Supplemental materials for this article are available through the JCGS web page at http://www.amstat.org/publications.

Simulation of the Matrix Bingham–von Mises–Fisher Distribution, With Applications to Multivariate and Relational Data

Peter D. HOFF

Orthonormal matrices play an important role in reduced-rank matrix approximations and the analysis of matrix-valued data. A matrix Bingham–von Mises–Fisher distribution is a probability distribution on the set of orthonormal matrices that includes linear and quadratic terms in the log-density, and arises as a posterior distribution in latent factor models for multivariate and relational data. This article describes rejection and Gibbs sampling algorithms for sampling from this family of distributions, and illustrates their use in the analysis of a protein–protein interaction network. Supplemental materials, including code and data to generate all of the numerical results in this article, are available online.

Key Words: Bayesian inference; Eigenvalue decomposition; Markov chain Monte Carlo; Random matrix; Social network; Stiefel manifold.

1. INTRODUCTION

Unit vectors and orthonormal matrices play an important role in directional statistics, multivariate analysis, and matrix decomposition methods. For $R \leq m$ the set of $m \times R$ matrices having orthonormal columns is called the Stiefel manifold and is denoted $\mathcal{V}_{R,m} = \{X \in \mathbb{R}^{m\times R} : X^T X = I_{R\times R}\}$. Probability distributions and statistical inference for data from this manifold have been developed primarily in the literature on directional data, particularly for the case of points on a sphere ($R = 1, m = 3$). Theoretical treatments of probability distributions on higher dimensional manifolds have been given in Gupta and Nagar (2000) and Chikuse (2003).

Many of the commonly used probability distributions on $\mathcal{V}_{R,m}$ have exponential family forms. For example, a flexible class of probability densities having linear and quadratic terms is given by

$$p_{\text{BMF}}(X|A, B, C) \propto \text{etr}(C^T X + BX^T AX), \tag{1.1}$$

Peter D. Hoff is Professor, Departments of Statistics, Biostatistics and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA 98195-4322 (E-mail and Web: *hoff@stat.washington.edu*, *http://www.stat.washington.edu/~hoff*).

Journal of Computational and Graphical Statistics, Volume 18, Number 2, Pages 438–456
DOI: 10.1198/jcgs.2009.07177

where A and B are generally taken to be symmetric matrices and $\text{etr}(Z)$ is $\exp(\sum_i Z_{[i,i]})$, the exponential of the trace of Z. This family of densities, introduced by Khatri and Mardia (1977), is called the matrix Bingham–von Mises–Fisher (BMF) family or the matrix Langevin–Bingham family. The density (1.1) is expressed with respect to the uniform probability measure on X, which is simply the measure proportional to surface area on the $mR - R(R+1)/2$ dimensional surface of $\mathcal{V}_{R,m}$ in $\mathbb{R}^{mR}$. Special subfamilies of the BMF densities include the Bingham (B) densities for which $C = 0$ and the von Mises–Fisher (MF) densities for which A or B is zero. In what follows we will sometimes refer to these densities and their corresponding distributions as $\text{BMF}(A, B, C)$, $\text{B}(A, B)$, and $\text{MF}(C)$.

Relevant to the modeling of directional data, Wood (1987) described a method for simulating from densities of this type in cases where $R = 1$ and $m = 3$, and Kent, Constable, and Er (2004) described methods for a complex version of this distribution that is feasible for $R = 1$ and small values of m. Kume and Walker (2006) described a Gibbs sampling scheme for the case $R = 1$ and $C = 0$. However, there is a need for statistical and computational tools for data from higher dimensional manifolds: Large, matrix-variate datasets are frequently analyzed and described using matrix decomposition techniques, in which heterogeneity across rows and columns is represented by low-rank orthonormal eigenvector matrices. Probability models for these matrices provide a framework for describing variability and uncertainty in matrix-variate data. In particular, distributions of the form (1.1) arise as posterior distributions in many models for multivariate and relational data.

Example (Factor analysis)**:** Let Y be an $n \times p$ data matrix representing n observations from a p-variate distribution. If p is large or the columns are highly correlated, it may be desirable to represent y_i, the p measurements within row i, as linear functions of $R < p$ latent factors $u_i = (u_{i,1}, \ldots, u_{i,R})^T$:

$$y_{i,j} = u_i^T D v_j + \epsilon_{i,j},$$
$$Y = UDV^T + E.$$

In this parameterization, D is an $R \times R$ diagonal matrix of positive numbers and the $n \times R$ and $p \times R$ matrices U and V can be assumed to be orthonormal matrices, elements of $\mathcal{V}_{R,n}$ and $\mathcal{V}_{R,p}$, respectively. In situations involving ordinal or missing data it may be desirable to take a likelihood-based approach to estimation of U, D, and V. If the error matrix E is made up of independent and identically distributed normal variates, then uniform prior distributions for U and V imply that

$$p(U|Y, D, V) \propto \text{etr}([YVD]^T U/\sigma^2),$$
$$p(V|Y, D, U) \propto \text{etr}([Y^T UD]^T V/\sigma^2),$$

which are matrix von Mises–Fisher densities, $\text{MF}(YVD/\sigma^2)$ and $\text{MF}(Y^T UD/\sigma^2)$, respectively. Joint posterior inference for U and V can be obtained by iteratively sampling from these two distributions.

Example (Principal components)**:** Again, let Y be an $n \times p$ data matrix where the rows are assumed to be independent observations from a mean-zero p-variate normal population with covariance matrix Σ. Writing Σ via its eigenvalue decomposition $\Sigma = U\Lambda U^T$,

the probability density of the data is

$$p(Y|\Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\,\mathrm{etr}(-Y\Sigma^{-1}Y^T/2)$$
$$= (2\pi)^{-np/2}\prod_{j=1}^{p}\lambda_j^{-n/2}\,\mathrm{etr}(-\Lambda^{-1}U^TY^TYU/2).$$

A uniform or matrix von Mises–Fisher prior distribution on $U \in \mathcal{V}_{p,p}$ results in a Bingham–von Mises–Fisher posterior distribution. This result could be useful if one wanted to use a nonstandard prior distribution for the eigenvalues of Σ, or if there was prior information about the principal components.

Example (Network data)**:** Network data consist of binary measurements on pairs of objects or nodes. Such data are often represented as a graph in which a link between two nodes indicates the presence of a relationship of some kind. Alternatively, the data can be represented with a binary matrix Y so that $y_{i,j}$ is the 0–1 indicator of a link between nodes i and j (the diagonal of Y is generally undefined). One approach to modeling such data is to use a latent factor model with a probit link:

$$y_{i,j} = \delta_{(c,\infty)}(z_{i,j}),$$
$$z_{i,j} = u_i^T\Lambda u_j + \epsilon_{i,j},$$
$$Z = U\Lambda U^T + E,$$

where E is modeled as a symmetric matrix of independent standard normal noise, Λ is a diagonal matrix, and U is an element of $\mathcal{V}_{R,m}$, with R generally taken to be much smaller than m. Such a model can be thought of as a latent eigenvalue decomposition for the graph Y. Given a uniform prior distribution for U, we have

$$p(U|Z,\Lambda) \propto \mathrm{etr}(Z^TU\Lambda U^T/2)$$
$$= \mathrm{etr}(\Lambda U^TZU/2),$$

which is a Bingham density $\mathrm{B}(Z/2,\Lambda)$.

Section 2 of this article describes a rejection sampling method for the MF distribution, and a Gibbs sampling algorithm is provided for cases in which the rejection method is infeasible. Section 3 presents a Gibbs sampling algorithm for generating random matrices of arbitrary dimension from the BMF distribution. Specifically, I show how to construct a Markov chain in $X \in \mathcal{V}_{R,m}$ having values that converge in distribution to p_{BMF}. Section 4 implements the sampling algorithms in the context of a data analysis of the interaction network of 270 proteins. In this example the ability to sample from the BMF distribution allows for Bayesian inference and estimation. A discussion follows in Section 5.

2. SAMPLING FROM THE VON MISES–FISHER DISTRIBUTION

When $R = 1$ the Stiefel manifold $\mathcal{V}_{1,m}$ is simply the unit sphere $\mathcal{S}_m$ in $\mathbb{R}^m$, and we denote an element of $\mathcal{V}_{1,m}$ simply as a unit vector x. We refer to the MF distribution on

$\mathcal{V}_{1,m}$ as the vector von Mises–Fisher distribution, and the distribution on $\mathcal{V}_{R,m}$, $R > 1$ as the matrix von Mises–Fisher distribution.

The vector von Mises–Fisher distribution over $\mathcal{S}_m$ has a density with respect to the uniform distribution given by

$$p_{\mathrm{MF}}(x|c) = \frac{\|c/2\|^{m/2-1}}{\Gamma(m/2)I_{m/2-1}(\|c\|)} \exp\{c^T x\}, \qquad x \in \mathcal{S}_m$$

(Fisher 1953; Watson and Williams 1956), where $I_\nu(\cdot)$ is a modified Bessel function of the first kind. Wood (1994) provided a sampling scheme for this distribution that is relatively straightforward to implement: Writing $c = \kappa\mu$ for a positive scalar κ and a unit vector μ, Wood's procedure is to first sample a unit vector v uniformly from the sphere in $\mathbb{R}^{m-1}$ and a scalar random variable $w \in (-1, 1)$ from the density proportional to $(1 - w^2)^{(m-3)/2} e^{\kappa w}$ using rejection sampling. The unit vector $z = ([1 - w^2]^{1/2} v^T, w)^T$ then has the von Mises–Fisher distribution with parameter $\tilde{c} = (0, \ldots, 0, \kappa)^T$, and the vector $x = Az$ has the von Mises–Fisher distribution with parameter $c = \kappa\mu$ if A is any orthogonal matrix whose last column is μ. In this section we show how the ability to sample from this vector MF distribution can be used to sample from the matrix MF distribution, having density

$$p_{\mathrm{MF}}(X|C) = \left[{}_0F_1\left(\tfrac{1}{2}m; \tfrac{1}{4}D^2\right)\right]^{-1} \operatorname{etr}(C^T X), \qquad X \in \mathcal{V}_{R,m},$$

where D is the diagonal matrix of singular values of C and ${}_0F_1(\frac{1}{2}m; \frac{1}{4}D^2)$ is a hypergeometric function with a matrix argument (Herz 1955). We first present a rejection sampling approach in which a rejection envelope is constructed out of a product of vector MF densities. A second approach is by iterative Gibbs sampling of the columns of X.

2.1 A REJECTION SAMPLING SCHEME

An important reparameterization of the matrix MF density on $\mathcal{V}_{R,m}$ is via the singular value decomposition of C, whereby $C = UDV^T$, with U and V being $m \times R$ and $R \times R$ orthonormal matrices, and D a diagonal matrix with positive entries. Using these parameters, the density of X can be written as $p_{\mathrm{MF}}(X|U, D, V) \propto \operatorname{etr}(VDU^T X) = \operatorname{etr}(DU^T XV)$. This density is maximized at $X = UV^T$, which can be interpreted as the modal orientation of the population. The entries of D can be interpreted as concentration parameters, describing how close a random matrix X is likely to be to the mode.

As pointed out in section 2.5 of Chikuse (2003), a random draw from the matrix MF distribution can be generated by rejection sampling using a uniform envelope: Because the density is maximized in X by UV^T, the ratio between p_{MF} and the uniform density g_u on $\mathcal{V}_{R,m}$ is bounded by

$$\frac{p_{\mathrm{MF}}(X)}{g_u(X)} < \frac{\operatorname{etr}(DU^T[UV^T]V)}{{}_0F_1(\frac{1}{2}m; \frac{1}{4}D^2)} = \frac{\operatorname{etr}(D)}{{}_0F_1(\frac{1}{2}m; \frac{1}{4}D^2)}.$$

If independent pairs $X \sim g_u$ and $u \sim \mathrm{uniform}(0, 1)$ are repeatedly sampled until $u < \operatorname{etr}(C^T X - D)$, then the resulting X has density p_{MF}. As Chikuse pointed out, such a procedure will be extremely inefficient for most C of interest, due to the poor approximation of p_{MF} by g_u.

A much better approximation to p_{MF} can be constructed from a product of vector von Mises–Fisher densities, which can be used to put the columns of X near the columns of the modal orientation with high probability: If H is an $m \times R$ matrix with orthogonal columns, then the mode of $\mathrm{MF}(H)$ is H and so a random draw of X from $\mathrm{MF}(H)$ will generally be such that $X_{[,r]}$ is near $H_{[,r]}$ for each r. A value of $X_{[,r]}$ near $H_{[,r]}$ can be generated from an $\mathrm{MF}(H_{[,r]})$ distribution, but doing this independently for each r would not generate an orthonormal matrix X. However, we can generate an orthonormal X near H using a sequence of dependent vector MF distributions, so that each $X_{[,r]}$ is orthogonal to $X_{[,1]}, \ldots, X_{[,r-1]}$, and generated to be in the direction of $H_{[,r]}$ to the extent that $X_{[,1]}, \ldots, X_{[,r-1]}$ do not already lie in this direction. This suggests generating a proposal value $X \in \mathcal{V}_{R,m}$ as follows:

1. Sample $X_{[,1]} \sim \mathrm{MF}(H_{[,1]})$.
2. For $r \in \{2, \ldots, R\}$:
   (a) construct N_r, an orthonormal basis for the null space of $X_{[,(1,\ldots,r-1)]}$;
   (b) sample $z \sim \mathrm{MF}(N_r^T H_{[,r]})$;
   (c) set $X_{[,r]} = N_r z$.

Recall that the null space of a matrix A is the set of vectors x such that $Ax = 0$. Here and in what follows, an orthonormal basis for the null space of an $m \times k$ matrix A with rank k refers to any $m \times (m-k)$ orthonormal matrix N such that $N^T A = 0_{(m-k)\times k}$. Any vector x that is orthogonal to A can be expressed as a linear combination of the columns of N.

Straightforward calculations show that the probability density of the matrix X generated in steps 1 and 2 above can be expressed as

$$g(X) = \left\{ \prod_{r=1}^{R} \frac{\|N_r^T H_{[,r]}/2\|^{(m-r-1)/2}}{\Gamma\left(\frac{m-r+1}{2}\right) I_{(m-r-1)/2}(\|N_r^T H_{[,r]}\|)} \right\} \mathrm{etr}\{H^T X\},$$

with respect to the uniform density on $\mathcal{V}_{R,m}$. The ratio of the matrix MF density to g is then

$$\begin{aligned} \frac{p_{\mathrm{MF}}(X)}{g(X)} &= {}_0F_1\left(\frac{1}{2}m; \frac{1}{4}D^2\right)^{-1} \\ &\quad \times \left\{ \prod_{r=1}^{R} 2^{(m-r-1)/2} \Gamma\left(\frac{m-r+1}{2}\right) \frac{I_{(m-r-1)/2}(\|N_r^T H_{[,r]}\|)}{\|N_r^T H_{[,r]}\|^{(m-r-1)/2}} \right\}. \end{aligned}$$

Because $I_k(x)/x^k = 2^{-k} \sum_{l=0}^{\infty} (x^2/4)^l / [\Gamma(k+l+1)\Gamma(l+1)]$ is increasing in $|x|$, and $\|N_r H_{[,r]}\| \le \|H_{[,r]}\|$, we have

$$\begin{aligned} \frac{p_{\mathrm{MF}}(X)}{g(X)} &< {}_0F_1\left(\frac{1}{2}m; \frac{1}{4}D^2\right)^{-1} \\ &\quad \times \left\{ \prod_{r=1}^{R} 2^{(m-r-1)/2} \Gamma\left(\frac{m-r+1}{2}\right) \frac{I_{(m-r-1)/2}(\|H_{[,r]}\|)}{\|H_{[,r]}\|^{(m-r-1)/2}} \right\} \\ &= K(H). \end{aligned}$$

For any particular X, the ratio of $p_{\mathrm{MF}}(X)/g(X)$ to this upper bound is

$$\frac{p_{\mathrm{MF}}(X)}{g(X)K(H)} = \prod_{r=2}^{R} \frac{I_{(m-r-1)/2}(\|N_r^T H_r\|)}{I_{(m-r-1)/2}(\|H_r\|)} \left(\frac{\|H_r\|}{\|N_r^T H_r\|} \right)^{(m-r-1)/2}.$$

We use this bound and ratio in the rejection sampler described below. If H does not have orthogonal columns, then although the bound is sharp the ratio may be extremely small: Consider the case in which $H_{[,1]}$ and $H_{[,2]}$ are both equal to the vector a. In this case, a draw of X from p_{MF} will have one or the other of its first two columns close to the direction of a with equal probability, whereas draws from g will generally have $X_{[,1]}$ close to the direction of a and $X_{[,2]}$ orthogonal to it. In this case, $\|N_2^T H_{[,2]}\|$ will be much smaller than $\|H_{[,2]}\|$, making the ratio quite small. The remedy to this problem is quite simple: To sample from the matrix MF distribution with a concentration matrix C having nonorthogonal columns, first compute the singular value decomposition $C = UDV^T$, sample a matrix $Y \sim \mathrm{MF}(UD)$ using the above described rejection scheme, then set $X = YV^T$. This procedure is summarized as follows:

1. Obtain the singular value decomposition UDV^T of C and let $H = UD$.
2. Sample pairs $\{u, Y\}$ until $u < \frac{p_{\mathrm{MF}}(Y)}{g(Y)K(H)}$, using the following scheme:
   (a) Sample $u \sim \mathrm{uniform}(0, 1)$.
   (b) Sample $Y_{[,1]} \sim \mathrm{MF}(H_{[,1]})$, and for $r \in \{2, \ldots, R\}$ consecutively,
      i. let N_r be an orthonormal basis for the null space of $Y_{[,(1,\ldots,r-1)]}$;
      ii. sample $z \sim \mathrm{MF}(N_r^T H_{[,r]})$;
      iii. set $Y_{[,r]} = N_r z$.
3. Set $X = YV^T$.

To examine the feasibility of this rejection scheme a small simulation study was performed for $m \in \{10, 20, 200\}$ and $R \in \{2, 4, 6\}$. For each combination of m and R, three $m \times R$ matrices were constructed, each having R singular values all equal to $m/2, m$, or $2m$. One hundred observations from each MF distribution were generated using the above rejection scheme, and the average number of rejected draws is recorded in Table 1. In general, the results indicate that the rejection sampler is a feasible method for sampling from the MF distribution for a broad range of values of m, R, and D. However, we see that the average number of rejected draws is generally increasing in the magnitude of the elements of D and the ratio R/m. For large values of this latter ratio, $\|N_r^T H_{[,r]}\|$ is typically a small fraction of $\|H_{[,r]}\|$ and the ratio $p_{\mathrm{MF}}(X)/g(X)$ is then rarely close to the bound. Similarly, large D leads to large differences between $\|N_r^T H_{[,r]}\|$ and $\|H_{[,r]}\|$ for higher values of r.

2.2 A GIBBS SAMPLING SCHEME

For large values of D or R the rejection sampler described above may be prohibitively slow. As an alternative, a simple Gibbs sampling scheme can be used to construct a de-

Table 1. Average number of rejected draws needed to generate a single observation from MF(C), for various values of m, R, and singular values d.

	$m=10$				$m=20$				$m=200$		
d	$R=2$	$R=4$	$R=6$	d	$R=2$	$R=4$	$R=6$	d	$R=2$	$R=4$	$R=6$
5	0.07	0.76	5.08	10	0.11	0.63	3.80	100	0.03	0.68	2.28
10	0.09	2.40	26.52	20	0.24	1.78	15.25	200	0.20	1.45	10.04
20	0.38	4.01	77.74	40	0.30	3.65	42.70	400	0.35	3.40	36.52

pendent Markov chain $\{X^{(1)}, X^{(2)}, \ldots\}$ such that $X^{(s)}$ converges in distribution to p_{MF} as $s \to \infty$. Such a sequence can be constructed by iteratively sampling the columns of X from their full conditional distributions.

The density of $X \sim \text{MF}(C)$ can be expressed in terms of a product over the columns of X:

$$p_{\text{MF}}(X|C) \propto \text{etr}(C^T X) = \prod_{r=1}^{R} \exp\bigl(C_{[,r]}^T X_{[,r]}\bigr).$$

The columns are not statistically independent of course, because they are orthogonal with probability 1. As such, we can rewrite X as $X = \{X_{[,1]}, X_{[,-1]}\} = \{Nz, X_{[,-1]}\}$, where $z \in \mathcal{S}_{m-R+1}$ and N is an $m \times (m-R+1)$ orthonormal basis for the null space of $X_{[,-1]}$. Note that $N^T N = I_{m-R+1}$ and so $z = N^T X_{[,1]}$. Following chapter 3 of Chikuse (2003), the conditional distribution of z given $X_{[,-1]}$ is given by

$$p\bigl(z|X_{[,-1]}\bigr) \propto \exp\bigl(C_{[,1]}^T N z\bigr) = \exp(\tilde{c}^T z),$$

which is a vector MF density. This fact can be exploited to implement a Gibbs sampler that generates a Markov chain in X. Given a current value $X^{(s)} = X$, the sampler proceeds by construction of a new value $X^{(s+1)}$ as follows:

Given $X^{(s)} = X$, perform steps 1, 2, and 3 for each $r \in \{1, \ldots, R\}$ in random order:

1. let N be an orthonormal basis for the null space of $X_{[,-r]}$;
2. sample $z \sim \text{MF}(N^T C_{[,r]})$ using the sampling scheme of Wood (1994);
3. set $X_{[,r]} = Nz$.

Set $X^{(s+1)} = X$.

Iteration of this algorithm generates a reversible aperiodic Markov chain $\{X^{(1)}, X^{(2)}, \ldots\}$. If $m > R$, this Markov chain is also irreducible and $X^{(s)}$ converges in distribution to MF(C) as $s \to \infty$. To see that the chain is irreducible, we will first show how it is possible to move between any values of $(X_{[,1]}, X_{[,2]})$ in the null space of $(X_{[,3]}, \ldots, X_{[,R]})$, changing only one of the two columns at a time using steps 1, 2, and 3 of the Gibbs sampler. Letting $(X_{[,1]}^a, X_{[,2]}^a)$ and $(X_{[,1]}^b, X_{[,2]}^b)$ be two potential values of $(X_{[,1]}, X_{[,2]})$, $X_{[,1]}$ can first move from $X_{[,1]}^a$ to any point $\tilde{X}_{[,1]}$ in the null space of $(X_{[,2]}^a, X_{[,2]}^b, X_{[,3]}, \ldots, X_{[,R]})$. Such a move is possible because the dimension of this null space is $m - R > 0$ and the only

constraint on $\tilde{X}_{[,1]}$ in the Gibbs sampler is that it is orthogonal to $(X^a_{[,2]}, X_{[,3]}, \ldots, X_{[,R]})$. The next step is to move $X_{[,2]}$ from $X^a_{[,2]}$ to $X^b_{[,2]}$. This move is possible because $X^b_{[,2]}$ is orthogonal to $(\tilde{X}_{[,1]}, X_{[,3]}, \ldots, X_{[,R]})$. Finally, $X_{[,1]}$ can move from $\tilde{X}_{[,1]}$ to $X^b_{[,1]}$. Proceeding by induction, it can be shown that it is possible to go from any state of $(X_{[,1]}, \ldots, X_{[,k]})$ to any other for $k \leq R$, giving the result.

However, if $m - R = 0$, then the chain is reducible. This is because in this case the null space of $X_{[,-r]}$ is one dimensional and therefore the value of z in step 2 of the algorithm satisfies $z \in \{-1, 1\}$, and so states in the Markov chain will remain fixed up to column-wise changes in sign. The remedy for this situation is not too difficult: An irreducible Markov chain for the case $m = R$ can be constructed by sampling two columns of X at a time. Details of such a procedure are given in the context of the more general Bingham–von Mises–Fisher distribution in the next section.

Finally, we note that nonorthogonality among the columns of C can add to the autocorrelation in the Gibbs sampler. For example, if $C_{[,1]} = C_{[,2]} = a$ for some vector a with a large magnitude, then $X_{[,1]}$ and $X_{[,2]}$ have equally high probabilities of being in the direction of a. But because the Gibbs sampler works by generating values from full conditionals and $X^T_{[,1]}X_{[,2]} = 0$ with probability 1, once $X_{[,1]}$ is in the direction of a then $X_{[,2]}$ must be sampled as orthogonal to $X_{[,1]}$ and thus nearly orthogonal to a. If $|a|$ is large, then moving away from this local mode is unlikely, resulting in a poorly mixing Markov chain. The simple remedy to this situation is similar to that taken for the rejection sampler in Section 2.1: Perform the Gibbs sampler on $Y \sim \text{MF}(UD)$ where $X = UDV^T$, and record values of $X = YV^T$.

3. SAMPLING FROM THE BINGHAM–VON MISES–FISHER DISTRIBUTION

In this section a Markov chain Monte Carlo method for sampling from p_{BMF} is derived. Similar to the approach in Section 2.2, the method involves iteratively resampling each column of X given the others using full conditional distributions, thereby generating a Markov chain $\{X^{(1)}, X^{(2)}, \ldots\}$ whose stationary distribution is p_{BMF}. We first present a method for sampling from the vector BMF distribution on $x \in \mathcal{S}_m$, and then show how this can be used iteratively to sample from the matrix BMF distribution.

3.1 THE VECTOR BINGHAM DISTRIBUTION

The Bingham distribution on the unit sphere in $\mathbb{R}^m$ has a density with respect to the uniform distribution given by

$$p_B(x|A) \propto \exp(x^T Ax), \qquad x \in \mathcal{S}_m.$$

Because $x^T Ax = x^T A^T x = \frac{1}{2}x^T(A + A^T)x$, A can be assumed to be symmetric. Let the eigenvalue decomposition of A be $A = E\Lambda E^T$, and let $y = E^T x$. Because E is orthogonal

the change of variables formula gives the density of y as

$$\begin{aligned} p(y|E,\Lambda) &= c(\Lambda)\exp(y^T E^T E\Lambda E^T Ey) \\ &= c(\Lambda)\exp(y^T\Lambda y) \\ &\propto \exp\left(\sum_{i=1}^m \lambda_i y_i^2\right). \end{aligned}$$

Again, this density is with respect to the uniform density on the sphere. If we write $y_m^2 = 1-\sum_{i=1}^{m-1} y_i^2$, the uniform density in terms of $\{y_1,\ldots,y_{m-1}\}$ is proportional to $|y_m|^{-1} = (1-\sum_{i=1}^{m-1} y_i^2)^{-1/2}$ and so the above density with respect to Lebesgue measure on $(y_1,\ldots,y_{m-1})$ becomes

$$p(y|E,\Lambda) \propto \exp\left(\sum_{i=1}^m \lambda_i y_i^2\right)|y_m|^{-1}, \qquad y_m^2 = 1-\sum_{i=1}^{m-1} y_i^2 \in [0,1].$$

We now consider MCMC sampling for the vector y. One possibility would be to construct a Gibbs sampler by iteratively sampling components of y from their full conditional distributions. Although the full conditionals are nonstandard, Kume and Walker (2006) gave a clever auxiliary variable method that allows for Gibbs sampling and requires only that we are able to sample uniformly from an interval. However, although easy to implement for the vector Bingham distribution, the auxiliary variable approach is difficult to extend to the more complicated Bingham–von Mises–Fisher distribution, discussed in the next subsection. Additionally, the auxiliary variable Gibbs sampler can also result in a slowly mixing Markov chain because the full conditionals can be highly constrained. For example, the full conditional density of y_1 given $y_2,\ldots,y_{m-1}$ is nonzero only on $y_1^2 < 1-\sum_2^{m-1} y_i^2$.

We now consider an alternative to Gibbs sampling that will result in a better-mixing Markov chain and that can also be easily extended to sample from the BMF distribution. The idea is to reparameterize $\{y_1,\ldots,y_m\}$ for each i so that we can update each y_i in a less constrained way. For example, to update y_1 let $\theta = y_1^2$ and $q = \{y_1^2/(1-y_1^2),\ldots,y_m^2/(1-y_1^2)\}$, so that $\{y_1^2,\ldots,y_m^2\} = \{\theta,(1-\theta)q_{-1}\}$. Rather than sampling y_1^2 given $y_2^2,\ldots,y_{m-1}^2$, we can sample a new value of $y_1^2 = \theta \in (0,1)$ conditional on q_{-1}, thereby allowing for larger redistributions of the "mass" of $\{y_1^2,\ldots,y_m^2\}$ while ensuring that $\sum_1^m y_i^2 = 1$.

Keeping in mind that $y_m^2 = 1-\sum_1^{m-1} y_i^2$ and $q_m = 1-\sum_2^{m-1} q_i$, the joint density of $\{\theta,q_2,\ldots,q_{m-1}\}$ can be obtained from that of $\{y_1,\ldots,y_{m-1}\}$ as follows:

$$\left|\frac{d\theta}{dy_1}\right| = 2|y_1| = 2\theta^{1/2}, \qquad \left|\frac{dq_i}{dy_i}\right| = 2\frac{|y_i|}{1-y_1^2} = 2q_i^{1/2}(1-\theta)^{-1/2}, \qquad i>1$$

and so

$$\begin{aligned} p(\theta,q_{-1}) &\propto \exp(\theta\lambda_1 + (1-\theta)q_{-1}^T\lambda_{-1}) \times \left|\frac{d\{\theta,q_{-1}\}}{d\{y_1,\ldots,y_{m-1}\}}\right|^{-1}[(1-\theta)q_m]^{-1/2} \\ &\propto \exp(\theta\lambda_1 + (1-\theta)q_{-1}^T\lambda_{-1}) \times \theta^{-1/2}(1-\theta)^{(m-3)/2}\prod_{i=2}^m q_i^{-1/2}, \\ p(\theta|q_{-1}) &\propto \exp(\theta[\lambda_1 - q_{-1}^T\lambda_{-1}]) \times \theta^{-1/2}(1-\theta)^{(m-3)/2}. \end{aligned}$$

Sampling $\theta \in (0, 1)$ can proceed with a rejection scheme. The density we need to sample from is of the form $p(\theta) \propto \theta^{-1/2}(1-\theta)^k e^{\theta a}$. If k is larger than a, then a beta$(1/2, k)$ distribution works well as an envelope. If on the other hand a is much larger than k, then $p(\theta)$ has a substantial local mode close to 1. I have found that a beta$(1/2, 1 + k \wedge [(k - a) \vee -1/2])$ envelope distribution works well for a wide range of values of a and k. Further details of a rejection sampling scheme based on this envelope distribution for θ are available in the R-software companion to this article.

Iteration of the procedure described above, with $\theta = y_i^2$ for each $i \in \{1, \ldots, m\}$ and a similarly redefined q, generates a Markov chain in $\{y_1^2, \ldots, y_m^2\}$ with a stationary distribution equal to $p(y_1^2, \ldots, y_m^2|E, \Lambda)$. The signs of the y_i's do not affect the density and can each be randomly and independently assigned to be positive or negative with equal probability. The value of x is then obtained from $x = Ey$. To summarize, our Markov chain Monte Carlo algorithm for generating observations from $p_B(x|A)$ is to iterate the following algorithm:

Given $A = E^T \Lambda E$ and a current value of $x^{(s)} = x$,

1. Compute $y = E^T x$.
2. Perform steps (a)–(d) for each $i \in \{1, \ldots, m\}$ in random order:
   (a) let $\{q_1, \ldots, q_m\} = \{y_1^2/(1 - y_i^2), \ldots, y_m^2/(1 - y_i^2)\}$;
   (b) sample $\theta \in (0, 1)$ from the density proportional to $e^{\theta(\lambda_i - q_{-i}^T \lambda_{-i})} \times \theta^{-1/2}(1 - \theta)^{(m-3)/2}$;
   (c) sample s_i uniformly from $\{-1, +1\}$;
   (d) set $y_i = s_i \theta^{1/2}$ and for each $j \neq i$, set $y_j^2 = (1 - \theta)q_j$ leaving the sign unchanged.
3. Set $x = Ey$.

Set $x^{(s+1)} = x$.

Although $\theta = y_i^2$ is being sampled from a full conditional distribution, a different parameterization of $\{y_1^2, \ldots, y_m^2\}$ is being used for each i and so this is not a standard Gibbs sampling algorithm. For each i, the algorithm above can be seen as sampling from the distribution of y conditional on $f_i(y) = \{y_1^2/(1 - y_i^2), \ldots, y_{i-1}^2/(1 - y_i^2), y_{i+1}^2/(1 - y_i^2), \ldots, y_m^2/(1 - y_i^2)\}$. For each i the steps defined in (a)–(d) therefore define a transition kernel $J_i(\tilde{y}|y) = p_B(\tilde{y}|\Lambda, f_i(\tilde{y}) = f_i(y))$, each of which has the target distribution as its invariant distribution. Iteratively sampling from each J_i as in step 2 defines a composite transition kernel which has the same invariant distribution as each J_i but is also irreducible. As described in Tierney (1994), section 2, this type of procedure defines a valid Metropolis–Hastings algorithm.

A small simulation study was run to compare the Metropolis–Hastings algorithm described in steps 1, 2, and 3 above to the auxiliary variable Gibbs sampler from Kume and Walker (2006). For each combination of values of $k \in \{2, 4, 8\}$ and $g \in \{0, 5, 10\}$ we generate observations from $p_B(x|\Lambda_{k,g})$ where $\Lambda_{k,g}$ is an $m \times m$ diagonal matrix of eigenvalues

Table 2. Efficiency comparisons of the Metropolis–Hastings algorithm to the auxiliary Gibbs method.

	$m = 20$		$m = 40$		$m = 80$	
g	$n_{\text{eff}}^{\text{mh}}/n_{\text{eff}}^{\text{gibbs}}$	$r^{\text{mh}}/r^{\text{gibbs}}$	$n_{\text{eff}}^{\text{mh}}/n_{\text{eff}}^{\text{gibbs}}$	$r^{\text{mh}}/r^{\text{gibbs}}$	$n_{\text{eff}}^{\text{mh}}/n_{\text{eff}}^{\text{gibbs}}$	$r^{\text{mh}}/r^{\text{gibbs}}$
0	5.25	1.90	5.90	2.12	5.42	2.01
5	6.11	2.14	6.54	2.34	5.62	2.12
10	6.54	2.17	7.32	2.68	5.85	2.19

NOTE: For both methods, the average effective sample sizes of $x_1^2, \ldots, x_m^2$ from 5,000 autocorrelated observations were computed and their ratio appears in the first column for each m. Effective sample size per unit time was also computed, and the ratio appears in the second column for each m.

with $m = 10 \times k$. For each g, the first k eigenvalues are set equal to $g + (m, m-1, \ldots, m-k+1)/10$, and the remaining eigenvectors are set to $(m-k, \ldots, 1)/10$. In other words, for $g > 0$, $\Lambda_{k,g}$ has k big eigenvectors and $m-k$ small ones. For example, the diagonal of $\Lambda_{2,10}$ is equal to $(12.0, 11.9, 1.8, 1.7, \ldots, 0.2, 0.1)$. The Metropolis–Hastings algorithm and Kume and Walker's approach were both iterated 5,000 times to generate 5,000 dependent observations of x for each value of k and g. Densities for a variety of functions of x were compared visually across the two methods and no discrepancies were found. The efficiencies of the two algorithms can be compared using the idea of effective sample size. The effective sample size of a sequence $z_1, \ldots, z_n$ of correlated random variables having a common marginal distribution is the value n_{eff} such that $\text{var}[\bar{z}_n] = \text{var}[z_i]/n_{\text{eff}}$. For each combination of k and g and each of the two algorithms, we approximate the effective sample sizes of $x_1^2, \ldots, x_m^2$ using the function `effectiveSize` in the `coda` package of the R programming environment, and compute their average. For the Gibbs sampler of Kume and Walker, the average effective sample size was less than 1000 for all combinations of k and g, whereas the values for the Metropolis–Hastings algorithm were all around 5,000, indicating substantially less autocorrelation using this approach. The ratios of the average effective samples sizes are given in Table 2. The Metropolis–Hastings approach is roughly five to seven times more efficient than the Gibbs sampler in terms of effective sample size. However, the Metropolis–Hastings algorithm relies on a rejection sampler for θ, which may slow down the algorithm in real time. To study this, we also compute the "rate" r of effective sample size per unit system time for each of the two methods, and give their ratio in the second columns of Table 2 for each m. Using this efficiency criterion, the Metropolis–Hastings approach is roughly twice as efficient as the Gibbs approach. Although producing a less dependent Markov chain, the Metropolis–Hastings approach may take more actual time to implement. Which algorithm to implement in practice may depend on the particular Bingham distribution being simulated.

3.2 THE VECTOR BINGHAM–VON MISES–FISHER DISTRIBUTION

The vector Bingham–von Mises–Fisher density adds a linear term to the quadratic of the vector Bingham log-density, so that $p(x|A) \propto \exp(c^T x + x^T A x)$. A Gibbs sampling algorithm for this distribution can proceed in nearly the same way as for the vector Bingham distribution. For the BMF distribution, the signs of the y_i's are not uniformly distributed

on $\{-1, +1\}$, and so we parameterize y in terms of θ and q as above but additionally let $s_i = \text{sign}(y_i)$. For a given value of the vector s the transformation between (θ, q_{-1}) and y is one-to-one, and the conditional density of $\{\theta, s_1\}$ given q_{-1} and s_{-1} is

$$p(\theta, s_1|q_{-1}, s_{-1}) \propto \left\{e^{\theta(\lambda_1 - q_{-1}^T\lambda_{-1})} \times \theta^{-1/2}(1-\theta)^{(m-3)/2}\right\} \times \exp\left(\theta^{1/2}s_1 d_1 + (1-\theta)^{1/2}\left(s_{-1} \circ q_{-1}^{1/2}\right)^T d_{-1}\right),$$

where "$\circ$" is the Hadamard product (element-wise multiplication) and $d = E^T c$. A value of $\{\theta, s_1\}$ can be sampled from its full conditional distribution by first sampling $\theta \in (0, 1)$ from $p(\theta|q_{-1}, s_{-1})$ and then sampling s_1 conditional on θ. This results in the following modification of steps 2(b) and 2(c) above:

(b) sample θ from the density proportional to $p(\theta, s_i = -1|q_{-i}, s_{-i}) + p(\theta, s_i = +1|q_{-i}, s_{-i})$;

(c) sample $s_i \in \{-1, +1\}$ with probabilities proportional to $\{e^{-\theta^{1/2}d_i}, e^{+\theta^{1/2}d_i}\}$.

The density we need to sample θ from is of the form

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^k e^{\theta a + (1-\theta)^{1/2}b} \times \left(e^{-\theta^{1/2}c} + e^{\theta^{1/2}c}\right).$$

Sampling from this density can be achieved using rejection sampling. If the terms involving a or k dominate this density (as they do for the data analysis in Section 4), then the envelope distribution used in the previous subsection works well. Otherwise, using a mixture of a beta$(1/2, 1)$ and another beta density that matches the mode and curvature of $p(\theta)$ can be used.

3.3 THE MATRIX BINGHAM–VON MISES–FISHER DISTRIBUTION

To simplify notation, in what follows we assume B is a diagonal matrix. If B is symmetric but not diagonal, then the procedure can be applied to $\tilde{X} = XE$, where E are the eigenvectors of B.

Expressing the density $p_{\text{BMF}}(X|A, B, C)$ in terms of a product over the columns of X, we have

$$p_{\text{BMF}}(X|A, B, C) \propto \text{etr}(C^T X + BX^T AX) \propto \prod_{r=1}^{R} \exp\left(C_{[,r]}^T X_{[,r]} + b_{r,r} X_{[,r]}^T A X_{[,r]}\right).$$

As in Section 2, the columns are not statistically independent because they are orthogonal with probability 1. We rewrite X as $X = \{Nz, X_{[,-1]}\}$ as before, where $z \in \mathcal{S}_{m-R+1}$ and N is an $m \times (m - R + 1)$ orthonormal basis for the null space of $X_{[,-1]}$. The conditional density of z given $X_{[,-1]}$ is

$$p\left(z|X_{[,-1]}\right) \propto \exp\left(C_{[,1]}^T Nz + b_{1,1} z^T N^T ANz\right) = \exp(\tilde{c}^T z + z^T \tilde{A} z),$$

which is a vector BMF density. A Markov chain in X with stationary distribution BMF(A, B, C) can therefore be constructed as follows:

Given $X^{(s)} = X$, perform steps 1–4 for each $r \in \{1, \ldots, R\}$ in random order:

1. let N be an orthonormal basis for the null space of $X_{[,-r]}$ and let $z = N^T X_{[,r]}$;
2. compute $\tilde{c} = N^T C_{[,r]}$ and $\tilde{A} = b_{r,r} N^T A N$;
3. update the elements of z using the Gibbs sampler for the vector BMF$(\tilde{A}, \tilde{c})$ density;
4. set $X_{[,r]} = Nz$.

Set $X^{(s+1)} = X$.

Iteration of this algorithm generates a reversible aperiodic Markov chain $\{X^{(1)}, X^{(2)}, \ldots\}$ which is irreducible for $m > R$ for the same reasons described in Section 2.2, and thus converges in distribution to BMF(A, B, C). However, if $m = R$, then the chain is reducible. As described in Section 2.2, this is because in this case the null space of $X_{[,-r]}$ is one dimensional and the states in the Markov chain will remain fixed up to column-wise multiplication by -1. The remedy for this situation is to sample multiple columns of X at a time. In particular, the full conditional distribution of two columns is easy to derive. For example, an orthonormal basis N for the null space of $X_{[,-(1,2)]}$ is an $m \times 2$ matrix, and so $X_{[,(1,2)]} = NZ$ where Z is a 2×2 orthonormal matrix. The density of Z given $X_{[,-(1,2)]}$ is

$$p(Z) \propto \text{etr}(\tilde{C}^T Z + \tilde{B} Z^T \tilde{A} Z),$$

where $\tilde{C} = N^T C_{[,(1,2)]}$, $\tilde{B} = \text{diag}(b_{1,1}, b_{2,2})$, and $\tilde{A} = N^T A N$. Because Z is orthogonal, we can parameterize it as

$$Z = \begin{pmatrix} \cos\phi & s\sin\phi \\ \sin\phi & -s\cos\phi \end{pmatrix}$$

for some $\phi \in (0, 2\pi)$ and $s = \pm 1$. The second column $Z_{[,2]}$ of Z is a linear function of the first column $Z_{[,1]}$, and the uniform density on the circle is constant in ϕ, so the joint density of (ϕ, s) is simply $p(Z(\phi, s))$. Sampling from this distribution can be accomplished by first sampling $\phi \in (0, 2\pi)$ from a density proportional to $p(Z(\phi, -1)) + p(Z(\phi, +1))$, and then sampling s conditional on ϕ. To summarize, the Gibbs sampling scheme for the case $m = R$ is as follows:

Given $X^{(s)} = X$, perform steps 1–5 for each pair $(r_1, r_2) \subset \{1, \ldots, R\}$ in random order:

1. let N be an orthonormal basis for the null space of $X_{[,-(r_1,r_2)]}$;
2. compute $\tilde{C} = N^T C_{[,(r_1,r_2)]}$, $\tilde{B} = \text{diag}(b_{r_1,r_1}, b_{r_2,r_2})$, and $\tilde{A} = N^T A N$;
3. sample $\phi \in (0, 2\pi)$ from the density proportional to $p(Z(\phi, -1)) + p(Z(\phi, +1))$;
4. sample $s \in \{-1, +1\}$ with probabilities proportional to $\{p(Z(\phi, -1)), p(Z(\phi, +1))\}$;
5. set $Z = Z(\phi, s)$ and $X_{[,(r_1,r_2)]} = NZ$.

Set $X^{(s+1)} = X$.

3.4 A REJECTION SAMPLER FOR SMALL SQUARE BINGHAM MATRICES

In the case that $m = R$ and R is not too large (i.e., $R = 8$ or less) it is often feasible to implement a rejection sampler for the Bingham(A, B) distribution by generating proposals from the Wishart distribution: Let $W \sim \text{Wishart}(\nu, [\delta I - A]^{-1}/2)$, where I is the identity matrix and δ is large enough so that $[\delta I - A]$ is positive definite. Let $W = XLX^T$ be the eigenvalue decomposition of W, in which each column of X is multiplied by either -1 or 1 with equal probability (this eliminates systematic but arbitrary ways that numerical algorithms align the eigenvectors). The Wishart distribution on W induces a joint distribution on (L, X) given by $g(L, X) = g(L)g(X|L)$, where

$$\begin{aligned} g(X|L) &= \text{etr}([A - \delta I]XLX^T) \\ &= \text{etr}(LX^TAX - \delta L) \propto \text{etr}(LX^TAX). \end{aligned}$$

Thus the conditional density of X given L is a Bingham(A, L) density (see also Takemura and Sheena 2005). The appropriate acceptance ratio is given by

$$\begin{aligned} r &= \frac{p_B(X|A, B)}{g(X|L)\max_X\{p_B(X|A, B)/g(X|L)\}} \\ &= \text{etr}([B - L]X^TAX)e^{-\lambda}, \end{aligned}$$

where $\lambda = \max_X \text{tr}([B - L]X^TAX)$. It is not too hard to show that the maximum value of λ is given by $\lambda = \sum_{r=1}^{R} a_{(r)}d_{(r)}$, where $a_{(1)}, \ldots, a_{(R)}$ are the eigenvalues of A and $d_{(1)}, \ldots, d_{(R)}$ are the diagonal values of $D = B - L$, both in order from largest to smallest. The rejection algorithm is to repeatedly sample $u \sim \text{uniform}(0, 1)$ and pairs (L, X) from g until $u < r$. To see that this works, note that the joint conditional probability density of (L, X) given $u < r$ is

$$\begin{aligned} p(L, X|u < r) &\propto \Pr(u < r|L, X)g(L, X) \\ &= r \times g(L, X) \\ &= \frac{p_B(X|A, B)}{g(X|L)}e^{-\lambda} \times g(L)g(X|L) \\ &= p_B(X|A, B) \times [g(L)e^{-\lambda}], \end{aligned}$$

where λ depends on L but not on X. Thus, conditional on $u < r$, L and X are independent and the distribution of X is $p_B(X|A, B)$. In this algorithm, the values of ν and δ can be adjusted to improve the acceptance rate. Suggested values of ν and δ for given A and B are available in the online supplementary notes for this article.

4. EXAMPLE: EIGENMODEL ESTIMATION FOR NETWORK DATA

In this section we use the model for network data described in the Introduction to analyze a symmetric binary matrix of protein–protein interaction data, originally described in Butland et al. (2005). For these data, $y_{i,j} = 1$ if proteins i and j bind together and $y_{i,j} = 0$

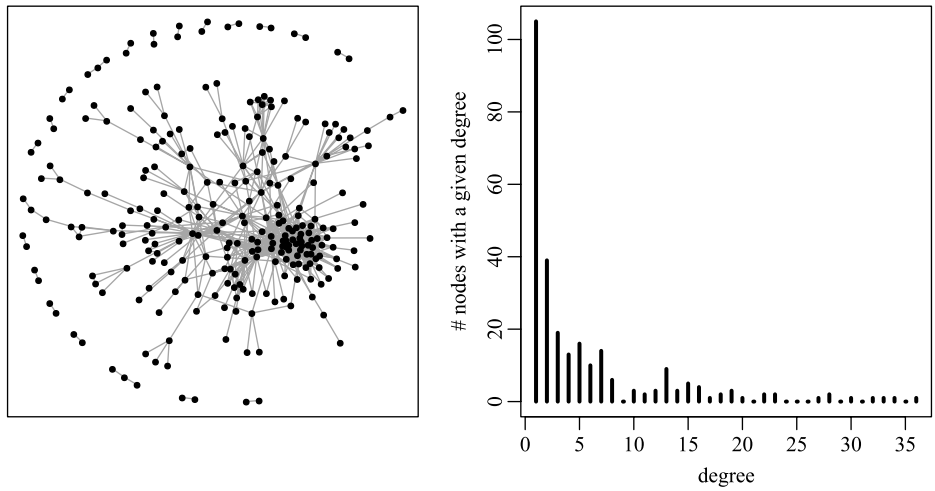

Figure 1. Descriptive plots of the protein interaction network. The first panel shows the complete dataset. The second panel gives a histogram of the degree distribution.

otherwise. The data consist of pairwise measurements among $m = 270$ essential proteins of *E. coli*. The interaction rate is $\bar{y} = 0.02$, with most nodes (53%) having only 1 or 2 links. Nevertheless, the large connected component of the graph consists of 230 of the 270 nodes, as shown in the first panel of Figure 1.

As described in the Introduction, our model for these data is essentially a latent factor model with a probit link:

$$y_{i,j} = \delta_{(c,\infty)}(z_{i,j}),$$
$$z_{i,j} = u_i^T \Lambda u_j + \epsilon_{i,j},$$
$$Z = U\Lambda U^T + E.$$

With $U \in \mathcal{V}_{R,m}$, this model can be thought of as an R-dimensional latent eigenvalue decomposition for the graph Y. Hoff (2005) discussed parameter estimation for a version of this model, and Hoff and Ward (2004) used such a model to describe networks of international relations. However, the approaches in these two articles use normal distributions for the latent factors, leading to some identifiability issues. For example, the magnitudes of the factors are confounded with the eigenvalues, and attempts to fix the scale of the u_i's lead to complicated constraints among their multivariate means, variances, and covariances.

A cleaner approach to modeling such data is to estimate $\{u_1, \ldots, u_m\}$ as being the rows of an $m \times R$ orthonormal matrix U. The probability density of a symmetric matrix Z with mean $U\Lambda U^T$ and off-diagonal unit variance is

$$p(Z|U, L) \propto \text{etr}[-(Z - U\Lambda U^T)^T (Z - U\Lambda U^T)/4]$$
$$= \text{etr}(-Z^T Z/4)\,\text{etr}(Z^T U\Lambda U^T/2)\,\text{etr}(-\Lambda^2/4).$$

We call this model a latent eigenmodel, as the parameters U and Λ are the eigenvectors and values of the mean matrix of Z. The $-1/4$ has replaced the usual $-1/2$ because Z

is symmetric. Additionally, the diagonal elements of Z have variance 2, but do not correspond to any observed data as the diagonal of Y is undefined. These diagonal elements are integrated over in the Markov chain Monte Carlo estimation scheme described below.

Using a uniform prior distribution on U and independent normal$(0, \tau^2)$ prior distributions for the elements of Λ gives

$$p(\Lambda|Z,U) = \prod_{r=1}^{R} \text{dnorm}\big(\lambda_r : \text{mean} = \tau^2 U_r^T Z U_r/(2+\tau^2), \text{var} = 2\tau^2/(2+\tau^2)\big),$$

$$\begin{aligned} p(U|Z,\Lambda) &\propto \text{etr}(Z^T U \Lambda U^T/2) \\ &= \text{etr}(\Lambda U^T Z U/2) \\ &= \text{dBMF}(U : A = Z/2, B = \Lambda, C = 0), \end{aligned}$$

where "dnorm" and "dBMF" denote the normal and BMF densities with the corresponding parameters. For any given prior $p(c)$, all $z_{i,j}$'s such that $y_{i,j} = 0$ must be below c, and all $z_{i,j}$'s for which $y_{i,j} = 1$ must be above, so the conditional distribution of c given Z and Y is proportional to $p(c)$ but restricted to the interval $(\max\{z_{i,j} : y_{i,j} = 0\}, \min\{z_{i,j} : y_{i,j} = 1\})$. For convenience we use a normal$(0, 100)$ prior distribution on c, resulting in a constrained normal as a full conditional distribution. Approximate posterior inference for U, Λ, and c can be obtained via iterative Gibbs sampling of $\{U, \Lambda, Z, c\}$ from their full conditional distributions given the data Y. One iteration of the sampling scheme consists of the following:

Given $\{U, \Lambda, Z, c\}^{(s)} = \{U, \Lambda, Z, c\}$,

1. sample the columns of U from their full conditional distributions under BMF$(Z/2, \Lambda, 0)$;
2. sample the elements of Λ from their normal conditional distributions given above;
3. sample the elements of Z from normal densities with mean $U\Lambda U^T$ but constrained to be above or below c depending on Y;
4. sample c from a constrained normal distribution.

Set $\{U, \Lambda, Z, c\}^{(s+1)} = \{U, \Lambda, Z, c\}$.

A natural choice of the prior parameter τ^2 is m, as this is roughly the variance of the eigenvalues of an $m \times m$ matrix of independent standard normal noise.

There are several reasons for fitting a statistical model to these data. First of all, the undefined diagonal $\{y_{i,i}\}$ precludes a standard eigenvalue decomposition of the original data. Second, even if the diagonal could be reasonably defined, the data are binary and so a decomposition on this raw data scale may be inappropriate. Additionally, a statistical model provides measures of uncertainty and predictive probabilities. The latter can be particularly useful in terms of outlier analysis: $\{i, j\}$ pairs for which $y_{i,j} = 0$ but $\hat{\Pr}(y_{i,j} = 1)$ is large might indicate a "missing link" and could warrant further investigation.

A three-dimensional eigenmodel was fit to the protein interaction data using the Gibbs sampling scheme described above. Specifically, two independent Gibbs samplers of length

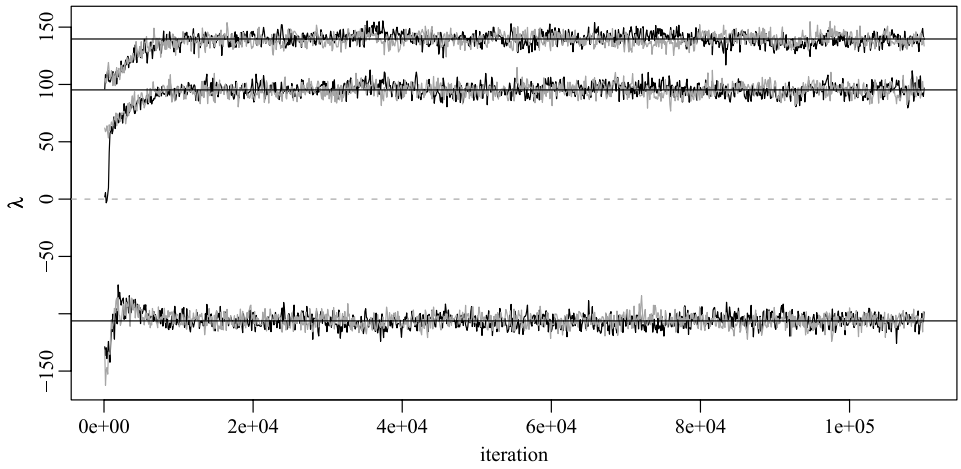

Figure 2. Samples of Λ from the two independent Markov chains.

110,000 were constructed, one initiated with random starting values and the other with values obtained from the eigenvalue decomposition of a rank-based transformation of Y, in which the ranks of tied values were randomly assigned. Both chains converged to the same region of the parameter space after a few thousand iterations. A plot of the sequences of eigenvalues from each of the chains is given in Figure 2, indicating one negative and two positive eigenvalues. For posterior analysis the first 10,000 iterations of each chain were dropped and only every 100th iteration was retained, leaving 1000 values from each of the two chains for each parameter value. From these values we can calculate the posterior mean value of $U\Lambda U^T$, which is not exactly a rank 3 matrix but is very close—its first three eigenvectors accounted for 99.95% of the sum of squares of the posterior mean. The eigenvectors corresponding to the largest eigenvalues of this mean matrix can be reasonably thought of as a posterior point estimate of U.

The eigenvectors corresponding to the two positive eigenvalues are plotted in the first panel of Figure 3, along with links between interacting protein pairs. Proteins with large values of $u_{i,1}^2 + u_{i,2}^2$ are plotted using their names. For positive eigenvalues, the interpretation of the parameters is that $(u_{i,1}, u_{i,2})$ and $(u_{j,1}, u_{j,2})$ being in the same direction contributes to the tendency for there to be an interaction between nodes i and j. Additionally, in this model a network "hub" having many connections is modeled as having a large value of $u_{i,1}^2 + u_{i,2}^2$ and makes most of its connections to proteins having factors of smaller magnitude but in the same direction.

The second panel of Figure 3 displays a different aspect of the protein network. The plot identifies two groups of proteins having large positive and large negative values of $\{u_{i,3}\}$, respectively. Members of each group are similar in the sense that they primarily interact with members of the opposite group but not with each other. The model captures this pattern with a negative eigenvalue λ_3, so that $u_{i,3}, u_{j,3}$ being large and of opposite sign is associated with a high probability of interaction between i and j. In this way, the latent eigenmodel is able to represent subnetworks that resemble bipartite graphs.

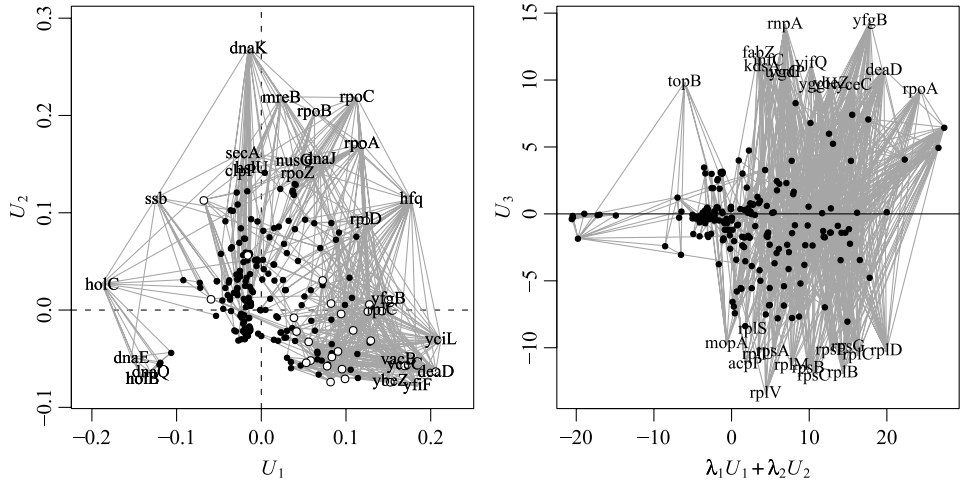

Figure 3. Plots of the latent eigenvectors. In the first panel, nodes with large values of $|u_{i,3}|$ are plotted with white circles.

For a detailed biological interpretation of the different hubs and clusters of the network, the reader is referred to Butland et al. (2005).

5. DISCUSSION

Distributions over the Stiefel manifold play an important role in directional data analysis, multivariate analysis, and random matrix theory. An important family of probability distributions over this manifold is the matrix Bingham–von Mises–Fisher family, which generalizes the vector and matrix von Mises–Fisher and Bingham distributions. This article has developed a rejection sampling scheme for the matrix MF distribution and a Gibbs sampling scheme for the matrix BMF distribution, thereby providing a useful tool for studying these complicated multivariate probability distributions.

Additionally, it has been shown that members of the BMF family of distributions arise as conditional posterior distributions in Gaussian and probit models for multivariate data. Likelihood-based approaches to multivariate modeling may be necessary when the data are ordinal, missing, or otherwise nonstandard, and being able to sample from the BMF family allows for parameter estimation in these situations.

SUPPLEMENTAL MATERIALS

Supplemental materials for this article are available at the *JCGS* website and include the following:

Replication files: R-code and datasets used to generate all of the numeric results in this article, including the figures and tables. (replication-files.zip, zip archive)

Supplementary notes: Further notes and R-code for the rejection sampler for square Bingham matrices, as described in Section 3.4. (squarebingsampler.pdf, pdf file)

Files are also available at my website: *http://www.stat.washington.edu/~hoff*.

ACKNOWLEDGMENTS

The author thanks the editor, associate editor, and two referees for their suggestions on improving the readability and consistency of this article. This work was partially funded by NSF Grant SES-0631531.

[Received December 2007. Revised February 2009.]

REFERENCES

Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J., and Emili, A. (2005), "Interaction Network Containing Conserved and Essential Protein Complexes in *Escherichia coli*," *Nature*, 433, 531–537.

Chikuse, Y. (2003), *Statistics on Special Manifolds*. *Lecture Notes in Statistics*, Vol. 174, New York: Springer-Verlag, ISBN 0-387-00160-3.

Fisher, R. (1953), "Dispersion on a Sphere," *Proceedings of the Royal Society*, Ser. A, 217, 295–305, ISSN 0962-8444.

Gupta, A. K., and Nagar, D. K. (2000), *Matrix Variate Distributions*. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*, Vol. 104, Boca Raton, FL: Chapman & Hall/CRC, ISBN 1-58488-046-5.

Herz, C. S. (1955), "Bessel Functions of Matrix Argument," *Annals of Mathematics (2)*, 61, 474–523, ISSN 0003-486X.

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100 (469), 286–295, ISSN 0162-1459.

Hoff, P. D., and Ward, M. D. (2004), "Modeling Dependencies in International Relations Networks," *Political Analysis*, 12 (2), 160–175.

Kent, J. T., Constable, P. D. L., and Er, F. (2004), "Simulation for the Complex Bingham Distribution," *Statistics and Computing*, 14 (1), 53–57, ISSN 0960-3174.

Khatri, C. G., and Mardia, K. V. (1977), "The von Mises–Fisher Matrix Distribution in Orientation Statistics," *Journal of the Royal Statistical Society*, Ser. B, 39 (1), 95–106, ISSN 0035-9246.

Kume, A., and Walker, S. G. (2006), "Sampling From Compositional and Directional Distributions," *Statistics and Computing*, 16 (3), 261–265, ISSN 0960-3174.

Takemura, A., and Sheena, Y. (2005), "Distribution of Eigenvalues and Eigenvectors of Wishart Matrix When the Population Eigenvalues Are Infinitely Dispersed and Its Application to Minimax Estimation of Covariance Matrix," *Journal of Multivariate Analysis*, 94 (2), 271–299, ISSN 0047-259X.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," (with discussion), *The Annals of Statistics*, 22 (4), 1701–1762, ISSN 0090-5364.

Watson, G. S., and Williams, E. J. (1956), "On the Construction of Significance Tests on the Circle and the Sphere," *Biometrika*, 43, 344–352, ISSN 0006-3444.

Wood, A. T. A. (1987), "The Simulation of Spherical Distributions in the Fisher–Bingham Family," *Communications in Statistics: Simulation and Computation*, 16 (3), 885–898, ISSN 0361-0918.

——— (1994), "Simulation of the von Mises–Fisher Distribution," *Communications in Statistics: Simulation and Computation*, 23, 157–164.