}, \phi, \mu, \psi)$$
$$= \frac{p(d_j \neq 0|\mathbf{\Theta}_{-j})}{p(d_j = 0|\mathbf{\Theta}_{-j})} \frac{p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j \neq 0, \phi, \mu, \psi)}{p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j = 0, \phi, \mu, \psi)}. \quad (2)$$

The first ratio is simply the prior conditional odds of $\{d_j \neq 0\}$ and can be derived from the prior distribution on the rank K. The second term in (2) can be viewed as a Bayes factor, evaluating the evidence in the data for additional structure in $E[\mathbf{Y}]$ beyond that provided by $\mathbf{\Theta}_{-j}$. Recall from the previous section that $\mathbf{Y} - \mathbf{UDV}' = \mathbf{E}_{-j} - d_j \mathbf{U}_{[, j]} \mathbf{V}'_{[, j]}$, and so we can write

$$p(\mathbf{Y}|\mathbf{U}, \mathbf{D}, \mathbf{V}, \phi, \mu, \psi)$$
$$= \left[\left(\frac{\phi}{2\pi}\right)^{mn/2} \exp\left\{-\frac{1}{2}\phi\|\mathbf{E}_{-j}\|^2\right\} \right]$$
$$\times \exp\left\{-\frac{1}{2}\phi d_j^2\right\} \exp\left\{\phi d_j \mathbf{U}'_{[, j]}\mathbf{E}_{-j}\mathbf{V}_{[, j]}\right\}$$
$$= p(\mathbf{Y}|\mathbf{\Theta}_{-j}, d_j = 0, \phi, \mu, \psi)$$
$$\times \exp\left\{-\frac{1}{2}\phi d_j^2\right\} \exp\left\{\phi d_j \mathbf{U}'_{[, j]}\mathbf{E}_{-j}\mathbf{V}_{[, j]}\right\}. \quad (3)$$

The first term in (3) is equal to the denominator of the Bayes factor, and is simply a product of normal densities, where the elements of $\mathbf{Y}$ have means given by $\mathbf{U}_{[, -j]}\mathbf{D}_{[-j, -j]}\mathbf{V}'_{[, -j]}$ and equal variances $1/\phi$. The numerator of the Bayes factor can be obtained by integrating (3) over $\mathbf{\Theta}_j$ with respect to its conditional distribution given $\mu, \psi, \mathbf{\Theta}_{-j}$, and $\{d_j \neq 0\}$. Integrating first with respect to $\mathbf{U}_{[, j]}, \mathbf{V}_{[, j]}$, we need to calculate $E[\exp\{\phi d_j \mathbf{U}'_{[, j]}\mathbf{E}_{-j}\mathbf{V}_{[, j]}\}|\mathbf{\Theta}_{-j}, d_j]$. Let $\tilde{m} = m - \sum_{k \neq j}\{d_k \neq 0\}$ and $\tilde{n} = n - \sum_{k \neq j}\{d_k \neq 0\}$. Recall that, conditional on $\mathbf{\Theta}_{-j}$, $\mathbf{U}_{[, j]} \stackrel{d}{=} \mathbf{N}_{\{-j\}}^u \mathbf{u}$ and $\mathbf{V}_{[, j]} \stackrel{d}{=} \mathbf{N}_{\{-j\}}^v \mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$

are uniformly distributed on the $\tilde{m}$ and $\tilde{n}$ spheres. Letting $\tilde{\mathbf{E}} = \mathbf{N}^{u\prime}_{\{-j\}}\mathbf{E}_{-j}\mathbf{N}^v_{\{-j\}}$, the required expectation can therefore be rewritten as $E_{\mathbf{uv}}[\exp\{\phi d_j \mathbf{u}'\tilde{\mathbf{E}}\mathbf{v}\}]$. This expectation is nonstandard and is derived in the Appendix. The result gives

$$p(\mathbf{Y}|\boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j)$$
$$= p(\mathbf{Y}|\boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j = 0)$$
$$\times \exp\left\{-\frac{1}{2}\phi d_j^2\right\} \sum_{l=0}^{\infty} \|\tilde{\mathbf{E}}\|^{2l}\phi^{2l}d_j^{2l}a_l, \qquad (4)$$

where the sequence $\{a_l\}_0^{\infty}$ can be computed exactly and is given in the Appendix.

The calculation of $p(\mathbf{Y}|\boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j \neq 0)$ is completed by integrating (4) over d_j with respect to $p(d_j|\boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j \neq 0)$, the normal density with mean μ and precision ψ. This integration simply requires calculating the even moments of a normal distribution, resulting in

$$p(\mathbf{Y}|\boldsymbol{\Theta}_{-j}, \phi, \psi, d_j \neq 0)$$
$$= p(\mathbf{Y}|\boldsymbol{\Theta}_{-j}, \phi, \psi, d_j = 0)\sum_{l=0}^{\infty}\|\tilde{\mathbf{E}}\|^{2l}a_l b_l, \quad (5)$$

where the sequence $\{b_l\}_0^{\infty}$ is given by

$$b_l = \phi^{2l}\left(\frac{\psi}{\phi+\psi}\right)^{1/2}\exp\left\{-\frac{1}{2}\frac{\mu^2\psi\phi}{\phi+\psi}\right\}$$
$$\times E\left[\left\{\frac{1}{\sqrt{\phi+\psi}}\left(Z + \frac{\mu\psi}{\phi+\psi}\right)\right\}^{2l}\right],$$

where Z is standard normal. The required moments can be calculated iteratively; see, for example, Smith (1995). The conditional odds of $\{d_j \neq 0\}$ is, therefore,

$$\text{odds}(d_j \neq 0|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi)$$
$$= \frac{p(d_j \neq 0|\boldsymbol{\Theta}_{-j})}{p(d_j = 0|\boldsymbol{\Theta}_{-j})}\sum_{l=0}^{\infty}\|\tilde{\mathbf{E}}\|^{2l}a_l b_l.$$

In practice, only a finite number of terms can be used to compute the preceding sums. The sum in (4) can be bounded above and below by modified Bessel functions, and the error in a finite-sum approximation can be bounded. This can also provide a guide as to how many terms to include in the approximation of (5). Details are given in the Appendix.

If $\{d_j \neq 0\}$ is sampled, it is still necessary to sample d_j, $\mathbf{U}_{[,j]}$, and $\mathbf{V}_{[,j]}$. Multiplying (4) by the prior for $d_j|\{d_j \neq 0\}$ shows that the required conditional distribution for d_j is proportional to

$$p(d_j|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, \{d_j \neq 0\})$$
$$\propto \exp\left(-\frac{1}{2}(d_j-\mu)^2\psi\right)\exp\left(-\frac{1}{2}d_j^2\phi\right)\sum_{l=0}^{\infty}\|\tilde{\mathbf{E}}\|^{2l}\phi^{2l}d_j^{2l}a_l,$$

which is an infinite mixture with the following components:

- mixture weights $w_l \propto \|\tilde{\mathbf{E}}\|^{2l}a_l b_l$
- mixture densities $f_l(d) \propto d^{2l}\exp\{-\frac{1}{2}(d-\tilde{\mu})^2\tilde{\psi}\}$, where $\tilde{\mu} = \mu\psi/(\phi+\psi)$ and $\tilde{\psi} = \phi+\psi$.

The density $f_l(d)$ is nonstandard, but can be sampled from quite efficiently using rejection sampling with a scaled and shifted t distribution as the approximating density (the tails of a normal distribution are not heavy enough).

To sample $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$, we first sample $\mathbf{u}$ and $\mathbf{v}$ from their joint distribution, and then set $\mathbf{U}_{[,j]} = \mathbf{N}^u_{\{-j\}}\mathbf{u}$ and $\mathbf{V}_{[,j]} = \mathbf{N}^v_{\{-j\}}\mathbf{v}$. Equation (3) indicates that the joint conditional density of $\{\mathbf{u}, \mathbf{v}\}$ is of the form

$$p(\mathbf{u}, \mathbf{v}|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j) = c(\mathbf{A})\exp\{\mathbf{u}'\mathbf{A}\mathbf{v}\}, \quad (6)$$

where $\mathbf{A} = \phi d_j\tilde{\mathbf{E}}$ and $c(\mathbf{A})^{-1} = c_{\tilde{m}}(0)^{-1}c_{\tilde{n}}(0)^{-1}\sum_{l=0}^{\infty}\|\mathbf{A}\|^{2l}\times a_l$. This density defines a joint distribution for two dependent unit vectors. A similar distribution for dependent unit vectors was discussed by Jupp and Mardia (1980), except their vectors were of the same length and the matrix $\mathbf{A}$ was assumed to be orthogonal. Some useful facts about the distribution in (6) include:

- The conditional distribution of $\mathbf{u}|\mathbf{v}$ is vMF($\mathbf{A}\mathbf{v}$) and that of $\mathbf{v}|\mathbf{u}$ is vMF($\mathbf{u}'\mathbf{A}$).
- The marginal distribution of $\mathbf{v}$ is proportional to $I_{\tilde{m}/2-1}(\|\mathbf{A}\mathbf{v}\|)/\|\mathbf{A}\mathbf{v}\|^{\tilde{m}/2-1}$.
- The joint density has local maxima at $\{\pm(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k), k = 1, \ldots, \tilde{n}\}$, where $(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k)$ are the kth singular vectors of $\mathbf{A}$.

I have implemented a number of rejection and importance samplers for this distribution, although making these schemes efficient is still a work in progress. A relatively fast approximate method that seems to work well for a variety of matrices $\mathbf{A}$ is first to sample $(\mathbf{u}, \mathbf{v})$ from the local modes $\{\pm(\hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k), k = 1, \ldots, \tilde{n}\}$ according to the exact relative probabilities and then use this value as a starting point for a small number of Gibbs samples, alternately sampling from $p(\mathbf{u}|\mathbf{A}, \mathbf{v})$ and $p(\mathbf{v}|\mathbf{A}, \mathbf{u})$.

We now outline a Gibbs sampling algorithm that moves within and between models of different dimensions, using the conditional distributions derived in this section and in Section 3:

A. Variable dimension sampler. For each $j \in \{1, \ldots, n\}$, sample $\boldsymbol{\Theta}_j = \{\mathbf{U}_{[,j]}, d_j, \mathbf{V}_{[,j]}\}$:

- sample $d_j|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi$
- sample $(\mathbf{U}_{[,j]}, \mathbf{V}_{[,j]})|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j$.

B. Fixed dimension sampler. For each $\{j : d_j \neq 0\}$,

- sample $\mathbf{U}_{[,j]}|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j, \mathbf{V}_{[,j]}$
- sample $\mathbf{V}_{[,j]}|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, d_j, \mathbf{U}_{[,j]}$
- sample $d_j|\mathbf{Y}, \boldsymbol{\Theta}_{-j}, \phi, \mu, \psi, \mathbf{U}_{[,j]}, \mathbf{V}_{[,j]}$.

C. Other terms.

- sample $\phi|\mathbf{Y}, \boldsymbol{\Theta}$
- sample $\mu|\mathbf{D}, \psi$
- sample $\psi|\mathbf{D}, \mu$.

The distributions required for the steps in A were outlined in this section; steps B and C were outlined in Section 3. Technically, iterating steps A and C alone will produce a Markov chain with the desired stationary distribution, but adding the steps in B can improve the within-model mixing of the Markov chain at a relatively low computational cost. Also, by conditioning on whether $d_j = 0$ for each j, the steps in B can be viewed as Gibbs sampling for all $j \in \{1, \ldots, n\}$, not just those for which $d_j \neq 0$. R code that implements this algorithm is available at my website.

5. SIMULATION STUDY

In this section we examine the sampling properties of the estimation procedure with a small simulation study. Each dataset in this study was simulated from the following model:

- $\mathbf{U} \sim \text{uniform}(\mathcal{V}_{5,m})$ and $\mathbf{V} \sim \text{uniform}(\mathcal{V}_{5,n})$.
- $\mathbf{D} = \text{diag}\{d_1, \ldots, d_5\}$, $\{d_1, \ldots, d_5\} \sim$ iid uniform$(\frac{1}{2}\mu_{mn}, \frac{3}{2}\mu_{mn})$.
- $\mathbf{Y} = \mathbf{UDV}' + \mathbf{E}$, where $\mathbf{E}$ is an $m \times n$ matrix of standard normal noise.

For each value of m and n, the sampling mean of $\{d_1, \ldots, d_5\}$ was taken to be $\mu_{mn} = \sqrt{n + m + 2\sqrt{nm}}$. Such a value should distribute the singular values $\{d_1, \ldots, d_5\}$ near the "cusp" of detectability: As shown in Edelman (1988), the largest singular value of an $m \times n$ matrix $\mathbf{E}$ of standard normal noise is approximately μ_{mn} for large m and n.

Three-hundred datasets were generated using the preceding model, one hundred for each of the three sample sizes $(m, n) \in \{(10, 10), (100, 10), (100, 100)\}$. These were generated in the R statistical computing environment using the integers 1–100 as random seeds for each of the three sample sizes. Code to generate these datasets is available from my website. Prior distributions for the parameters $\{\phi, \mu, \psi\}$ were taken as previously described with "prior sample sizes" of $\nu_0 = 2$ and $\eta_0 = 2$. This gives exponential prior distributions for ϕ and ψ. The values of σ_0^2, μ_0, and τ_0^2 were derived from empirical Bayes-type estimates obtained by averaging over different ranks as follows:

1. For each $k \in \{0, \ldots, n\}$:

   (a) Let $\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'$ be the least-squares projection of $\mathbf{Y}$ onto the set of rank-k matrices.
   (b) Let $\hat{\sigma}_k^2 = \|\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'\|^2/(nm)$.
   (c) Let $\hat{\mu}_k = \sum_{j=1}^{k} \hat{d}_j/k$ and $\hat{\tau}_k^2 = \sum_{j-1}^{k}(\hat{d}_j - \bar{\hat{d}})^2/k$.

2. Let $\sigma_0^2 = \frac{1}{n+1}\sum_{j=0}^{n}\hat{\sigma}_j^2$, $\mu_0 = \frac{1}{n+1}\sum_{j=0}^{n}\hat{\mu}_j$, $\nu_0^2 = \frac{1}{n}\sum_{j=0}^{n}(\hat{\mu}_j - \bar{\hat{\mu}})^2$, and $\tau_0^2 = \frac{1}{n+1}\sum_{j=0}^{n}\hat{\tau}_j^2$.

The resulting prior distributions are weakly centered around averages of empirical estimates, where the averaging is over ranks 0 through n. Finally, the prior distribution on the rank K of the mean matrix was taken to be uniform on $\{0, \ldots, n\}$.

For each of the 100×3 datasets, 20,000 iterations of the Gibbs sampling scheme described in Section 4.3 were run to obtain approximate samples from the posterior distribution of $\mathbf{UDV}'$. Parameter values were saved every 10th scan after dropping the first 10,000 scans to allow for burn-in, resulting in 1,000 Monte Carlo samples per simulation. All Markov chains were begun with $K = 0$ and $\{\phi, \mu, \psi\}$ set equal to their prior modes. Summaries of the posterior distributions for the three different values of (m, n) are displayed in Figure 2. The first column of each panel plots the MCMC approximation to the expected value of $p(K|\mathbf{Y})$ for each value of (m, n). The expectation $E_Y[p(K|\mathbf{Y})]$ is approximated by $\frac{1}{100}\sum_{s=1}^{100} p(K|\mathbf{Y}_{(s)})$, where $\mathbf{Y}_{(s)}$ is the sth simulated dataset for a given value of (m, n) [for the case $m = n = 100$, $p(K|\mathbf{Y})$ is plotted only for $K \leq 10$, although the distribution extends beyond this value]. These distributions are all peaked around the correct value of $K = 5$.

Also of interest is how frequently the posterior mode $\hat{K} = \arg\max p(K|\mathbf{Y})$ obtains the true value of K. This information is displayed in the second column of Figure 2, which gives the empirical distribution of $\hat{K}$ taken over each of the 100 datasets. As we see, the true value $K = 5$ is the most frequent value of the estimate in each dataset. For each dataset, we also computed three other estimates of K: $\hat{K}_l$, the Laplace approximation of Minka (2000); $\hat{K}_c$, the number of eigenvalues of the correlation matrix of $\mathbf{Y}$ that are larger than 1, often suggested in the factor analysis literature; and $\hat{K}_e$, the index of the largest gap in the eigenvalues of $\mathbf{Y}'\mathbf{Y}$, which is used in machine learning and clustering (see, e.g., Meila and Xu 2003). Descriptions of the sampling distributions of these estimators are presented in Table 1. The only case in which the peak of the sampling distribution for one of these estimators obtained the correct value was $\hat{K}_l$ in the case of $m = 100$ and $n = 10$, although the sampling distribution was less peaked around the true value than that of the Bayes estimate $[\Pr(\hat{K}_l = 5|\mathbf{Y}) = .46$ versus $\Pr(\hat{K} = 5|\mathbf{Y}) = .66]$. In the case of $m = 10$ and $n = 10$, $\hat{K}_l$ did poorly, with $\Pr(\hat{K}_l = 1|\mathbf{Y}) = .86$. Finally, we note that $\hat{K}_c$ behaved extremely poorly for the case $m = n = 100$, having a sampling distribution centered around $K = 36$. This is perhaps due to the fact that the asymptotic behavior of eigenvalues for square matrices is quite different than that of rectangular matrices (see Edelman 1988). R-code to obtain all estimators for these simulated data is available at my website.

Perhaps of more importance than an estimate of $\hat{K}$ is an accurate estimate of $\mathbf{M}$. In the last column of Figure 2 we compare the error of the model-averaged estimate of $\mathbf{M}$ to that of the least-squares estimate. For each simulated dataset, the posterior mean $\hat{\mathbf{M}} = E[\mathbf{M}|\mathbf{Y}]$ was obtained by averaging its value over the last 10^4 scans of the Gibbs sampler. The squared error in estimation, averaged over elements of the mean matrix was calculated as $\text{ASE}_B = \|\hat{\mathbf{M}} - \mathbf{M}\|^2/(mn)$, where $\mathbf{M}$ is the mean matrix that generated the data. This value is compared to ASE_{LS}, which is the corresponding average squared error of the least-squares projection of $\mathbf{Y}$ onto the space of rank-$\hat{K}$ matrices. The distribution of this ratio is mostly below 1 for the case $m = n = 10$ and strictly below 1 for the other two cases where there are more parameters to estimate. This corresponds with our intuition: The model-averaged estimates improve relative to the least-squares estimates as the number of parameters increases. These results indicate that simply obtaining a posterior estimate $\hat{K}$ of K and then using the corresponding rank-$\hat{K}$ least-squares estimate of $\mathbf{M}$ generally results in an estimate that can be substantially improved upon by model averaging, at least in terms of this error criterion.

As a simple summary of the mixing properties of these Markov chains, we examined the convergence and autocorrelation of the marginal samples of the error precision ϕ using Geweke's (1992) z test of stationarity, and by calculating the effective sample size of the 1,000 Monte Carlo samples saved for each chain. We declared a Markov chain simulation a "success" if Geweke's z statistic did not exceed 2 in absolute value and if the effective sample size was at least 100. Based on these criteria' the percentages of MCMC simulations that were successful were 82%, 86%, and 91% for the three sample sizes $\{(10, 10), (100, 10), (100, 100)\}$, respectively. The simulation
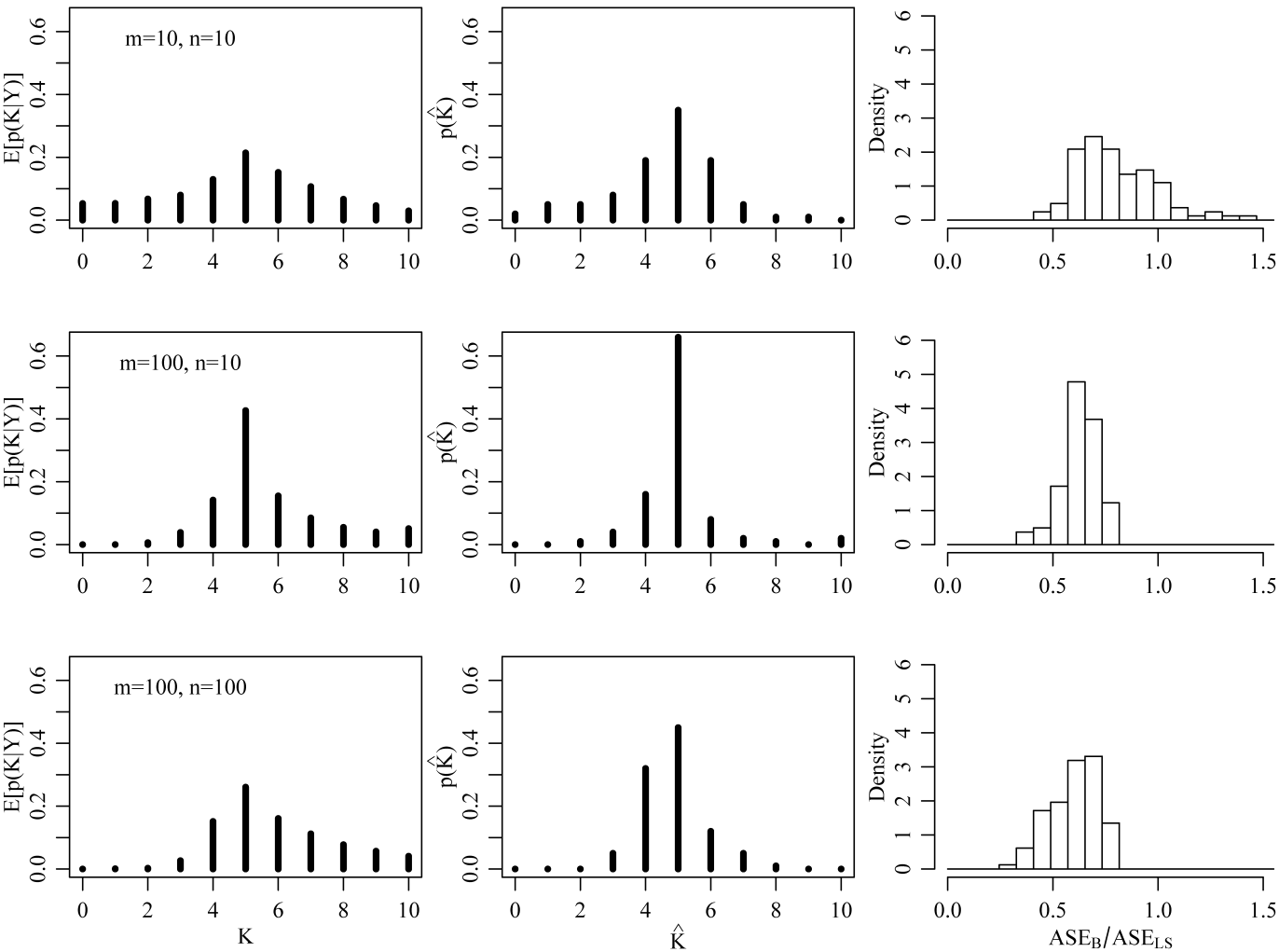
Figure 2. Results of the Simulation Study. Plots in the first column give the averages of p(K|**Y**) over 100 simulated datasets. The second column gives the empirical distribution of the posterior mode $\hat{K}$. The third column gives the distribution of the ratio of the squared error of the Bayes estimate of **M** to that of the least-squares estimate.

"failures" were not examined extensively, but we conjecture that running these chains longer is likely to result in improved estimation. Of course, in the analysis of a single dataset the usual recommendation is to assess the convergence and the autocorrelation of the Markov chain, and to adjust the length accordingly.

Finally, the performance of the model was examined using two additional prior distributions. The posterior analysis was first rerun using a "diffuse" conjugate prior distribution, in which $\phi \sim$ exponential(1), $\psi \sim$ exponential(1), and $\mu \sim$ normal($0, 1/\psi$). The modes of the sampling distribution of $\hat{K}$

Table 1. Comparison of Model Selection Procedures: Mode of the Sampling Distribution of $\hat{K}$ and the Probability That $\hat{K} = 5$ for Different Estimators of the Rank of **M**

Sample size	Estimator			
	$\hat{K}$	$\hat{K}_e$	$\hat{K}_c$	$\hat{K}_l$
(10, 10)	5, .35	2, .00	4, .14	1, .00
(100, 10)	5, .66	3, .09	4, .27	5, .46
(100, 100)	5, .45	6, .21	36, .00	4, .24

were 6, 5, and 5 for sample sizes (10, 10), (100, 10), and (100, 100), respectively, with $\Pr(\hat{K} = 5)$ equal to .21, .53, and .32 for the three different cases, indicating slightly worse performance than the empirical Bayes prior distribution, but better performance than the other approaches. A modification to this prior was also implemented, in which $\mu \sim$ normal ($\phi^{-1/2} \times \sqrt{n + m + 2\sqrt{nm}}, 1/\psi$) and the other distributions remained the same. This prior distribution focuses the search for nonzero d_j's to values that are as large as the largest singular values of normally distributed noise matrices. Such a prior distribution may be appealing in practice, because it requires that factors that enter into the model have a reasonable magnitude. Use of this prior distribution resulted in sampling distributions for $\hat{K}$ that have modes of 5 for each of the sample sizes, with $\Pr(\hat{K} = 5)$ equal to .32, .55, and .42. A more complicated alternative to these prior distributions would be to have the prior distributions for $\{\phi, \mu, \psi\}$ depend on K. For example, given $K = k$, the prior distributions for $\{\phi, \mu, \psi\}$ could be based on $\{\hat{\sigma}_k^2, \hat{\mu}_k, \hat{\tau}_k^2\}$. Such prior distributions would require some minor modifications to the variable dimension sampler outlined in the previous section.

6. EXTENSION AND EXAMPLE: ANALYSIS OF BINARY RELATIONAL DATA

A potentially useful extension of the model described in the preceding text is to a class of generalized bilinear models of the form

$$\theta_{i,j} = \boldsymbol{\beta}'\mathbf{x}_{i,j} + \mathbf{u}_i'\mathbf{D}\mathbf{v}_j + e_{i,j},$$

$$E[y_{i,j}|\boldsymbol{\Theta}] = g^{-1}(\theta_{i,j}),$$

where g is the link function. Such models allow for the analysis of a variety of data types. For example, binary data can be modeled as $y_{i,j} \sim \text{binary}(\exp\{\theta_{i,j}\}/(1+\exp\{\theta_{i,j}\}))$ and count data as $y_{i,j} \sim \text{Poisson}(\exp\{\theta_{i,j}\})$. Gabriel (1998) considered maximum-likelihood estimation for a variant of this model in situations where the dimension of $\mathbf{D}$ is fixed, and Hoff (2005) considered a symmetric version of this model for the analysis of social network data. Parameter estimation and dimension selection for the foregoing model can be made by sampling from a Markov chain generated by a modified version of the algorithm of Section 4.3. Given current values of $\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{U}, \mathbf{D}, \mathbf{V}$, sample new values as follows:

1. Let $\tilde{\mathbf{Y}} = \boldsymbol{\Theta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{UDV}' + \mathbf{E}$. Update $\mathbf{U}, \mathbf{D}$, and $\mathbf{V}$ from their conditional distribution given $\tilde{\mathbf{Y}}$ as described in Section 4.3.
2. Let $\tilde{\mathbf{Y}} = \boldsymbol{\Theta} - \mathbf{UDV}' = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$. Update $\boldsymbol{\beta}$ from its conditional distribution given $\tilde{\mathbf{Y}}$ (a multivariate normal distribution).
3. Sample $\boldsymbol{\Theta}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{UDV}' + \mathbf{E}^*$, where $\mathbf{E}^*$ is a matrix of normally distributed noise with zero mean and precision ϕ. Replace $\theta_{i,j}$ by $\theta_{i,j}^*$ with probability $(p(y_{i,j}|\theta_{i,j}^*))/(p(y_{i,j}|\theta_{i,j})) \wedge 1$.

We illustrate the use of such a model and estimation procedure with an analysis of binary relational data between 46 global service firms and 55 cities, obtained from the Globalization and World Cities study group (*http://www.lboro.ac.uk/gawc*). For these data, $y_{i,j} = 1$ if firm j has an office in city i and $y_{i,j} = 0$ otherwise. Standard practice is to represent within-row and within-column homogeneity with effects that are additive on the log-odds scale,

$$\log \text{odds}(y_{i,j} = 1) = \beta + a_i + b_j, \tag{7}$$

and so the effects $\mathbf{a} = \{a_1, \ldots, a_m\}$ and $\mathbf{b} = \{b_1, \ldots, b_n\}$ constitute a rank-2 structure. We look for evidence of higher order

structure by considering the model

$$\log \text{odds}(y_{i,j} = 1) = \beta + \gamma_{i,j},$$
$$\gamma_{i,j} = \mathbf{u}_i'\mathbf{D}\mathbf{v}_j + e_{i,j}. \tag{8}$$

The rank-2 structure of model (7) is easily incorporated into (9) by fixing $\mathbf{U}_{[,1]} = \frac{1}{\sqrt{m}}\mathbf{1}_{m \times 1}$ and $\mathbf{V}_{[,2]} = \frac{1}{\sqrt{n}}\mathbf{1}_{n \times 1}$ and modeling d_1 and d_2 to be nonzero with probability 1. The additive city and firm effects are then given by $\mathbf{a} = d_2\mathbf{U}_{[,2]}$ and $\mathbf{b} = d_1\mathbf{V}_{[,1]}$, respectively. Note that any remaining effects represented by $\mathbf{UDV}'$ will be orthogonal to these additive effects, and that the mean of the matrix $\mathbf{UDV}'$ is identically zero, making it unaliased with the intercept β. For the remainder of this analysis, the variable K will refer to the number of additional nonzero singular values of $\mathbf{UDV}'$ beyond the additive row and column effects.

We fix the error variance $1/\phi = 1$, because this scaling parameter is confounded with the magnitude of β and $\mathbf{UDV}'$. For simplicity, we use independent normal $(0, 100)$ prior distributions for β and the nonzero elements of $\mathbf{D}$, and a uniform prior distribution for K. A Markov chain of length 25,000 was constructed using the previously described algorithm, starting with $K = 0$. Mixing across ranks K was rapid as is shown in the first panel of Figure 3, which displays values of K every 100th scan of the Markov chain. Stationary of the Markov chain in K was not rejected at level .05 based on Geweke's z test, and the effective sample size for estimating the posterior distribution of K was 472. The Monte Carlo estimate of $p(K|\mathbf{Y})$, shown in the second panel, gives a posterior mode of $K = 6$ and suggests strong evidence for structure in the log odds beyond that of the additive row and column effects.

One of the practical motivations for selecting an appropriate model dimension is prediction. Many binary social network datasets include missing values, in which it is not known whether $y_{i,j} = 1$ or $y_{i,j} = 0$. In such cases it is often desirable to make predictions about missing values based on the observed data, and thus to base model selection on predictive performance. With this in mind, we compare the preceding results to the following 10-fold cross-validation procedure:

1. Randomly split the set of pairs $\{i, j\}$ into 10 test sets $A_1, \ldots, A_{10}$.
2. For $K = 0, 1, \ldots, K_{\max}$:
   (a) For $l = 1, \ldots, 10$:
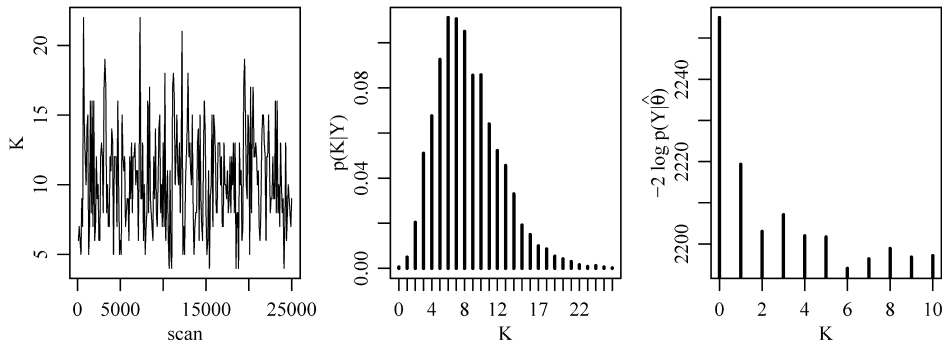

Figure 3. Posterior Estimation of K. The first panel plots values of K every 100th scan of the Markov chain. The second panel plots the Monte Carlo estimate of p(K|**Y**). The third panel gives the results of a cross-validation evaluation of K ∈ {0, . . . , 10}.

i. With the rank fixed at K, perform the MCMC algorithm using only $\{y_{i,j} : \{i,j\} \notin A_l\}$, but sample values of $\theta_{i,j}$ for all ordered pairs.

ii. Based on the Monte Carlo sample values $\{\theta_{i,j}^{(1)}, \ldots, \theta_{i,j}^{(S)}\}$, compute the posterior mean $\hat{\mu}_{i,j} = (1/S) \sum_{s=1}^{S} (\exp\{\theta_{i,j}^{(s)}\}/(1+\exp\{\theta_{i,j}^{(s)}\}))$ for $\{i,j\} \in A_l$ and the log predictive probability $\text{lpp}(A_l) = \sum_{\{i,j\} \in A_l} \log p(y_{i,j}|\hat{\mu}_{i,j})$.

(b) Measure the predictive performance for K as $\text{LPP}(K) = \sum_{l=1}^{K_{\max}} \text{lpp}(A_l)$.

The values of $-2\text{LPP}(K)$ for $K \in \{0, \ldots, 10\}$ are shown in the third panel of Figure 3. For the particular random partitioning of the data used here, the cross-validation procedure suggests a model rank of $K = 6$, which is the same value as the posterior mode of the Bayes solution. However, a comparison of N values of K using a 10-fold cross-validation procedure requires the construction of $10 \times N$ separate Markov chains, and further requires specification of the values of K to be compared. In contrast, the Bayesian procedure requires only one MCMC run and can potentially visit each value of $K \in \{1, \ldots, n\}$.

Finally we examine some of the patterns in the structure of $\mathbf{UDV}'$ beyond those of the additive effects. The posterior mean of $\mathbf{UDV}$, minus the additive effects, was obtained by averaging over scans of the Markov chain. The first two singular values and vectors of this matrix were obtained, and the values of the resulting row (city) effects are plotted in Figure 4. These values are strongly related to geography: U.S. cities cluster together, as do cities in Europe, Latin America, and the Pacific rim.

7. DISCUSSION

This article has presented a model-based version of the singular value decomposition, thereby extending a general data analysis tool that has a wide variety of data analysis purposes and interpretations. For example, in this article it is used for noise reduction and estimation of a mean matrix (Sec. 5), as well as prediction and data description (Sec. 6).

The approach taken in this paper is to model the data matrix $\mathbf{Y}$ as equal to a reduced-rank mean matrix $\mathbf{M}$ plus Gaussian noise, and to estimate $\mathbf{M}$ simultaneously along with its rank. The approach is Bayesian and the estimation procedure, based on Markov chain Monte Carlo methods, allows for a variety of model extensions, such as to the generalized bilinear models described in Section 6, estimation using replicate data matrices, and estimation subject to missing data. This latter extension may be of particular use in the analysis of relational data among a large number of nodes, where it may be too costly to make observations on all possible pairs. In such cases, the value of $y_{i,j}$ may be missing for many pairs, but one can make predictions based on estimates $\mathbf{u}_i$, $\mathbf{D}$, and $\mathbf{v}_j$ obtained from the observed data. Using this approach to predict missing links in social networks and protein–protein interaction networks is one of my current research areas. However, for large datasets with 1,000 nodes (10^6 observations) or more, the MCMC scheme in this article becomes prohibitively computationally expensive. I am currently studying methods for making approximate Bayesian inference for large relational datasets. These include Laplace approximations for various components of the MCMC
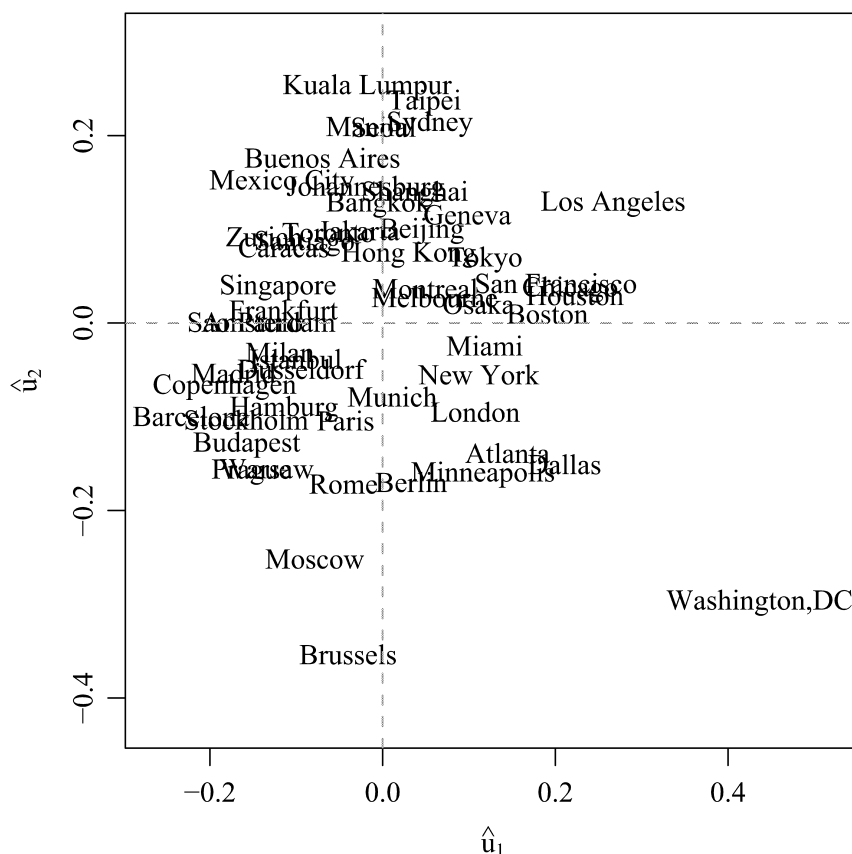

Figure 4. City Specific Effects: The First Two Left Singular Vectors of $E(\mathbf{M}|\mathbf{Y})$ Indicate Strong Geographic Patterns in the Data.

scheme in Sections 3 and 4, and using variational methods to approximate joint posterior distributions (Jordan, Ghahramani, Jaakkola, and Saul 1999).

Although based on the conceptually straightforward Gibbs sampler, the complexity of the full conditional distributions used in the Markov chain in Section 4 suggests we look for an alternative procedure. For example, we could model only d_j to be zero with nonzero probability, and sample from its full conditional distribution instead of sampling marginally over $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$. Unfortunately, an algorithm based on this approach will not mix well across ranks of $\mathbf{M}$ because d_j, $\mathbf{U}_{[,j]}$, and $\mathbf{V}_{[,j]}$ are dependent to an extreme: The probability of sampling $d_j \neq 0$ is essentially zero unless $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ are near a pair of local modes, but the probability of $\mathbf{U}_{[,j]}$ and $\mathbf{V}_{[,j]}$ being in such a state is essentially zero if $d_j = 0$. An alternative approach to sampling from distributions on complicated sample spaces is to use Metropolis–Hastings type algorithms. Based on my initial work on this problem, obtaining proposal distributions for these algorithms that achieve even minimal acceptance rates requires an extreme amount of tuning. In contrast, Gibbs sampling for this model is possible (as shown in this article), requires no tuning or prespecification of model dimensions to be considered, and, for the examples in this article, mixes well across matrices $\mathbf{M}$ of different ranks.

Computer code and data for all numerical results in this article are available at *www.stat.washington.edu/hoff*.

APPENDIX A: PROOF OF PROPOSITION 1

We first construct a sample from the uniform distribution on $\mathcal{V}_{K,m}$ and then show that it has the desired conditional distributions. Let $\mathbf{z}_1, \ldots, \mathbf{z}_K$ be iid multivariate normal $(0, \mathbf{I}_{m \times m})$. Let $\mathbf{x}_1 = \mathbf{z}_1$ and, for $j = 1, \ldots, K-1$, let

- $\mathbf{X}_j = (\mathbf{x}_1 \cdots \mathbf{x}_j)$
- $\mathbf{P}_j = \mathbf{I} - \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'$
- $\mathbf{x}_{j+1} = \mathbf{P}_j\mathbf{z}_{j+1}$.

Note that $\mathbf{P}_j$ is the symmetric, idempotent projection matrix of $\mathbb{R}^K$ onto the null space of $\mathbf{X}_j$, and so the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{j+1}$ are orthogonal. For each j, let $\mathbf{U}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1/2}$. For $j = K$, we have

$$\mathbf{X}_K'\mathbf{X}_K = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_K' \end{pmatrix} (\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_K)$$

$$= \begin{pmatrix} |\mathbf{x}_1|^2 & 0 & \cdots & 0 \\ 0 & |\mathbf{x}_2|^2 & \cdots & 0 \\ 0 & 0 & \cdots & |\mathbf{x}_K|^2 \end{pmatrix}$$

and so

$$\mathbf{U}_K = \left(\frac{\mathbf{x}_1}{|\mathbf{x}_1|}, \frac{\mathbf{x}_2}{|\mathbf{x}_2|}, \ldots, \frac{\mathbf{x}_K}{|\mathbf{x}_K|} \right)$$

is a matrix of K orthonormal vectors in $\mathbb{R}^m$. The proof will be complete if we can show the following lemmas hold:

Lemma A.1. The distribution of $\mathbf{U}_K$ is the uniform distribution on $\mathcal{V}_{K,m}$.

Lemma A.2. We have $\mathbf{U}_{[,k+1]}|\mathbf{U}_k \overset{d}{=} \mathbf{N}_k\mathbf{u}_{k+1}$, where $\mathbf{N}_k$ is an orthonormal basis for the null space of $\mathbf{U}_k$ and $\mathbf{u}_{k+1}$ is distributed uniformly on the $(m-k)$-dimensional sphere.

Proof of Lemma A.1. By theorem 2.2.1(iii) of Chikuse (2003), an $m \times K$ matrix of the form $\mathbf{X}_K(\mathbf{X}_K'\mathbf{X}_K)^{-1/2}$ is uniformly distributed on $\mathcal{V}_{K,m}$ if $\mathbf{X}_K$ is an $m \times K$ random matrix with rank K a.s. and has a distribution that is invariant under left-orthogonal transformations. For each $\mathbf{X}_k$ constructed, we will show left invariance by induction. Let $\mathbf{H}: \mathbb{R}^m \to \mathbb{R}^m$ be an orthogonal transformation and note that $\mathbf{HX}_1 = \mathbf{Hx}_1 \overset{d}{=} \mathbf{x}_1 = \mathbf{X}_1$. Now suppose $\mathbf{HX}_k \overset{d}{=} \mathbf{X}_k$. The distribution of $\mathbf{HX}_{k+1}$ is determined by its characteristic function:

$$E\left[\exp\left\{ i \sum_{j=1}^{k+1} \mathbf{t}_j'\mathbf{Hx}_j \right\} \right]$$

$$= E\left[\exp\left\{ i \sum_{j=1}^{k} \mathbf{t}_j'\mathbf{Hx}_j \right\} E[\exp\{i\mathbf{t}_{k+1}'\mathbf{Hx}_{k+1}\}|\mathbf{X}_k] \right].$$

Note that $\mathbf{t}_{k+1}'\mathbf{Hx}_{k+1} = (\mathbf{P}_k'\mathbf{H}'\mathbf{t}_{k+1})'\mathbf{z}_{k+1}$, where $\mathbf{z}_{k+1}$ is a vector of independent standard normals and is independent of $\mathbf{X}_k$. Thus the characteristic function can be rewritten as

$$E\left[\exp\left\{ i \sum_{j=1}^{k} \mathbf{t}_j'\mathbf{Hx}_j \right\} \exp\left\{ -\frac{1}{2}\mathbf{t}_{k+1}'\mathbf{HP}_k\mathbf{P}_k'\mathbf{H}'\mathbf{t}_{k+1} \right\} \right]$$

$$= E\left[\exp\left\{ i \sum_{j=1}^{k} \mathbf{t}_j'\tilde{\mathbf{x}}_j \right\} \exp\left\{ -\frac{1}{2}\mathbf{t}_{k+1}'\tilde{\mathbf{P}}_k\mathbf{t}_{k+1} \right\} \right], \quad \text{(A.1)}$$

where $\tilde{\mathbf{x}}_j = \mathbf{Hx}_j$ and

$$\tilde{\mathbf{P}}_k = \mathbf{HP}_k\mathbf{P}_k'\mathbf{H}' = \mathbf{HP}_k\mathbf{H}'$$

$$= \mathbf{H}(\mathbf{I} - \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k')\mathbf{H}'$$

$$= \mathbf{I} - \mathbf{HX}_k((\mathbf{HX}_k)'(\mathbf{HX}_k))^{-1}(\mathbf{HX}_k)'.$$

A similar calculation shows that the distribution of $\mathbf{X}_j$ is characterized by

$$E\left[\exp\left\{ i \sum_{j=1}^{k+1} \mathbf{t}_j'\mathbf{x}_j \right\} \right]$$

$$= E\left[\exp\left\{ i \sum_{j=1}^{k} \mathbf{t}_j'\mathbf{x}_j \right\} \exp\left\{ -\frac{1}{2}\mathbf{t}_{k+1}'\mathbf{P}_k\mathbf{t}_{k+1} \right\} \right]. \quad \text{(A.2)}$$

By assumption, $\mathbf{X}_k \overset{d}{=} \mathbf{HX}_k$, and so $\{\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{P}_k\} \overset{d}{=} \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k, \tilde{\mathbf{P}}_k\}$ and the expectations (A.1) and (A.2) are equal. Because the characteristic functions specify the distributions, $\mathbf{HX}_{k+1} \overset{d}{=} \mathbf{X}_{k+1}$ and the lemma is proved.

Proof of Lemma A.2. The vector $\mathbf{U}_{[,k+1]}$ is constructed as $\mathbf{U}_{[,k+1]} = \mathbf{P}_k\mathbf{z}_{k+1}/|\mathbf{P}_k\mathbf{z}_{k+1}|$. The matrix $\mathbf{P}_k$ has $m-k$ eigenvalues of 1, the rest being 0, giving the eigenvalue decomposition $\mathbf{P}_k = \mathbf{N}_k\mathbf{N}_k'$, where $\mathbf{N}_k$ is a $m \times (m-k)$ matrix whose columns form an orthonormal basis for the null space of $\mathbf{U}_k$. Substituting in $\mathbf{N}_k\mathbf{N}_k'$ for $\mathbf{P}_k$ gives

$$\mathbf{U}_{[,k+1]} = \frac{\mathbf{N}_k\mathbf{N}_k'\mathbf{z}_{k+1}}{|\mathbf{N}_k\mathbf{N}_k'\mathbf{z}_{k+1}|}$$

$$= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}_{k+1}}{(\mathbf{z}'\mathbf{N}_k\mathbf{N}_k'\mathbf{N}_k\mathbf{N}_k'\mathbf{z})^{1/2}}$$

$$= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}_{k+1}}{(\mathbf{z}'\mathbf{N}_k\mathbf{N}_k'\mathbf{z})^{1/2}}$$

$$= \mathbf{N}_k \frac{\mathbf{N}_k'\mathbf{z}}{|\mathbf{N}_k'\mathbf{z}|}.$$

Note that for each k, $\mathbf{U}_k = \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1/2}$, and so the projection matrix $\mathbf{P}_k$ can be written as $\mathbf{I} - \mathbf{U}_k\mathbf{U}_k'$, a function of $\mathbf{U}_k$. Therefore, given $\mathbf{U}_k$, $\mathbf{U}_{[,k+1]}$ is equal in distribution to $\mathbf{N}_k$ (a function of $\mathbf{U}_k$) multiplied by $\mathbf{N}_k'\mathbf{z}/|\mathbf{N}_k'\mathbf{z}|$. The distribution of $\mathbf{N}_k'\mathbf{z}$ can be found via its characteristic function: For an $m - k$ vector $\mathbf{t}$,

$$E\big[\exp\{i\mathbf{t}'(\mathbf{N}_k'\mathbf{z})\}\big] = E\big[\exp\{i(\mathbf{N}_k\mathbf{t})'\mathbf{z}\}\big]$$

$$= \exp\left\{-\frac{1}{2}\mathbf{t}'\mathbf{N}_k'\mathbf{N}_k\mathbf{t}\right\}$$

$$= \exp\left\{-\frac{1}{2}\mathbf{t}'\mathbf{t}\right\},$$

so we see that $\mathbf{N}_k\mathbf{z}_{k+1}$ is equal in distribution to an $m - k$ vector of independent standard normal random variables and that $\mathbf{N}_k\mathbf{z}_{k+1}/|\mathbf{N}_k \times \mathbf{z}_{k+1}|$ is uniformly distributed on the $m - k$ sphere.

APPENDIX B: EXPECTATION OF $e^{\mathbf{u}'\mathbf{A}\mathbf{v}}$

In this section we compute $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$ for uniformly distributed unit vectors $\mathbf{u}$ and $\mathbf{v}$, and an arbitrary $m \times n$ matrix $\mathbf{A}$. Integrating with respect to $\mathbf{v}$ can be accomplished by noting that as a function of $\mathbf{v}$, $e^{\mathbf{u}'\mathbf{A}\mathbf{v}}$ is proportional to the von Mises–Fisher distribution on the n sphere S_n, with parameter $\mathbf{u}'\mathbf{A}$; that is,

$$\int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} p(\mathbf{v})\,dS_n(\mathbf{v}) = \int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} c_n(0)\,dS_n(\mathbf{v})$$

$$= \frac{c_n(0)}{c_n(\|\mathbf{u}'\mathbf{A}\|)}\int e^{\mathbf{u}'\mathbf{A}\mathbf{v}} c_n(\|\mathbf{u}'\mathbf{A}\|)\,dS_n(\mathbf{v})$$

$$= \frac{c_n(0)}{c_n(\|\mathbf{u}'\mathbf{A}\|)}$$

$$= \Gamma\left(\frac{n}{2}\right)\left(\frac{2}{\|\mathbf{u}'\mathbf{A}\|}\right)^{n/2-1} I_{n/2-1}(\|\mathbf{u}'\mathbf{A}\|),$$

where I_ν is the modified Bessel function of the first kind. The series expansion of $I_{n/2-1}(\|\mathbf{u}'\mathbf{A}\|)$ gives

$$\Gamma\left(\frac{n}{2}\right)\left(\frac{2}{\|\mathbf{u}'\mathbf{A}\|}\right)^{n/2-1} I_{n/2-1}(\|\mathbf{u}'\mathbf{A}\|)$$

$$= \sum_{l=0}^{\infty} \|\mathbf{u}'\mathbf{A}\|^{2l} \frac{\Gamma(n/2)}{\Gamma(l+1)\Gamma(l+n/2)4^l}.$$

All the terms in the sum are positive, so $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$ can be found by replacing $\|\mathbf{u}'\mathbf{A}\|^{2l}$ with its expectation in the preceding equation. To compute this expectation, let $\mathbf{A} = \mathbf{L}\mathbf{\Lambda}^{1/2}\mathbf{R}'$ be the singular value decomposition of $\mathbf{A}$, where $\mathbf{L}'\mathbf{L} = \mathbf{R}'\mathbf{R} = \mathbf{I}$ and $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{A}'\mathbf{A}$. Then

$$\|\mathbf{u}'\mathbf{A}\|^2 = \mathbf{u}'\mathbf{A}\mathbf{A}'\mathbf{u}$$

$$= \mathbf{u}'\mathbf{L}\mathbf{\Lambda}^{1/2}\mathbf{R}'\mathbf{R}\mathbf{\Lambda}^{1/2}\mathbf{L}'\mathbf{u}$$

$$= \mathbf{u}'\mathbf{L}\mathbf{\Lambda}\mathbf{L}'\mathbf{u}$$

$$\equiv \tilde{\mathbf{u}}'\mathbf{\Lambda}\tilde{\mathbf{u}}$$

$$= \sum_{j=1}^{n} \tilde{u}_j^2 \lambda_j,$$

where $\tilde{\mathbf{u}} = \mathbf{L}'\mathbf{u}$. We will now identify the distribution of the vector $\{\tilde{u}_1^2, \ldots, \tilde{u}_n^2\}$. Let $\mathbf{B} = \{\mathbf{L}, \mathbf{L}^\perp\}$ be an orthonormal basis for $\mathbb{R}^m$. Because the uniform distribution on the sphere is rotationally invariant, $\mathbf{B}'\mathbf{u}$ is equal in distribution to $\mathbf{u}$ and so $\mathbf{L}'\mathbf{u}$ is equal in distribution to the first n coordinates of $\mathbf{u}$. Recall that a uniformly distributed vector

$\mathbf{u}$ can be generated by sampling $z_1, \ldots, z_m$ independently from a standard normal distribution and then dividing each term by $|\sum z_i^2|^{1/2}$. Therefore,

$$\{\tilde{u}_1^2, \ldots, \tilde{u}_n^2\} \stackrel{d}{=} \frac{\{z_1^2, \ldots, z_n^2\}}{\sum_{j=1}^m z_j^2}$$

$$= \left(\frac{\sum_{j=1}^n z_j^2}{\sum_{j=1}^m z_j^2}\right)\left(\frac{\{z_1^2, \ldots, z_n^2\}}{\sum_{j=1}^n z_j^2}\right)$$

$$\stackrel{d}{=} \theta\mathbf{q},$$

where $\theta \sim \text{beta}(n/2, (m-n)/2)$, $\mathbf{q} \sim \text{Dirichlet}_n(1/2, \ldots, 1/2)$, and θ and $\mathbf{q}$ are independent. Therefore, $\|\mathbf{u}'\mathbf{A}\|^2 \stackrel{d}{=} \theta\lambda'\mathbf{q}$, where λ is the diagonal of $\mathbf{\Lambda}$ and $\mathbf{q}$ are the eigenvalues of $\mathbf{A}'\mathbf{A}$. The required expectation is then

$$E[\|\mathbf{u}'\mathbf{A}\|^{2l}] = E[\theta^l]E[(\lambda'\mathbf{q})^l].$$

The first expectation is given by $E[\theta^l] = [\Gamma(n/2+l)\Gamma(m/2)]/[\Gamma(m/2+l)\Gamma(n/2)]$. The second expectation is the lth moment of a Dirichlet average, which results in a type of multiple hypergeometric function denoted as $R_l(\lambda, \frac{1}{2}\mathbf{1})$. This expectation and its generalizations have been studied by Carlson (1977, chap. 5), Dickey (1983), and others. An algorithm for recursively computing $R_1, \ldots, R_l$ exactly from a generating function is provided in the next section.

To make the result of the calculation a little more intuitive, let $\tilde{\lambda} = \lambda/\sum\lambda_j \equiv \lambda/\|\mathbf{A}\|^2$ and make use of the fact that $E[(\lambda'\mathbf{q})^l] = \|\mathbf{A}\|^{2l}E[(\tilde{\lambda}'\mathbf{q})^l]$. Combining the results gives

$$E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}] = \sum_{l=0}^{\infty} \|\mathbf{A}\|^{2l} E[(\tilde{\lambda}'\mathbf{q})^l]\frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}$$

$$\equiv \sum_{l=0}^{\infty} \|\mathbf{A}\|^{2l} a_l,$$

and so we see how the expectation is related to the norm of $\mathbf{A}$ via $\|\mathbf{A}\|^{2l}$ and the variability in relative sizes of the squared singular values via $E[(\tilde{\lambda}'\mathbf{q})^l]$.

To get bounds on a finite-sum approximation to $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}]$, note that $\lambda_{\min}^l < E[(\lambda'\mathbf{q})^l] < \lambda_{\max}^l$, so

$$\sum_{l=r+1}^{\infty} \lambda_{\min}^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}$$

$$< \sum_{l=r+1}^{\infty} E[(\lambda'\mathbf{q})^l]\frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}$$

$$< \sum_{l=r+1}^{\infty} \lambda_{\max}^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}.$$

The outer sums can be computed as

$$\sum_{l=r+1}^{\infty} \lambda^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l}$$

$$= \left(\frac{2}{\sqrt{\lambda}}\right)^{m/2-1} I_{m/2-1}(\sqrt{\lambda})\Gamma\left(\frac{m}{2}\right)$$

$$- \sum_{l=1}^{r} \lambda^l \frac{\Gamma(m/2)}{\Gamma(m/2+l)\Gamma(1+l)4^l},$$

and so bounds on $E[e^{\mathbf{u}'\mathbf{A}\mathbf{v}}] - \sum_{l=0}^{r}\|\mathbf{A}\|^{2l}a_l$ can be obtained.

APPENDIX C: COMPUTING $E[(\boldsymbol{\lambda}'\mathbf{q})^l]$

Let $\mathbf{q} \sim \text{Dirichlet}_n(\alpha_1, \ldots, \alpha_n)$. Carlson (1977, sec. 6.6) showed that

$$\prod_{i=1}^{n}(1 - t\lambda_i)^{-\alpha_i} = \sum_{l=0}^{\infty} \frac{\Gamma(\boldsymbol{\alpha}'\mathbf{1} + l)}{\Gamma(\boldsymbol{\alpha}'\mathbf{1})\Gamma(l+1)} t^l E[(\boldsymbol{\lambda}'\mathbf{q})^l].$$

Let

$$c_l = \frac{\Gamma(\boldsymbol{\alpha}'\mathbf{1} + l)}{\Gamma(\boldsymbol{\alpha}'\mathbf{1})\Gamma(l+1)} E[(\boldsymbol{\lambda}'\mathbf{q})^l].$$

We now show how to calculate c_{k+1} based on $c_1, \ldots, c_k$. Let $f(t) = \sum_{l=0}^{\infty} c_l t^l$ be the right side of the equation and let $g(t) = -\sum_{i=1}^{n} \alpha_i \log(1 - t\lambda_i)$ be the log of the left side. By taking derivatives with respect to t and evaluating at zero, we have

$$f^{(l)}(0) = \Gamma(l+1)c_l, \qquad g^{(l)}(0) = \Gamma(l)\sum_{i=1}^{n} \alpha_i \lambda_i^l.$$

Because $f(t) = e^{g(t)}$, we have

$$f^{(k+1)}(0) = \sum_{l=0}^{k} \binom{k}{l} f^{(l)}(0) g^{(k+1-l)}(0).$$

Plugging the values of $f^{(l)}(0)$ into the sum gives

$$c_{k+1} = \sum_{l=0}^{k} \left[c_l \binom{k}{l} \frac{\Gamma(l+1)\Gamma(k+1-l)}{\Gamma(k+2)} \left(\sum_{i=1}^{n} \alpha_i \lambda_i^{k+1-l} \right) \right].$$

Simplifying gives

$$E[(\boldsymbol{\lambda}'\mathbf{q})^{k+1}]$$
$$= \sum_{l=0}^{k} \left[E[(\boldsymbol{\lambda}'\mathbf{q})^l] \frac{\Gamma(\mathbf{1}'\boldsymbol{\alpha} + l)\Gamma(k+1)}{\Gamma(\mathbf{1}'\boldsymbol{\alpha} + k + 1)} \left(\sum_{i=1}^{n} \alpha_i \lambda_i^{k+1-l} \right) \right].$$

C code with an R interface to calculate $\{E[(\boldsymbol{\lambda}'\mathbf{q})^l] : l = 0, \ldots, k\}$ is available at my website.

[Received February 2006. Revised August 2006.]

REFERENCES

Abramowitz, M., and Stegun, I. A. (eds.) (1972), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, New York: Dover.

Carlson, B. C. (1977), *Special Functions of Applied Mathematics*, New York: Academic.

Chikuse, Y. (2003), *Statistics on Special Manifolds, Lecture Notes in Statistics*, Vol. 174, New York: Springer-Verlag.

Dickey, J. M. (1983), "Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses," *Journal of the American Statistical Association*, 78, 628–637.

Edelman, A. (1988), "Eigenvalues and Condition Numbers of Random Matrices," *SIAM Journal on Matrix Analysis and Applications*, 9, 543–560.

Gabriel, K. R. (1971), "The Biplot Graphic Display of Matrices With Application to Principal Component Analysis," *Biometrika*, 58, 453–467.

——— (1978), "Least Squares Approximation of Matrices by Additive and Multiplicative Models," *Journal of the Royal Statistical Society*, 40, 186–196.

——— (1998), "Generalised Bilinear Regression," *Biometrika*, 85, 689–700.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 169–193.

Gower, J. C., and Hand, D. J. (1996), *Biplots, Monographs on Statistics and Applied Probability*, Vol. 54, London: Chapman and Hall.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

——— (2003), "Trans-Dimensional Markov Chain Monte Carlo," in *Highly Structured Stochastic Systems*, eds. P. J. Green, N. L. Hjort, and S. Richardson, *Oxford Statistical Science Series*, Vol. 27, New York: Oxford University Press, pp. 179–206. Includes part A by S. J. Godsill and part B by J. Heikkinen.

Harshman, R. A., Green, P. E., Wind, Y., and Lundy, M. E. (1982), "A Model for the Analysis of Asymmetric Data in Marketing Research," *Marketing Science*, 1, 205–242.

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286–295.

Householder, A. S., and Young, G. (1938), "Matrix Approximation and Latent Roots," *American Mathematical Monthly*, 45, 165–171.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233.

Jupp, P. E., and Mardia, K. V. (1980), "A General Correlation Coefficient for Directional Data and Related Regression Problems," *Biometrika*, 67, 163–173.

Lawley, D. N., and Maxwell, A. E. (1971), *Factor Analysis as a Statistical Method* (2nd ed.), New York: Elsevier.

Lopes, H. F., and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41–67.

Meila, M., and Xu, L. (2003), "Multiway Cuts and Spectral Clustering," Technical Report 442, University of Washington.

Minka, T. P. (2000), "Automatic Choice of Dimensionality for PCA," Technical Report 514, Massachusetts Institute of Technology, Media Lab.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression" (with comments and rejoinder), *Journal of the American Statistical Association*, 83, 1023–1036.

Rajan, J. J., and Rayner, P. J. W. (1997), "Model Order Selection for the Singular Value Decomposition and the Discrete Karhunen–Loeve Transform Using a Bayesian Approach," *Vision, Image and Signal Processing, IEE Proceedings*, 144, 116–123.

Smith, P. J. (1995), "A Recursive Formulation of the Old Problem of Obtaining Moments From Cumulants and vice versa," *The American Statistician*, 49, 217–218.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion and rejoinder), *The Annals of Statistics*, 22, 1701–1762.