Bayesian analysis of matrix data with **rstiefel**

Peter D. Hoff *

December 5, 2016

Abstract

We illustrate the use of the R-package **rstiefel** for matrix-variate data analysis in the context of two examples. The first example considers estimation of a reduced-rank mean matrix in the presence of normally distributed noise. The second example considers the modeling of a social network of friendships among teenagers. Bayesian estimation for these models requires the ability to simulate from the matrix-variate von Mises-Fisher distributions and the matrix-variate Bingham distributions on the Stiefel manifold.

1 Exponential families on the Stiefel manifold

The set of $m \times R$ matrices $\mathbf{U}$ for which $\mathbf{U}^T\mathbf{U} = \mathbf{I}_R$ is called the $m \times R$ Stiefel manifold and is denoted $\mathcal{V}_{R,m}$. The densities of a quadratic exponential family on this manifold (with respect to the uniform measure) are given by

$$p(\mathbf{U}|\mathbf{A},\mathbf{B},\mathbf{C}) \propto \text{etr}(\mathbf{C}^T\mathbf{U} + \mathbf{B}\mathbf{U}^T\mathbf{A}\mathbf{U}), \tag{1}$$

*Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98195-4322. This research was supported in part by NI-CHD grant 1R01HD067509-01A1.

where $\mathbf{C} \in \mathbb{R}^{m \times R}$, $\mathbf{B}$ is an $R \times R$ diagonal matrix and $\mathbf{A}$ is a symmetric matrix. Since $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, the density is unchanged under transformations of the form $\mathbf{A} \to \mathbf{A} + a\mathbf{I}$ or $\mathbf{B} \to \mathbf{B} + b\mathbf{I}$. Additionally, it is convenient to restrict the diagonal entries of $\mathbf{B}$ to be in decreasing order. If $\mathbf{B}$ is not ordered in this way, there exists a reparameterization $(\mathbf{A}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}})$ giving the same distribution as $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ but where $\tilde{\mathbf{B}}$ has ordered diagonal entries. More details on the Stiefel manifold and these distributions can be found in Chikuse (2003), Hoff (2009a), Hoff (2009b) and the references therein.

Distributions of this form were originally studied in the case $R = 1$, so that the manifold was just the surface of the m-sphere. In this case, $\mathbf{B}$ reduces to a scalar and can be absorbed into the matrix $\mathbf{A}$. The quadratic exponential family then has densities of the form

$$p(\mathbf{u}|\mathbf{c}, \mathbf{A}) \propto \exp(\mathbf{c}^T\mathbf{u} + \mathbf{u}^T\mathbf{A}\mathbf{u}). \tag{2}$$

The case that $\mathbf{A} = \mathbf{0}$ was studied by von Mises, Fisher and Langevin, and so a distribution with density proportional to $\exp(\mathbf{c}^T\mathbf{u})$ is often called a von Mises-Fisher or Langevin distribution on the sphere. The case that $\mathbf{c} = \mathbf{0}$ and $\mathbf{A} \neq 0$ was studied by Bingham (1974), and is called the Bingham distribution. This distribution has "antipodal symmetry" in that $p(\mathbf{u}|\mathbf{A}) = p(-\mathbf{u}|\mathbf{A})$, and so may be appropriate as a model for random axes, rather than random directions.

In recognition of the work of the above mentioned authors, we refer to distributions with densities given by (2) and (1) as vector-variate and matrix-variate Bingham-von Mises-Fisher distributions, respectively. This is a rather long name, however, so in this vignette I will refer to them as BMF distributions. The case that $\mathbf{A}$ (or $\mathbf{B}$) is the zero matrix will be referred to as an MF distribution, and the case that $\mathbf{C}$ is zero will be referred to as a

Bingham distribution. More descriptive names might be L, Q and LQ to replace the names MF, Bingham, and BMF, respectively, the idea being that the "L" and "Q" refer to the presence of linear and quadratic components of the density.

2 Model-based SVD

It is often useful to model an $m \times n$ rectangular matrix-variate dataset $\mathbf{Y}$ as being equal to some reduced rank matrix $\mathbf{M}$ plus i.i.d. noise, so that $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, with the elements $\{\epsilon_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n\}$ of $\mathbf{E}$ assumed to be i.i.d. with zero mean and some unknown variance σ^2. The singular value decomposition states that any rank-R matrix $\mathbf{M}$ can be expressed as $\mathbf{M} = \mathbf{UDV}^T$, where $\mathbf{U} \in \mathcal{V}_{R,m}$, $\mathbf{V} \in \mathcal{V}_{R,n}$ and $\mathbf{D}$ is an $R \times R$ diagonal matrix. If we are willing to assume normality of the errors, the model can then be written as

$$\begin{aligned} \mathbf{Y} &= \mathbf{UDV}^T + \mathbf{E} \\ \mathbf{E} &= \{\epsilon_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n\} \sim \text{ i.i.d. normal}(0, \sigma^2). \end{aligned}$$

Bayesian rank selection for this model was considered in Hoff (2007). In this vignette we consider estimation for a specified rank R, in which case the unknown parameters in the model are $\{\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2\}$. Given a suitable prior distribution over these parameters, Bayesian inference can proceed via construction of a Markov chain with stationary distribution equal to the conditional distribution of the parameters given $\mathbf{Y}$, i.e. the distribution with density $p(\mathbf{U}, \mathbf{D}, \mathbf{V}, \sigma^2|\mathbf{Y})$. In particular, conjugate prior distributions allow the construction of a Markov chain via the Gibbs sampler, which iteratively simulates each parameter from its full conditional distribution. If the prior

distribution for $\mathbf{U}$ is uniform on $\mathcal{V}_{R,m}$, then its full conditional density is given by

$$\begin{aligned} p(\mathbf{U}|\mathbf{Y},\mathbf{D},\mathbf{V},\sigma^2) &\propto p(\mathbf{Y}|\mathbf{U},\mathbf{D},\mathbf{V},\sigma^2) \\ &\propto \text{etr}(-[\mathbf{Y}-\mathbf{U}\mathbf{D}\mathbf{V}^T]^T[\mathbf{Y}-\mathbf{U}\mathbf{D}\mathbf{V}^T]/(2\sigma^2)) \\ &\propto \text{etr}([\mathbf{Y}\mathbf{V}\mathbf{D}/\sigma^2]^T\mathbf{U}), \end{aligned}$$

which is the density of an MF($\mathbf{Y}\mathbf{V}\mathbf{D}/\sigma^2$) distribution. Similarly, the full conditional distribution of $\mathbf{V}$ under a uniform prior is MF($\mathbf{Y}^T\mathbf{U}\mathbf{D}/\sigma^2$). For this vignette, we will use the following prior distributions for $\{d_1,\ldots,d_R,\sigma^2\}$:

$$\begin{aligned} \{d_1,\ldots,d_R|\tau^2\} &\sim \text{ i.i.d. normal}(0,\tau^2) \\ 1/\tau^2 &\sim \text{gamma}(\eta_0/2,\eta_0\tau_0^2/2) \\ 1/\sigma^2 &\sim \text{gamma}(\nu_0/2,\nu_0\sigma_0^2/2.) \end{aligned}$$

The corresponding full conditional distributions are

$$\begin{aligned} \{d_j|\mathbf{U},\mathbf{V},\mathbf{Y},\mathbf{d}_{-j},\sigma^2,\tau^2\} &\sim \text{normal}(\tau^2\mathbf{u}_j^T\mathbf{Y}\mathbf{v}_j/[\sigma^2+\tau^2],\tau^2\sigma^2/[\tau^2+\sigma^2]) \\ \{1/\tau^2|\mathbf{U},\mathbf{D},\mathbf{V},\mathbf{Y},\sigma^2\} &\sim \text{gamma}([\eta_0+R]/2,[\eta_0\tau_0^2+\textstyle\sum d_j^2]/2) \\ \{1/\sigma^2|\mathbf{U},\mathbf{D},\mathbf{V},\mathbf{Y},\tau^2\} &\sim \text{gamma}([\nu_0+mn]/2,[\nu_0\sigma_0^2+||\mathbf{Y}-\mathbf{U}\mathbf{D}\mathbf{V}^T||^2]/2). \end{aligned}$$

2.1 Simulated data

We now randomly generate some parameters and data according to the model above:

```
> library(rstiefel)
> set.seed(1)
> m<-60 ; n<-40 ; R0<-4
> U0<-rustiefel(m,R0)
```

```
> V0<-rustiefel(n,R0)
> D0<-diag(sort(rexp(R0),decreasing=TRUE))*sqrt(m*n)
> M0<-U0%*%D0%*%t(V0)
> Y<-M0 + matrix(rnorm(n*m),m,n)
```

The only command from the rstiefel package used here is `rustiefel`, which generates a uniformly distributed random orthonormal matrix. Note that `rustiefel(m,R)` gives a matrix with m rows and R columns, and so the arguments are in the reverse of their order in the symbolic representation of the manifold $\mathcal{V}_{R,m}$.

2.2 Gibbs sampler

Now we try to recover the true values of the parameters $\{\mathbf{U}_0, \mathbf{V}_0, \mathbf{D}_0, \sigma^2\}$ from the observed data $\mathbf{Y}$. Just for fun, let's estimate these parameters with a presumed rank $R > R_0$ that is larger than the actual rank. Equivalently, we can think of $\mathbf{U}_0, \mathbf{V}_0, \mathbf{D}_0$ as having dimension $m \times R$, $n \times R$ and $R \times R$, but with the last $R - R_0$ diagonal entries of $\mathbf{D}_0$ being zero.

The prior distributions for $\mathbf{U}$ and $\mathbf{V}$ are uniform on their respective manifolds. We set our hyperparameters for the other priors as follows:

```
> nu0<-1 ; s20<-1      #inverse-gamma prior for the error variance s2
> eta0<-1 ; t20<-1     #inverse-gamma prior for the variance t2 of the sing vals
```

Construction of a Gibbs sampler requires starting values for all (but one) of the unknown parameters. An natural choice is the MLE:

```
> R<-6
> tmp<-svd(Y) ; U<-tmp$u[,1:R] ; V<-tmp$v[,1:R] ; D<-diag(tmp$d[1:R])
> s2<-var(c(Y-U%*%D%*%t(V)))
> t2<-mean(diag(D^2))
```

Let's compare the MLE of $\mathbf{D}$ to the true value:

```
> d.mle<-diag(D)
> d.mle

[1] 40.05172 25.00226 19.70827 13.43382 13.10381 12.64942

> diag(D0)

[1] 38.514216 24.015791 17.352783  1.169442
```

The values of the MLE are, as expected, larger than the true values, especially for the smaller values of $\mathbf{D}_0$. Now let's see if the Bayes estimate provides some shrinkage.

```
> MPS<-matrix(0,m,n) ; DPS<-NULL
> for(s in 1:2500)
+ {
+   U<-rmf.matrix(Y%*%V%*%D/s2)
+   V<-rmf.matrix(t(Y)%*%U%*%D/s2)
+
+   vd<-1/(1/s2+1/t2)
+   ed<-vd*(diag(t(U)%*%Y%*%V)/s2 )
+   D<-diag(rnorm(R,ed,sqrt(vd)))
+
+   s2<-1/rgamma(1, (nu0+m*n)/2 , (nu0*s20 + sum((Y-U%*%D%*%t(V))^2))/2 )
+   t2<-1/rgamma(1, (eta0+R)/2, (eta0*t20 + sum(D^2))/2)
+
+   ### save output
+   if(s%%5==0)
```

```
+   {
+     DPS<-rbind(DPS,sort(diag(abs(D)),decreasing=TRUE))
+     M<-U%*%D%*%t(V)
+     MPS<-MPS+M
+   }
+ }
```

This generates a Gibbs sampler of 2500 iterations. Here, we save the values of $\mathbf{D}$ every 5th iteration, resulting in a sample of $\mathbf{D}$-values of size 500 with which to estimate $p(\mathbf{D}|\mathbf{Y})$. Additionally, we can obtain a posterior mean estimate of $\mathbf{M}_0 = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^T$ via the sample average of $\mathbf{UDV}^T$. Note that this estimate is not of rank R, as the set matrices of less than full rank is not convex. If we want a rank R estimate, we could take the rank-R approximation of the posterior mean.

Let's look at the squared error for the MLE, the posterior expectation of $\mathbf{M}_0$, and the rank-R approximation to the posterior expectation:

```
> tmp<-svd(Y) ; M.ml<-tmp$u[,1:R]%*%diag(tmp$d[1:R])%*%t(tmp$v[,1:R])
> M.b1<-MPS/dim(DPS)[1]
> tmp<-svd(M.b1) ; M.b2<-tmp$u[,1:R]%*%diag(tmp$d[1:R])%*%t(tmp$v[,1:R])
> mean( (M0-M.ml)^2 )

[1] 0.3563462

> mean( (M0-M.b1)^2 )

[1] 0.1306738

> mean( (M0-M.b2)^2 )
```

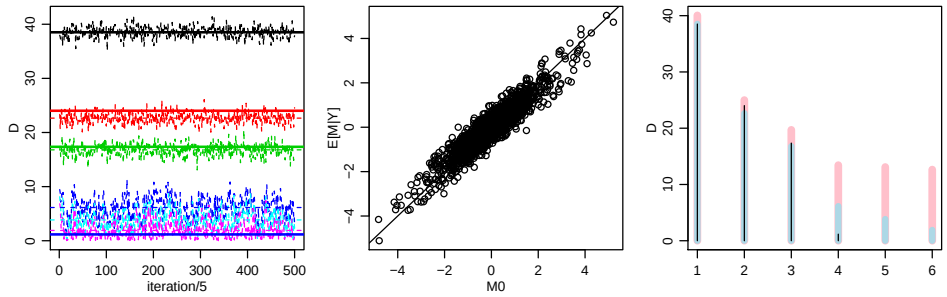

Figure 1: Some output of the Gibbs sampler.

```
[1] 0.1304145
```

Not surprisingly, the MLE has a much larger loss than the Bayes estimates. The squared error for the two Bayes estimates are nearly identical. This is because although the posterior mean has full rank $m \wedge n$, it is very close to its rank-R approximation.

Finally, let's make some plots based on the output of the Gibbs sampler. The left-most plot of Figure 1 gives simulated values of $\mathbf{D}$, with the values of $\mathbf{D}_0$ given in thick lines. The mixing of the Markov chain looks pretty reasonable. The center plot gives $\mathbf{M}_0$ versus its posterior expectation, approximated from the MCMC sample average of $\mathbf{UDV}^T$. The right plot gives the MLEs of $\mathbf{D}_0$ in pink, the posterior expectations of $\mathbf{D}_0$ in light blue, and the true values in thin black lines. The posterior estimates are very accurate for the large singular values of $\mathbf{D}_0$, but are overestimates for the smallest values (the last $R - R_0$ of which are zero). However, these Bayes estimates are much better than the unregularized MLEs.

3 Network analysis

In this section we analyze a dataset on the social network and some health behaviors of a group of $n = 50$ Scottish teenage girls. These data were derived from the data available at `http://www.stats.ox.ac.uk/~snijders/siena/s50_data.htm` and described in Michell and Amos (1997).

3.1 An eigenmodel for symmetric networks

Let $\mathbf{Y}$ be the $n \times n$ symmetric adjacency matrix corresponding to this network, with off-diagonal entry $y_{i,j}$ equal to the binary indicator of a friendship between actors i and j, as reported by one or both actors. In this vignette we will derive a model-based representation of these data using the following reduced-rank probit model:

$$\begin{aligned} z_{i,j} &= \theta + \mathbf{u}_i^T \Lambda \mathbf{u}_j + \epsilon_{i,j} \\ y_{i,j} &= 1_{(0,\infty)}(z_{i,j}), \end{aligned} \tag{3}$$

where $\{\epsilon_{i,j} = \epsilon_{j,i}\} \sim$ i.i.d. normal$(0, 1)$, $\Lambda = \text{diag}(\lambda_1, \lambda_2)$ and the matrix $\mathbf{U}$ with row vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ lies in the Stiefel manifold $\mathcal{V}_{R,n}$. This model is a type of two-way latent factor model in which the relationship between actors i and j is modeled in terms of their unobserved latent factors $\mathbf{u}_i$ and $\mathbf{u}_j$. This model and its relationship to other latent variable network models are described more fully in Hoff (2008).

Convenient prior distributions for $\{\mathbf{U}, \Lambda, \theta\}$ are as follows:

$$\begin{aligned} \theta &\sim \text{normal}(0, \tau_\theta^2) \\ (\lambda_1, \lambda_2) &\sim \text{ i.i.d. normal}(0, \tau_\lambda^2) \\ \mathbf{U} &\sim \text{uniform}(\mathcal{V}_{R,n}) \end{aligned}$$

Conditional on the observed network $\mathbf{Y}$, posterior inference can proceed via a Gibbs sampling scheme for the unknown quantities $\{\mathbf{Z}, \mathbf{U}, \Lambda, \theta\}$. Under model (3), observing $y_{i,j} = 0$ or 1 implies that $z_{i,j}$ is less than or greater than zero, respectively. Thus conditional on $\{\mathbf{Y}, \mathbf{U}, \Lambda, \theta\}$, the distribution of $\mathbf{Z}$ is that of a random symmetric normal matrix with mean $\theta + \mathbf{U}\Lambda\mathbf{U}^T$ and independent entries that are constrained to be positive or negative depending on the entries of $\mathbf{Y}$. Given $\mathbf{Z}$, the full conditional distributions of $\{\mathbf{U}, \Lambda, \theta\}$ do not depend on $\mathbf{Y}$, and can be obtained from the corresponding prior distributions and the density for the matrix $\mathbf{Z}$, given by

$$\begin{aligned} p(\mathbf{Z}|\mathbf{U}, \Lambda) &\propto \text{etr}(-[\mathbf{Z} - \theta\mathbf{1}\mathbf{1}^T - \mathbf{U}\Lambda\mathbf{U}^T]^T[\mathbf{Z} - \theta\mathbf{1}\mathbf{1}^T - \mathbf{U}\Lambda\mathbf{U}^T]/4) \\ &= \text{etr}(-\mathbf{E}^T\mathbf{E}/4) \times \text{etr}(\Lambda\mathbf{U}^T\mathbf{E}\mathbf{U}/2) \times \text{etr}(-\Lambda^2/4), \qquad (4) \end{aligned}$$

where $\mathbf{E} = \mathbf{Z} - \theta\mathbf{1}\mathbf{1}^T$ has mean $\mathbf{U}\Lambda\mathbf{U}^T$ and off-diagonal variances of 1. The diagonal elements of $\mathbf{E}$ (and $\mathbf{Z}$) have variance 2, but do not correspond to any observed data as the diagonal of $\mathbf{Y}$ is undefined. These diagonal elements are integrated over in the Markov chain Monte Carlo estimation scheme described below. From (4), the full conditional distribution of $\mathbf{U}$ is easily seen to be a Bingham$(\mathbf{E}/2, \Lambda)$ distribution. Full conditional distributions for the other quantities are available via standard calculations, and are given in Hoff (2009a) and in the code below.

3.2 Gibbs sampler

The data for this example are stored as a list:

```
> YX_scots<-dget("YX_scots") ; Y<-YX_scots$Y ; X<-YX_scots$X
```

The $n \times 2$ matrix $\mathbf{X}$ provides a binary indicator of drug use and smoking behavior for each actor during the period of the study. Understanding the

relationship between these health behaviors and the social network can be facilitated by examining the relationship between $\mathbf{X}$ and the latent factors $\mathbf{U}$ that represent the network via the model given in (3).

We specify the dimension of the latent factors and the values of the hyperparameters as follows:

```
> ## priors
> R<-2 ; t2.lambda<-dim(Y)[1] ; t2.theta<-100
```

A value of $\tau_\lambda^2 = n$ allows the prior magnitude of the latent factor effects to increase with n, but not as fast as the residual variance: Letting $\mathbf{U}_1$ be the first column of $\mathbf{U}$, we have $\mathrm{E}[||\lambda_1 \mathbf{U}_1 \mathbf{U}_1^T||^2] = \mathrm{E}[\lambda_1^2] = n$. On the other hand, letting $\mathcal{E}$ be the matrix of residuals $\{\epsilon_{i,j}\}$, we have $\mathrm{E}[||\mathcal{E}||^2] = (n+1)n$.

For brevity, we consider simple, naive starting values for the unknown parameters:

```
> ## starting values
> theta<-qnorm(mean(c(Y),na.rm=TRUE))
> L<-diag(0,R)
> set.seed(1)
> U<-rustiefel(dim(Y)[1],R)
```

Better starting values could be obtained from a few iterations of an EM or block coordinate descent algorithm, although these naive starting values are adequate for this example.

We are now ready to run the Gibbs sampler. We will store simulated values of Λ and θ in the objects `LPS` and `TPS`, respectively. Instead of saving values of $\mathbf{U}$, we will just compute the sum of $\mathbf{U}\Lambda\mathbf{U}^T$ across iterations of the Markov chain. Dividing by the number of iterations, this sum provides an

approximation to the posterior mean of $\mathbf{U}\Lambda\mathbf{U}^T$. A rank-$R$ eigendecomposition of the posterior mean can be used to provide an estimate of $\mathbf{U}$.

```
> ## MCMC
> LPS<-TPS<-NULL ; MPS<-matrix(0,dim(Y),dim(Y))
> for(s in 1:10000)
+ {
+
+   Z<-rZ_fc(Y,theta+U%*%L%*%t(U))
+
+   E<-Z-U%*%L%*%t(U)
+   v.theta<-1/(1/t2.theta + choose(dim(Y)[1],2))
+   e.theta<-v.theta*sum(E[upper.tri(E)])
+   theta<-rnorm(1,e.theta,sqrt(v.theta))
+
+   E<-Z-theta
+   v.lambda<-2*t2.lambda/(2+t2.lambda)
+   e.lambda<-v.lambda*diag(t(U)%*%E%*%U/2)
+   L<-diag(rnorm(R,e.lambda,sqrt(v.lambda)))
+
+   U<-rbing.matrix.gibbs(E/2,L,U)
+
+   ## output
+   if(s>100 & s%%10==0)
+   {
+     LPS<-rbind(LPS,sort(diag(L))) ; TPS<-c(TPS,theta) ; MPS<-MPS+U%*%L%*%t(U)
+   }
```

```
+ }
```

Note that this code uses a function `rZ_fc`, which simulates from the full conditional distribution of $\mathbf{Z}$ given $\{\mathbf{Y}, \mathbf{U}, \Lambda, \theta\}$, which is that of independent constrained normal random variables. The code for this function can be obtained from the LaTeX source file for this document.

A summary of the posterior distribution is provided in Figure 2. The first panel plots the posterior density of θ, and the second plots the (marginal) posterior densities of the ordered values of (λ_1, λ_2). This plot strongly suggests that the values of λ_1 and λ_2 are both positive. Since the probability of a friendship between i and j is increasing in $\mathbf{u}_i^T \Lambda \mathbf{u}_j$, the results posit that friendships are more likely between individuals with similar values for their latent factors (this effect is sometimes referred to as homophily). The third panel plots the observed network with the node positions obtained from the estimates of $\mathbf{u}_1, \ldots, \mathbf{u}_n$ based on the rank-2 approximation of the posterior mean of $\mathbf{U}\Lambda\mathbf{U}^T$. The plotting colors and characters for the nodes are determined by the drug and smoking behaviors: Non-smokers are plotted in green and smokers in red, non-drug users are plotted as circles and drug users as triangles. The plot indicates a separation between students with no drug or tobacco use (green circles) from the other students in terms of their latent factors, suggesting a relationship between these health behaviors and the social network.

References

Christopher Bingham. An antipodally symmetric distribution on the sphere. *Ann. Statist.*, 2:1201–1225, 1974. ISSN 0090-5364.

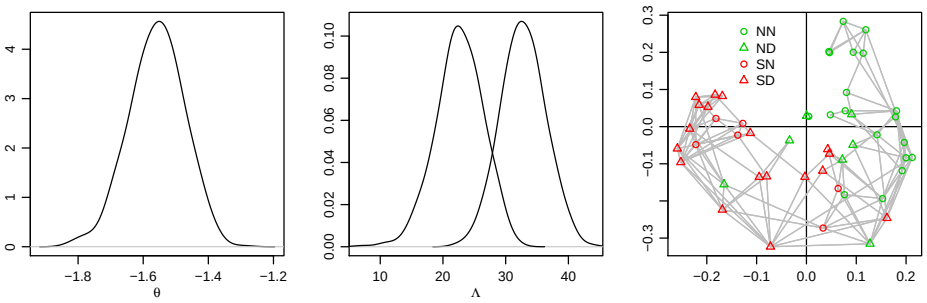

Figure 2: Some output of the Gibbs sampler.

Yasuko Chikuse. *Statistics on special manifolds*, volume 174 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2003. ISBN 0-387-00160-3.

Peter D. Hoff. Model averaging and dimension selection for the singular value decomposition. *J. Amer. Statist. Assoc.*, 102(478):674–685, 2007. ISSN 0162-1459.

Peter D. Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 657–664. MIT Press, Cambridge, MA, 2008. URL `http://cran.r-project.org/web/packages/eigenmodel/`.

Peter D. Hoff. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009a.

Peter D. Hoff. A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):971–992, 2009b.

L. Michell and A. Amos. Girls, pecking order and smoking. *Social Science & Medicine*, 44(12):1861–1869, 1997.