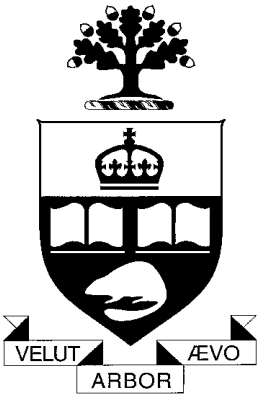

Bayesian Factor Analysis via Concentration

by

Y. Cao
Department of Statistics
University of Toronto

and

M. Evans
Department of Statistics
University of Toronto

and

I. Guttman
Department of Mathematics and Statistics
SUNY at Buffalo

Technical Report No. 1003 April 12, 2010

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Bayesian Factor Analysis via Concentration

Cao[*], Y., Evans[*], M. and Guttman[†], I.

Abstract: We consider factor analysis when we assume the distribution form is known up to its mean and variance. A prior is placed on the mean and variance and then inference is made as to whether or not any latent factors exist. Inference is carried out by comparing the concentrations of the prior and posterior about various subsets of the parameter space that are specified by hypothesizing factor structures. An importance sampling algorithm is developed to handle the case where the prior on the correlation matrix is uniform, independent of the prior on the location and scale parameters.

Key words and phrases: factor analysis, Bayesian inference, concentration, importance sampling.

1 Introduction

Suppose that $\mathbf{y} \in R^p$ has unknown mean $\mu \in R^p$ and variance $\mathbf{\Sigma} \in R^{p\times p}$. A factor model corresponds to saying that $\mathbf{\Sigma}$ possesses a particular structure, namely,

$$\mathbf{\Sigma} = \mathbf{\Gamma}_q \mathbf{\Gamma}_q' + \mathbf{\Psi} \tag{1}$$

[*]Department of Statistics, University of Toronto
[†]Department of Mathematics and Statistics, SUNY at Buffalo

where $0 \leq q \leq p$, $\mathbf{\Gamma}_q \in R^{p \times q}$ of rank q and $\mathbf{\Psi}$ is diagonal with nonnegative entries. The structure (1) arises from the existence of latent factors $\mathbf{f} \in R^q$, having distribution with mean $\mathbf{0}$ and variance $\mathbf{I}$, and unique variables $\mathbf{e} \in R^p$, uncorrelated with $\mathbf{f}$ and having mean $\mathbf{0}$ and variance $\mathbf{\Psi}$, such that $\mathbf{y} = \mu + \mathbf{\Gamma}_q \mathbf{f} + \mathbf{e}$. Note that when $q = p$, then we can take $\mathbf{\Gamma}_q = \mathbf{\Sigma}^{1/2}$ (the symmetric square root of $\mathbf{\Sigma}$) and $\mathbf{\Psi} = 0$ and so (1) is always correct for some q. When $q = 0$, then $\mathbf{\Gamma}_q = 0$ and the response variables are independent. The point of a factor analysis is to identify the smallest q such that (1) holds and and further choose $\mathbf{\Gamma}_q$ to help in providing interpretations for the latent factors.

For any orthogonal matrix $\mathbf{Q} \in R^{q \times q}$, we have that when (1) holds, then $\mathbf{\Sigma} = \mathbf{\Gamma}_q \mathbf{\Gamma}_q' + \mathbf{\Psi} = (\mathbf{\Gamma}_q \mathbf{Q})(\mathbf{\Gamma}_q \mathbf{Q})' + \mathbf{\Psi}$ and so $\mathbf{\Gamma}_q$ is not unique. We partially avoid the nonuniqueness by noting that $\mathbf{\Gamma}_q$ can be written uniquely as $\mathbf{\Gamma}_q = \mathbf{T}\mathbf{Q}$ where $\mathbf{Q} \in R^{q \times q}$ is orthogonal and $\mathbf{T} \in R^{p \times q}$ is lower triangular with nonnegative diagonal elements. We require $\mathbf{\Gamma}_q$ to be lower triangular with nonnegative diagonal elements hereafter, effectively taking $\mathbf{Q} = \mathbf{I}$, although we will weaken this restriction slightly for the computations. After $\mathbf{\Gamma}_q$ is estimated, rotations can be applied as is usual in factor analysis.

Likelihood methods are often used for factor analysis but these are well-known to suffer from various computational challenges. In fact, without sometimes fairly arbitrary restrictions being placed on the estimates, it can happen that software does not produce sensible answers (see, for example, Cao (2010)). So we consider the use of Bayesian methodology for this problem.

One Bayesian approach to factor analysis proceeds as follows. Let π_q^* be the prior probability that the model with q common factors is correct for $q = 0, \ldots, p$, and let π_q be a prior on (μ, Ψ, Γ_q). After observing a sample $\mathbf{Y} = (\mathbf{y}_1 \cdots \mathbf{y}_n)$, we have the posterior distributions $\pi_q(\cdot \mid \mathbf{Y})$ for inferences about $(\mu, \mathbf{\Psi}, \mathbf{\Gamma}_q)$, and the posterior probability $\pi_q^*(\mathbf{Y})$ that the model with q factors is correct. One common method of selecting q is based on the $\pi_q^*(\mathbf{Y})$, e.g., choose the submodel with largest posterior probability. Alternatively we could choose the

model with the largest Bayes factor. When the prior model probabilities are uniform, then these two criteria agree.

There are several difficulties associated with this approach. First, we need to specify $p+1$ prior distributions and prior model probabilities. This requires that, for the model with q factors, we have information about the relevant $\mathbf{\Gamma}_q$, as specified in a $q[p-(q-1)/2]$ dimensional distribution, and we need this for $q = 0, \ldots, p$. This involves a demanding amount of elicitation. To avoid this, we might place default priors on these quantities, but such priors are typically improper. This leads to ambiguities concerning the proper interpretation of posterior model probabilities and Bayes factors due to the dependence of these quantities on arbitrary constants multiplying improper priors. Further, these factor models are nested and, as such, considerations such as those discussed in Ghosh, Delampady and Samanta (2006), Section 6.7, are relevant.

While various solutions to these problems have been proposed, we take a very simple approach here that avoids these difficulties. We place a single prior on $(\mu, \mathbf{\Sigma})$ and assess a submodel by comparing the concentration of the prior around the subset of the parameter space specified by the submodel, with the concentration of the posterior about this subset. Intuitively, if the posterior concentrates much more about this subset than the prior, then we have evidence via the data, of the plausibility of the submodel. We discuss how to measure concentration and how to compare the prior and posterior concentrations in Section 2. In Section 3 we discuss a measure of concentration in factor analysis models. In Section 4 we develop a computational approach using a prior for which elicitation is relatively straightforward, namely, we place a uniform prior on the correlation matrix. In Section 5 we consider inferences and some applications involving simulated and real data.

Lee (2007) is an excellent, up-to-date discussion of current developments in Bayesian factor analysis. Also, relevant material can be found in Bartholomew (1987), Press and Shigemasu (1989), Lee and Press (1998), and Lopes and West (2004).

2 Concentration and Hypothesis Assessment

Suppose that we have a probability measure P on a set $\mathcal{X}$ and we want to measure to what extent P concentrates about $C \subset \mathcal{X}$. This concentration cannot be measured simply by $P(C)$ as, for two sets with the same probability content the remaining probability may be much more widely dispersed for one than the other. For the applications we consider, the probability content of subsets C of interest are equal to 0 and so it is how the probability is distributed on C^c that is relevant. In such a case the most obvious measure of concentration is $E_P(d(x, C))$ where $d(x, C) = \inf\{d(x, y) : y \in C\}$ for some distance measure $d(x, y)$ on $\mathcal{X} \times \mathcal{X}$. This can be seen as a generalization of the concept of variance by taking $C = \{E_P(X)\}$ and $d(x, y) = \|x - y\|^2$, where $\|\cdot\|$ is the Euclidean norm. Clearly the more the distribution of $d(x, C)$ concentrates near 0 the more concentrated P is about C.

Now suppose that we have a statistical model $\{P_\theta : \theta \in \Theta\}$, a prior Π on Θ and we want to assess the hypothesis that the true value of $\theta \in H_0 \subset \Theta$ after observing the data x. Perhaps the most natural method of assessing this hypothesis is to compute the posterior probability $\Pi(H_0 \,|\, x)$ and regard this as evidence against H_0 when it is small. A difficulty with this approach is that, when $\Pi(H_0)$ is small, then a large amount of data may be needed to make $\Pi(H_0 \,|\, x)$ large enough to be convincing. In fact, if Π assigns 0 probability to H_0, simply because it is a lower dimensional set of Θ, then $\Pi(H_0 \,|\, x) = 0$ no matter what data is obtained.

If we are making inference about a marginal parameter Ψ defined on Θ, then $H_0 = \Psi^{-1}\{\psi_0\}$ for some specified value ψ_0, and various approaches can be taken to deal with the problem caused by $\Pi(H_0) = 0$. When no Ψ naturally exists, as in the factor analysis problem we are discussing, then we have to choose one and taking $\Psi(\theta) = d(\theta, H_0)$ seems like a reasonable choice. Clearly, the degree to which Π (or $\Pi(\cdot \,|\, x)$) concentrates about H_0, is a measure of our belief in H_0 *a priori* (or *a posteriori*) and the prior (or posterior)

concentration of the distribution of $d(\theta, H_0)$ about 0 is measuring this.

Accordingly, we look at comparing the concentrations of the prior and the posterior probability measures about H_0. If the null hypothesis is true, i.e., the true value $\theta^* \in H_0$, then we expect that the data will lead to a greater concentration of the posterior distribution about H_0 than the prior distribution. In fact, under weak conditions, the posterior distribution will concentrate on θ^* as we increase the amount of data, and so the posterior distribution of $d(\theta, H_0)$ will converge to $d(\theta^*, H_0) = 0$ when $\theta^* \in H_0$, and converge to a nonzero value otherwise. So no matter how we choose d, for large amounts of data we can expect the posterior distribution of d to indicate whether or not H_0 holds.

To calibrate to what degree the posterior distribution of $d(\theta, H_0)$ is concentrating about 0, we compare its posterior distribution to its prior distribution. So, it is necessary to decide how we will compare the prior and posterior distributions of d. For this we look at the ratio of the posterior density to the prior density of d at 0, namely, $\pi_d(0 \mid x)/\pi_d(0)$ which measures the change in belief in H_0 from *a priori* to *a posteriori*. To calibrate this we compute the P-value

$$\Pi\left(\frac{\pi_d(d(\theta, H_0) \mid x)}{\pi_d(d(\theta, H_0))} \leq \frac{\pi_d(0 \mid x)}{\pi_d(0)} \mid x\right) \tag{2}$$

which is the posterior probability of an increase in belief, from *a priori* to *a posteriori,* in the distance of the true value of θ from H_0, that is no greater than that obtained when the hypothesis is true. So when (2) is near 0 the increase in belief is much greater for values of $\theta \notin H_0$ and we interpret this as evidence that H_0 is false. We note that if we were to change d to $e(d)$, where e is any 1-1, smooth transformation, then (2) does not change its value. Further, Evans and Shakhatreh (2008), establish various optimality properties for inferences based on (2) in the class of all Bayesian inferences.

Comparing the concentration of the prior and posterior measures about a hypothesis of interest, as a method for assessing this hypothesis, has been previously discussed in the

literature. For example, Evans, Gilula and Guttman (1993) and Evans, Gilula, Guttman and Swartz (1997) used this idea in the context of contingency tables. In those papers, however, the concentrations were compared simply via graphing the prior and posterior densities of d. In this paper we use a more precise comparison via (2).

3 Concentration for Factor Analysis Models

Basically we are trying to identify the smallest number of factors that effectively explain the correlations among the response variables. To reflect this we adopt a slightly different parameterization. For this let $\mathbf{\Sigma} = \Delta^{1/2}\mathbf{\Xi}\Delta^{1/2}$ where $\mathbf{\Xi}$ is the correlation matrix and $\Delta = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. Then we can write $\Delta^{-1/2}(\mathbf{y} - \mu) = \mathbf{\Gamma}_q\mathbf{f} + \mathbf{e}$ where $\mathbf{e}$ and $\mathbf{f}$ are as before but now $\mathbf{\Gamma}_q$ is lower triangular with nonnegative diagonal elements and $\sum_{k=1}^{\min(i,q)} \gamma_{ik}^2 + \psi_i = 1$ for $i = 1, \ldots, p$. We note that the ψ_i values are now determined by the values in $\mathbf{\Gamma}_q$. This leads to the equation

$$\mathbf{\Xi} = \mathbf{\Gamma}_q\mathbf{\Gamma}_q' + \mathbf{\Psi} \tag{3}$$

and we set H_0^q equal to the set of all $p \times p$ matrices of the form given by (3) where $\mathbf{\Gamma}_q$ is lower triangular with nonnegative diagonal elements and $\sum_{k=1}^{\min(i,q)} \gamma_{ik}^2 + \psi_i = 1$ for $i = 1, \ldots, p$. While our methods will address H_0^q as given by (3) our approach can also be applied to H_0^q as given by (1), but we believe (3) is more relevant for factor analysis.

We will now proceed to discuss how we will measure the concentration of the prior or posterior about H_0^q. For this we need a measure of the distance of an arbitrary correlation matrix $\mathbf{\Xi}$ from H_0^q. While there are many possible choices, it seems natural to base this on the squared Frobenius distance which, for positive semidefinite matrices $\Sigma^{(1)}, \Sigma^{(2)}$, is given by

$$d(\mathbf{\Sigma}^{(1)}, \mathbf{\Sigma}^{(2)}) = \sum_{i=1}^{p} \left(\sigma_{ii}^{(1)} - \sigma_{ii}^{(2)}\right)^2 + 2\sum_{i<j} \left(\sigma_{ij}^{(1)} - \sigma_{ij}^{(2)}\right)^2.$$

This simplifies for correlation matrices to $d(\boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}) = 2\sum_{i<j}(\xi_{ij}^{(1)} - \xi_{ij}^{(2)})^2$. So, for a given correlation matrix $\boldsymbol{\Xi}$ we want to find $d(\boldsymbol{\Xi}, H_0^q)$ and for this we need to be able to solve the following optimization problem. For a given correlation matrix $\boldsymbol{\Xi}$ we need to find the minimum, as a function of Γ_q, of

$$\sum_{i<j}\left(\xi_{ij} - \sum_{k=1}^{\min(i,q)} \gamma_{ik}\gamma_{jk}\right)^2 \tag{4}$$

where $\gamma_{ii} \geq 0$ for $i = 1, \ldots, q$ and $\sum_{k=1}^{\min(i,q)} \gamma_{ik}^2 \leq 1$ for $i = 1, \ldots, p$. Note that (4) is invariant under multiplying any column of Γ_q by -1 so instead we can consider minimizing (4) subject to the simpler constraint that the i-th row of Γ_q lies in the unit ball $B^{\min(i,q)}(\mathbf{0})$ centered at the origin in $R^{\min(i,q)}$. Therefore, to minimize (4) as a function of Γ_q, we have to find the point in the Cartesian product of spheres $B^{\min(1,q)}(\mathbf{0}) \times \cdots \times B^{\min(p,q)}(\mathbf{0})$ where this minimum is attained. Note that (4) is continuous on this compact set and so an absolute minimum is attained.

When $q = 0$ then, since $\Gamma_0 = 0$, the minimized value of (4) is $\sum_{i<j}\xi_{ij}^2$. For other values of q, however, we cannot obtain a closed form expression and need to proceed iteratively. Consider the case when $q = 1$, so we need to minimize $\sum_{i<j}(\xi_{ij} - \gamma_{i1}\gamma_{j1})^2$. First we start with $(\gamma_{11}(0), \ldots, \gamma_{p1}(0)) \in B^1(\mathbf{0}) \times \cdots \times B^1(\mathbf{0})$ which is the p-fold Cartesian product of the interval $[-1, 1]$. We can write

$$\sum_{i<j}(\xi_{ij} - \gamma_{i1}(0)\gamma_{j1}(0))^2 = \left(\sum_{j=2}^{p}\gamma_{j1}^2(0)\right)\gamma_{11}^2(0) - 2\left(\sum_{j=2}^{p}\xi_{1j}\gamma_{j1}(0)\right)\gamma_{11}(0) + c$$

where c is a constant as a function of $\gamma_{11}(0)$. If $\sum_{j=2}^{p}\gamma_{j1}^2(0) \neq 0$, then this quadratic in $\gamma_{11}(0)$ is minimized by $\gamma_{11}(0)$ equal to

$$\sum_{j=2}^{p}\xi_{1j}\gamma_{j1}(0) / \sum_{j=2}^{p}\gamma_{j1}^2(0). \tag{5}$$

If (5) is not in $[-1, 1]$, then the quadratic is minimized over this interval by setting $\gamma_{11}(0)$ equal to -1, when (5) is less than -1, or setting $\gamma_{11}(0)$ equal to 1, when (5) is greater than 1. If $\sum_{j=2}^{p} \gamma_{j1}^2(0) = 0$, then $\sum_{j=2}^{p} \xi_{1j}\gamma_{j1}(0) = 0$, and there is no dependence on $\gamma_{11}(0)$ so we set $\gamma_{11}(1) = \gamma_{11}(0)$. In any case we replace $\gamma_{11}(0)$ by the value $\gamma_{11}(1)$ that minimizes the quadratic over $[-1, 1]$ or we don't change its value. After updating γ_{11} we next proceed to replace $\gamma_{21}(0)$ by the value $\gamma_{21}(1)$ that minimizes the quadratic in $\gamma_{21}(0)$ over $[-1, 1]$ based on the same argument. We continue cycling through the variables in this way. We call this algorithm *constrained univariate quadratic iteration*.

We see immediately that at each step of the iteration the value of (4) never increases. Since (4) is bounded below by 0, this implies that the iteration converges to a minimum value. The convergence is typically very fast. This minimum value, as we will see, depends on the starting value and so we are not guaranteed to obtain the absolute minimum from a given starting value. Accordingly, we proceed as follows. We select m i.i.d. starting points from the uniform distribution on $B^1(\mathbf{0}) \times \cdots \times B^1(\mathbf{0})$ and compute the m minima $d_1, \ldots, d_m$ via this iterative procedure applied to each starting value. We then estimate $d(\boldsymbol{\Xi}, H_0^1)$ by $d_{(1):m}$, i.e., the smallest order statistic. The values $d_1, \ldots, d_m$ comprise an i.i.d. sample from a distribution with compact support in R^1 and so we have that $d_{(1):m}$ converges in probability to the minimum value in this support, which is the absolute minimum of (4). Actually, computational experience indicates that there are typically a very small number of minima for a given $\boldsymbol{\Xi}$ and often there is only 1. So this represents an efficient method for computing $d(\boldsymbol{\Xi}, H_0^1)$.

The same iterative procedure works for general q with generating the starting values uniformly in $B^{\min(1,q)}(\mathbf{0}) \times \cdots \times B^{\min(p,q)}(\mathbf{0})$. The following result is proved in the Appendix.

Proposition 1. Constrained univariate quadratic iteration with $\boldsymbol{\Gamma}_q(0) \in B^{\min(1,q)}(\mathbf{0}) \times \cdots \times B^{\min(p,q)}(\mathbf{0})$, always gives a nonincreasing sequence of values of (4) and as such converges.

The convergence of $\boldsymbol{\Gamma}_q(k)\boldsymbol{\Gamma}_q'(k) + \boldsymbol{\Psi}(k)$ is not necessary for the assessment of H_0^q but in

our experience it is always the case that this occurs. In fact the $\mathbf{\Gamma}_q(k)$ sequence typically converges to a point in $B^{\min(1,q)}(\mathbf{0}) \times \cdots \times B^{\min(p,q)}(\mathbf{0})$ but this depends on $\mathbf{\Xi}$ and the starting value $\mathbf{\Gamma}_q(0)$. Since $B^{\min(1,q)}(\mathbf{0}) \times \cdots \times B^{\min(p,q)}(\mathbf{0})$ is compact this sequence always has a convergent subsequence but at this point we have not been able to prove that there is only one limit point.

We now consider some examples of using constrained univariate quadratic iteration to compute the distance.

Example 1. First we consider a correlation matrix of the form $\mathbf{\Xi} = \mathbf{\Gamma}_1\mathbf{\Gamma}_1' + \mathbf{\Psi}$, i.e., the correlation matrix arises from a 1-factor model, and we want to compute $d(\mathbf{\Xi}, H_0^1)$. The minimization algorithm should converge to the actual distance 0. Suppose we take $p = 6$ and $\mathbf{\Gamma}_1 = (1, 1, 1, 1, 1, 1)'$ so $\mathbf{\Psi}$ is a matrix of zeros. To assess the performance of the minimization algorithm, we generated 10^3 starting values $\mathbf{\Gamma}_1(0)$ uniformly in $B_1^1(\mathbf{0}) \times \cdots \times B_1^1(\mathbf{0})$ and applied the algorithm to each case. For the stopping rule, we iterated until the squared distance between two successive $\mathbf{\Gamma}_1(i)$ was less than 10^{-5}. In this case all the minimum distances were equal to 0. The mean number of iterations required to attain the minimum was 2.884 and the maximum number required was 4. The convergence was very fast.

Now suppose the correlation matrix possesses a two factor structure, namely, $\mathbf{\Xi} = \mathbf{\Gamma}_2\mathbf{\Gamma}_2' + \mathbf{\Psi}$ and we want to compute $d(\mathbf{\Xi}, H_0^2)$. The exact distance in this case is again 0. Suppose $p = 8$ and

$$\mathbf{\Gamma}_2 = \begin{pmatrix} 0.8 & 0 & 0.8 & 0.8 & 0 & 0 & 0.8 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.8 & 0.8 & 0 & 0.6 \end{pmatrix}'$$

so $\mathbf{\Psi}$ is not the zero matrix in this case. We used 10^3 starting values generated uniformly from a point in $B^1(\mathbf{0}) \times B^2(\mathbf{0}) \times \cdots \times B^2(\mathbf{0})$ and stopped the iteration when the precision 10^{-5} was reached. For a two factor model, the iterative process takes longer with a mean number of 20 and a maximum of 47 iterations but still it converges very quickly. The minimum distance obtained was $d_{(1):m} = 8.678564 \times 10^{-7}$ while the maximum distance was equal to

1.459. To 4 decimal places there were 999 instances of the distance equaling 0. So effectively we found only 2 minima with the absolute minimum equaling 0.

In practice we do not need anything like 10^3 starting values to compute the absolute minimum. It is the case, however, that as q increases the number of iterations required to achieve a given precision increases. We can offset this, however, with a generalization of constrained univariate quadratic iteration where we proceed by iterating on the rows of $\mathbf{\Gamma}_q$ rather than individual entries. This generalization is discussed in Cao (2010).

4 Prior and Posterior Computations

For the sampling model we will assume that the observed data $\mathbf{Y} \in R^{n\times p}$ is a sample from a $N_p(\mu, \mathbf{\Sigma})$ distribution with $(\mu, \mathbf{\Sigma})$ completely unknown. A commonly used proper prior is the conjugate prior given by

$$\mathbf{\Sigma}^{-1} \sim W_p(k_0, \mathbf{A}_0),\ \mu \,|\, \mathbf{\Sigma} \sim N_p(\mu_0, \sigma_0^2\mathbf{\Sigma}) \tag{6}$$

where W_p denotes the Wishart distribution on $p \times p$ matrices and μ, σ_0^2, k_0 and $\mathbf{A}_0$ are hyperparameters. While computations with (6) are straightforward, specifying the hyperparameters is not. Suppose, however, we write $\mathbf{\Sigma} = \Delta^{1/2}\mathbf{\Xi}\Delta^{1/2}$ where $\Delta = \text{diag}(\delta_1, \ldots, \delta_p)$, and specify a prior as

$$\begin{aligned} \mathbf{\Xi} &\sim \text{Uniform}(C^p), \\ \delta_i^{-1} &\sim \text{Gamma}\,(\alpha_{0i}, \beta_{0i}) \text{ for } i = 1, \ldots, p, \\ \mu \,|\, \mathbf{\Sigma} &\sim N_p(\mu_0, \sigma_0^2\mathbf{\Sigma}), \end{aligned} \tag{7}$$

where C^p is the space of all $p\times p$ correlation matrices, and the $(\alpha_{0i}, \beta_{0i})$ are hyperparameters. We note that C^p is a compact set and so the Uniform(C^p) prior is proper. Knowledge about

the manifest variables y_i leads to simple elicitation arguments for the μ_{0i}, σ_0^2, and $(\alpha_{0i}, \beta_{0i})$ hyperparameters (see Cao (2010)). Further (7) avoids any need to elicit a prior for the correlations and the prior is proper. The use of a uniform prior on $\mathbf{\Xi}$ has received some discussion in, for example, Bernard, McCulloch, and Meng (2000), but it is not commonly used.

Perhaps the primary reason for the lack of use of the uniform prior on $\mathbf{\Xi}$ concerns the computations. For our application we need to determine the distributions of $d(\mathbf{\Xi}, H_0^q)$ when $\mathbf{\Xi} \sim \text{Uniform}(C^p)$ for the prior and when $\mathbf{\Xi}$ follows the posterior distribution induced by (7). For the prior there is an excellent algorithm due to Ghosh and Henderson (2003), called the onion method, and so we can simulate from the prior distribution of $d(\mathbf{\Xi}, H_0^q)$ by first generating $\mathbf{\Xi} \sim \text{Uniform}(C^p)$, and then using constrained univariate quadratic iteration to obtain $d(\mathbf{\Xi}, H_0^q)$.

It is much more difficult, however, to simulate exactly or even approximately from the posterior distribution of $d(\mathbf{\Xi}, H_0^q)$. For example, it is not clear at all that there is a good Gibbs sampling algorithm for this distribution as it is very complicated to work with. Of course, our goal is not to simulate from this posterior but rather evaluate (2). We develop an effective importance sampler for this purpose that in fact allows for priors more general than (7). The following result is proved in the Appendix.

Proposition 2. Suppose that $\mathbf{Y} \in R^{n \times p}$ is a sample from a $N_p(\mu, \mathbf{\Sigma})$ distribution, $\mu \,|\, \mathbf{\Sigma} \sim N_p(\mu_0, \sigma_0^2 \mathbf{\Sigma})$ and let π_1 denote a prior on Δ and π_2 a prior on $\mathbf{\Xi}$ with Δ and $\mathbf{\Xi}$ *a priori* independent. Then the posterior density of $\mathcal{J} = \mathbf{\Sigma}^{-1}$ is proportional to a $W_p(n - p + 1, A^{-1}(\mathbf{Y}))$ density times

$$k(\mathcal{J}) = \pi_1(\text{diag}(\mathcal{J}^{-1}))\pi_2((\text{diag}(\mathcal{J}^{-1}))^{-1/2}\mathcal{J}^{-1}(\text{diag}(\mathcal{J}^{-1}))^{-1/2})|\text{diag}(\mathcal{J}^{-1})|^{-(p-1)/2}$$

where $\mathbf{A}(\mathbf{Y}) = (n-1)\mathbf{S} + n(n\sigma_0^2 + 1)^{-1}(\bar{\mathbf{y}} - \mu_0)(\bar{\mathbf{y}} - \mu_0)'$.

This shows that the posterior density of $\mathbf{\Sigma}^{-1}$ factors as a Wishart density times a function of $\mathcal{J}^{-1} = \mathbf{\Sigma}$ that does not involve the data $\mathbf{Y}$. This function only depends on the prior π_1 on Δ and the prior π_2 on $\mathbf{\Xi}$. Depending on how we choose π_1 and π_2, this can lead to a simple importance sampling algorithm for approximating integrals with respect to the posterior. For our importance sampler we will use the $W_p(\mathbf{A}^{-1}(\mathbf{Y}), n-p+1)$ distribution. This is typically the "hard" part of the density as it contains all the dependence on the data.

The following result is proved in the Appendix.

Corollary 3. If π_2 is the uniform prior on $\mathbf{\Xi}$, and

(i) π_1 is the product of inverse Gamma $(\alpha_{0i}, \beta_{0i})$ densities, then

$$k(\mathcal{J}) = \prod_{i=1}^{p} {\sigma_{ii}}^{-\alpha_{0i}-(p+1)/2} \exp\{-\beta_{0i}/\sigma_{ii}\},$$

(ii) π_1 is the product of log-$N(\alpha_{0i}, \beta_{0i})$ densities, then

$$k(\mathcal{J}) = \prod_{i=1}^{p} \beta_{0i}^{-1} \sigma_{ii}^{-1-(p+1)/2} \exp\{-(\log \sigma_{ii} - \alpha_{0i})^2/2\beta_{0i}^2\}.$$

Note that in both cases $k(\mathcal{J})$ is a fairly simple function of the σ_{ii}.

In the following section we present examples of computing the prior and posterior densities of $d(\mathbf{\Xi}, H_0^q)$ for various q and applying these to the computation of (2). One numerical problem that arises in the computation of (2) is the need to compute $\pi_d(0 \,|\, x)/\pi_d(0)$. This is difficult because, at least when calculating the densities with respect to length measure, both terms in the ratio are close to 0. What we really want, however, is $\lim_{\epsilon \searrow 0} \pi_d(\epsilon \,|\, x)/\pi_d(\epsilon)$. Accordingly, we approximate this limit by $\pi_d(d_\alpha \,|\, x)/\pi_d(d_\alpha)$ where d_α is the α-th prior quantile of $d(\mathbf{\Xi}, H_0^q)$ since $d_\alpha \to 0$ as $\alpha \to 0$. We choose α small, e.g., $\alpha = 0.01$, and the Monte Carlo sample size large enough so that an accurate estimate can be obtained for d_α. This procedure was found to work well in a wide variety of examples.

5 Inferences and Examples

For inferences for factor analysis we proceed as follows. First we assess the hypothesis H_0^0. If we obtain evidence against H_0^0, via the evaluation of (2), then we proceed to assess H_0^1, via the corresponding value of (2), etc. So proceeding iteratively in this fashion we stop at the $(q+1)$-st step when we obtain evidence against H_0^{q+1} and then proceed as if H_0^q is true. This approach seems natural as our goal is to fit the factor model with the minimum number of factors required to explain the observed correlations.

Suppose then that we have selected the factor model with q factors. This says that the true correlation matrix satisfies (1) and we must estimate Ξ based on this restriction. Naturally we want to choose our estimate to be in H_0^q. Since this set is not convex, it doesn't make sense to use the conditional expectation given that $\Xi \in H_0^q$, as we are not guaranteed this value is in H_0^q. The conditional posterior density of Ξ, given that $d(\boldsymbol{\Xi}, H_0^q) = 0$, is not available in closed form and this makes computing the conditional posterior mode virtually impossible.

An alternative approach is to maximize the integrated likelihood conditional on $\Xi \in H_0^q$ and this can be be shown to be equivalent to maximizing the likelihood, after integrating out μ and Δ, restricted to $\Xi \in H_0^q$. Again this is a very difficult computation and we can expect algorithmic difficulties similar to those that occur with likelihood methods. Our approach here, however, leads to a natural approximation to this quantity.

For this let $\hat{\Xi}$ denote the plug-in MLE of Ξ, i.e., the estimate obtained from the MLE of Σ given by $(n-1)n^{-1}S$. Now let $\hat{\Xi}_q$ denote the point in H_0^q that is closest to $\hat{\Xi}$ and note that this can be computed by our algorithm. We take $\hat{\Xi}_q$ as our estimate of Ξ. Certainly, when H_0^q is true, then $\hat{\Xi}$ converges to a point in H_0^q and so $\hat{\Xi}_q$ will also converge to the true value. Furthermore, we can quantify the uncertainty in $\hat{\Xi}_q$ by looking at the posterior distribution of $d(\boldsymbol{\Xi}, \hat{\Xi}_q)$ and comparing this to its prior distribution to assess how much the data have

increased our belief in the estimate.

Corresponding to $\hat{\Xi}_q$ there is a $\hat{\mathbf{\Gamma}}_q$ and so we could use this as an estimate of the factor loadings. It should be noted, however, that there is in general no guarantee that there is only one such lower triangular matrix of factor loadings that will satisfy $\hat{\Xi}_q = \hat{\mathbf{\Gamma}}_q \hat{\mathbf{\Gamma}}_q' + \hat{\Psi}$, where $\hat{\Psi}$ is determined from $\hat{\mathbf{\Gamma}}_q$ as previously described. Of course we can also postmultiply $\hat{\mathbf{\Gamma}}_q$ by a $q \times q$ orthogonal matrix to obtain more interpretable loadings.

We now consider implementing these inferences in several examples.

Example 2. *Simulated Data*

We first consider simulating the data Y from a distribution with correlation matrix $\mathbf{\Xi} \in H_0^2$ where $\mathbf{\Gamma}_2 \in R^{8\times 2}$ is as given in Example 1. In this case the entries in the correlation matrix above the main diagonal are as given in Table 1.

0.00	0.64	0.64	0.00	0.00	0.64	0.48
	0.00	0.00	0.64	0.64	0.00	0.48
		0.64	0.00	0.00	0.64	0.48
			0.00	0.00	0.64	0.48
				0.64	0.00	0.48
					0.00	0.48
						0.48

Table 1: The true correlation matrix (above diagonal) in Example 2.

The prior chosen was as given in (7) with $(\alpha_{0i}, \beta_{0i}) = (1, 1)$ for $i = 1, \ldots, p$, $\mu_0 = \mathbf{0}$ and $\sigma_0^2 = 1$. We generated a sample of $n = 200$ values from the $N_8(\mathbf{0}, \mathbf{\Gamma}_2\mathbf{\Gamma}_2' + \mathbf{\Psi})$ distribution. In Figure 1 we have plotted the prior and posterior densities for the $q = 0, 1, 2$ cases. We see that the posterior becomes increasingly concentrated about 0 as q increases. The P-values given by (2) equal 0 for $q = 0, 1.7 \times 10^{-5}$ for $q = 1$, and 1 for $q = 2$. So we have evidence against H_0^0 and H_0^1, and no evidence against H_0^2.

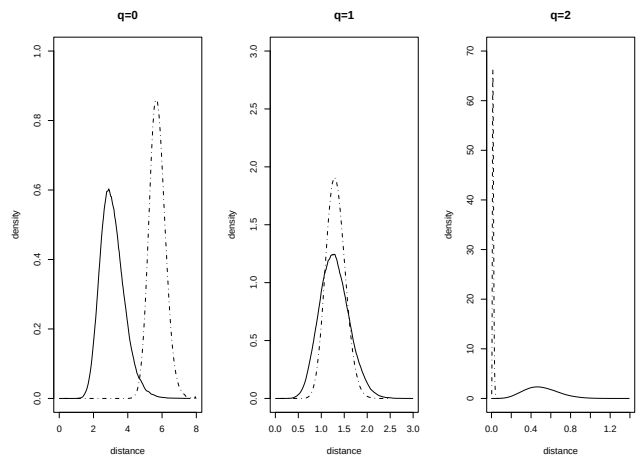

Figure 1: The prior (—) and the posterior (– –) densities of the distance for $q = 0, 1$, and 2 in Example 2.

We obtained

$$\hat{\mathbf{\Gamma}}_2 = \begin{pmatrix} 0.804 & 0.046 & 0.784 & 0.755 & 0.068 & 0.064 & 0.774 & 0.717 \\ 0.000 & 0.797 & 0.076 & 0.067 & 0.783 & 0.785 & -0.021 & 0.528 \end{pmatrix}'$$

and from this we obtain $\hat{\Xi}_2$ as provided in Table 2. We see that our procedure has performed quite well.

0.0370	0.630	0.607	0.054	0.052	0.622	0.576
	0.097	0.088	0.627	0.628	0.019	0.454
		0.597	0.112	0.110	0.605	0.602
			0.103	0.101	0.583	0.576
				0.618	0.036	0.462
					0.033	0.460
						0.544

Table 2: The estimated correlation matrix (above diagonal) in Example 2 when $q = 2$.

To assess the accuracy of the importance sampling, we estimated the coefficient of variation of the importance sampling estimator of the normalizing constant for the posterior. For a Monte Carlo sample of size N, the coefficient of variation is given by $CV =$

$N^{-1}(N\sum_{i=1}^{N} w_{*i}^2 - 1)$ where, for this case, $w_{*i} = k(\mathcal{J}_i)/\sum_{j=1}^{N} k(\mathcal{J}_i)$. A value of $\sum_{i=1}^{N} w_{*i}^2$ close to $1/N$ indicates that the importance sampling is working while a value near 1 indicates a failure. In this case, we used $N = 10^5$ for the posterior calculations and we obtained the values in Table 3. The results indicate that the importance sampling is working very well and this was found to be the case in a number of examples (see Cao (2010) for further discussion).

	$\sum_{i=1}^{N} w_{*i}^2$	CV
$q = 0$	1.938×10^{-5}	9.38×10^{-6}
$q = 1$	1.776×10^{-5}	7.76×10^{-6}
$q = 2$	1.780×10^{-5}	7.80×10^{-6}

Table 3: Sums of squared normalized importance sampling weights and coefficients of variation in Example 2.

Example 3. *Currency Exchange Data*

We now consider a data set involving monthly international exchanges rates ($n = 144$) available in West and Harrison (1997). These time series are the monthly changes in exchange rates in British pounds of the following $p = 6$ currencies: US dollar (US), Canadian dollar (CAN), Japanese yen (JAP), French franc (FRA), Italian lira (ITA) and the German Deutschmark (GER). The data span the period from January of 1975 to December of 1986 inclusive. For this data the correlation matrix is given by Table 4.

	CAN	JAP	FRA	ITA	GER
US	0.858	0.801	−0.453	−0.501	0.148
CAN		0.429	−0.144	−0.075	0.191
JAP			−0.446	−0.652	0.043
FRA				0.922	0.068
ITA					0.067

Table 4: Sample correlations in Example 3.

In West and Harrison (1997), it was determined that up to three principal components are needed by using various principal component analyses. Although the dimension has been

reduced to half, it offers no simplification based on a degrees of freedom argument. In other words, the factor model contains as many parameter as $\mathbf{\Sigma}$. Also the maximum likelihood approach failed to converge for this data set.

Lopes and West (2004) developed a reversible jump MCMC algorithm to handle the change in dimension as q changes, for the model with a prior on the model parameters for each q. They chose very diffuse but proper priors. They concluded that two factors are needed to explain the correlation structure.

Our analysis is based on the prior specified in (7). As noted this greatly simplifies the specification of the prior. For the $N_6(\mu_0, \sigma_0^2\mathbf{\Sigma})$ prior we took $\mu_0 = (1.8,\ 2.1,\ 430.6,\ 10.1,\ 1984,\ 12.4)'$, where the individual entries are the sample means of these quantities, and $\sigma_0^2 = 100$ was chosen to reflect ignorance about the locations parameters. We note that all other Bayesian factor analyses that we are aware of, have effectively set $\sigma_0^2 = 0$ so that all the prior probability for μ is concentrated at the sample means, while this is not a necessity in our approach. For the scaling parameters, we used $(\alpha_{0i}, \beta_{0i}) = (2.2, 0.1)$ for $i = 1, \ldots, 6$. These choices are those made in Lopes and West (2004) and are made here for comparison purposes. In an actual application we would want to elicit the values of μ_0, σ_0^2 and the $(\alpha_{0i}, \beta_{0i})$.

First we test the hypothesis H_0^0. The posterior and prior densities of $d(\mathbf{\Xi}, H_0^0)$ are plotted in in Figure 2.

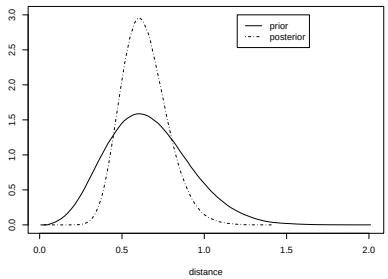

Figure 2: The posterior and prior densities of $d(\mathbf{\Xi}, H_0^0)$ in Example 3.

We see that the posterior leads to much less concentration around 0 than the prior. The P-value (2) equals 0 (to four decimal places) so we have strong evidence against the null hypothesis and conclude that $q > 0$, i.e., an independence model is not appropriate.

We now test H_0^1. The posterior and prior density comparison is presented in Figure 3.

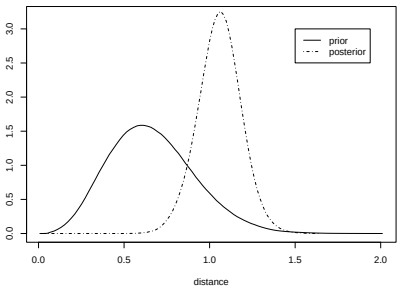

Figure 3: The posterior and prior densities of $d(\boldsymbol{\Xi}, H_0^1)$ in Example 3.

The plot again shows that posterior leads to much less concentration around 0 than the prior and the value of (2) is 0.0001. Thus we have strong evidence against the null hypothesis and conclude that $q > 1$, i.e., a model with more than 1 factor is needed.

We now proceed to assess H_0^2 and plot the prior and posterior densities in Figure 3.

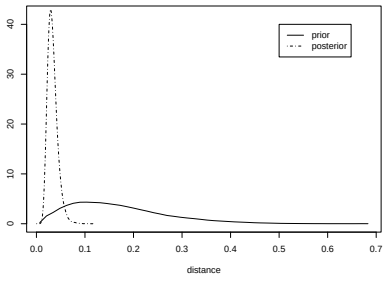

Figure 4: The posterior and prior densities of $d(\boldsymbol{\Xi}, H_0^2)$ in Example 3.

This plot shows that posterior concentrates much more around 0 than the prior does. The relative belief ratio at 0 is approximated by $\pi(d_\alpha|\mathbf{X})/\pi(d_\alpha) = 17.2$, where $d_\alpha = 0.018$ with $\alpha = 0.01$ and (2) equals 0.467. Therefore we do not have evidence against H_0^2 and we

conclude that a 2-factor model is reasonable. This conclusion agrees with Lopes and West (2004).

We now estimate Ξ, given that $\Xi \in H_0^2$. Based on the entries in Table 4 we obtained

$$\widehat{\mathbf{\Gamma}}_2 = \begin{pmatrix} 1.0 & 0.82345 & 0.72684 & -0.45423 & -0.50842 & 0.14552 \\ 0 & 0.37541 & -0.27513 & 0.74464 & 0.86111 & 0.17547 \end{pmatrix}',$$

and so the estimate $\hat{\Xi}_2$ is given by the entries of Table 5.

	CAN	JAP	FRA	ITA	GER
US	0.823	0.727	−0.454	−0.508	0.146
CAN		0.495	−0.094	−0.095	0.186
JAP			−0.535	−0.606	0.057
FRA				0.872	0.065
ITA					0.077

Table 5: Estimated correlations in Example 3 based on H_0^2.

We note that these are all relatively close to the values in Table 4 so the two factor model seems appropriate. In Figure 5 we plot the prior and posterior densities of the distance from $\hat{\Xi}_2$. This clearly shows a huge increase in belief for the estimate.

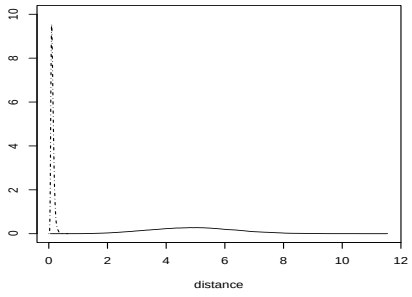

Figure 5: Plots of prior (—) and posterior (– –) densities of distance to $\hat{\Xi}_2$ in Example 3.

6 Conclusions

We have developed a Bayesian approach to factor analysis which has several attractive features. First we only need to place a prior on the full model parameter $(\mu, \mathbf{\Sigma})$ rather than a prior on each submodel. This cuts down on the need for extensive and difficult elicitation or the imposition of default improper priors. Further we have developed a computational approach that allows the use of a uniform prior on the correlation matrix. Accordingly, we are only required to elicit priors for location and scaling parameters which can be easily carried out in a variety of ways. The methodology is seen to work well in a variety of examples. Further we believe that the same approach can be applied to a number of other statistical problems and such applications are currently being developed.

7 Appendix

Proof of Proposition 1 We note first that if $ax^2 + bx + c$ is such that $a > 0$, then the minimum of the quadratic occurs at $-b/2a$. If this value is not in an interval $[l, u]$ then the minimum of the quadratic over the interval occurs at l or u.

Note that when $r < s$, then $\gamma_{rs} = 0$ and so we need only consider the case $r \geq s$. We see that γ_{rs} occurs in terms of (4) only when $i = r$ or $j = r$, so we can write (4) as

$$\sum_{j=r+1}^{p} \left(\sigma_{rj} - \sum_{k=1}^{\min(r,q)} \gamma_{rk}\gamma_{jk} \right)^2 + \sum_{i=1}^{r-1} \left(\sigma_{ir} - \sum_{k=1}^{\min(i,q)} \gamma_{ik}\gamma_{rk} \right)^2 + c \qquad \text{(A1)}$$

where c is some constant not involving γ_{rs}. Now when $s = r \leq q$, then the second term in (A1) does not involve γ_{rr} and so we need only consider the first term. We see immediately

that, as a function of γ_{rr}, the first term can be written as

$$\left(\sum_{j=r+1}^{p}\gamma_{jr}^2\right)\gamma_{rr}^2 - 2\left[\sum_{j=r+1}^{p}\left(\sigma_{rj}-\sum_{k=1}^{r-1}\gamma_{rk}\gamma_{jk}\right)\gamma_{jr}\right]\gamma_{rr}+c \tag{A2}$$

where c is a constant. Note that if $\sum_{j=r+1}^{p}\gamma_{jr}^2=0$, then the coefficient of γ_{rr} in (A2) is also 0. When $s<r$, then note that $s\le q$ and (A1) can be written as

$$\begin{array}{l}\left(\sum_{j=\max(r+1,s)}^{p}\gamma_{js}^2+\sum_{i=s}^{r-1}\gamma_{is}^2\right)\gamma_{rs}^2- \\ 2\left[\begin{array}{c}\sum_{j=\max(r+1,s+1)}^{p}\left(\sigma_{rj}-\sum_{k=1,k\neq s}^{\min(r,q)}\gamma_{rk}\gamma_{jk}\right)\gamma_{js}+ \\ \sum_{i=s}^{r-1}\left(\sigma_{ir}-\sum_{k=1,k\neq s}^{\min(i,q)}\gamma_{ik}\gamma_{rk}\right)\gamma_{is}\end{array}\right]\gamma_{rs}+c\end{array} \tag{A3}$$

where c is a constant. Again if the coefficient of γ_{rs}^2 in (A3) is 0, then the coefficient of γ_{rs} is 0. Now for $1\le r\le q$, then

$$\gamma_{rr}\in\left[-\left(\sigma_{rr}-\sum_{k=1}^{r-1}\gamma_{rk}^2\right)^{1/2},\left(\sigma_{rr}-\sum_{k=1}^{r-1}\gamma_{rk}^2\right)^{1/2}\right] \tag{A4}$$

and

$$\gamma_{rs}\in\left[-\left(\sigma_{rr}-\sum_{k=1,k\neq s}^{\min(r,q)}\gamma_{rk}^{2\,1/2}\right),\left(\sigma_{rr}-\sum_{k=1,k\neq s}^{\min(r,q)}\gamma_{rk}^2\right)^{1/2}\right] \tag{A5}$$

otherwise.

Now just as in the $q=1$ case we can select a starting $\Gamma_q(0)$ and then iterate using (A2) and (A3), to minimize each quadratic over the relevant interval as determined by (A4) and (A5). So we might start with $\gamma_{11}(0)$ replacing it by $\gamma_{11}(1)$, and then, with this new value for γ_{11}, replace $\gamma_{21}(0)$ by $\gamma_{21}(1)$, etc. At each step the distance (4) does not increase and so the sequence of distances converges to a minimum.

Proof of Proposition 2 The Jacobian of the transformation $(\Delta,\boldsymbol{\Xi})\to\boldsymbol{\Sigma}$ is given by

$|\mathrm{diag}(\mathbf{\Sigma})|^{-(p-1)/2}$ and so the joint prior on $(\mu, \mathbf{\Sigma})$ is proportional to

$$|\mathbf{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu-\mu_0)'\mathbf{\Sigma}^{-1}(\mu-\mu_0)\right\} \pi_1((\mathrm{diag}(\mathbf{\Sigma}))^{1/2})\times$$
$$\pi_2((\mathrm{diag}(\mathbf{\Sigma}))^{-1/2}\mathbf{\Sigma}(\mathrm{diag}(\mathbf{\Sigma}))^{-1/2})|\mathrm{diag}(\mathbf{\Sigma})|^{-(p-1)/2}.$$

Making the change of variable $\mathbf{\Sigma} \to \mathcal{J} = \mathbf{\Sigma}^{-1}$, which has Jacobian $|\mathbf{\Sigma}|^{p+1}$, the prior density of $(\mu, \mathcal{J})$ is

$$|\mathcal{J}|^{-(p+1/2)}\mathrm{etr}\left\{-\frac{1}{2\sigma_0^2}\mathcal{J}(\mu-\mu_0)(\mu-\mu_0)'\right\} \pi_1(\mathrm{diag}(\mathcal{J}^{-1}))\times$$
$$\pi_2((\mathrm{diag}(\mathcal{J}^{-1}))^{-1/2}\mathcal{J}^{-1}(\mathrm{diag}(\mathcal{J}^{-1}))^{-1/2})|\mathrm{diag}(\mathcal{J}^{-1})|^{-(p-1)/2}.$$

The likelihood function is proportional (up to a constant function of the data and parameters) to $|\mathbf{\Sigma}|^{-n/2}\mathrm{etr}\{-(n/2)\mathbf{\Sigma}^{-1}(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)' - (n-1)/2)\mathbf{S}\mathbf{\Sigma}^{-1}\}$, so the joint posterior of $(\mu, \mathcal{J})$ is similarly proportional to

$$|\mathcal{J}|^{n/2-(p+1/2)}\mathrm{etr}\left\{\begin{array}{c} -\frac{1}{2\sigma_0^2}\mathcal{J}(\mu-\mu_0)(\mu-\mu_0)' - \frac{n}{2}\mathcal{J}(\bar{\mathbf{y}}-\mu)(\bar{\mathbf{y}}-\mu)' \\ -\frac{(n-1)}{2}\mathbf{S}\mathcal{J} \end{array}\right\}\times$$
$$\pi_1(\mathrm{diag}(\mathcal{J}^{-1}))|\pi_2((\mathrm{diag}(\mathcal{J}^{-1}))^{-1/2}\mathcal{J}^{-1}(\mathrm{diag}(\mathcal{J}^{-1}))^{-1/2})|\mathrm{diag}(\mathcal{J}^{-1})|^{-(p-1)/2}.$$

Now we have that

$$\sigma_0^{-2}(\mu-\mu_0)'\mathcal{J}(\mu-\mu_0) + n(\mu-\bar{\mathbf{y}})'\mathcal{J}(\mu-\bar{\mathbf{y}})$$
$$= (\sigma_0^{-2}+n)\left(\mu - \frac{\sigma_0^{-2}\mu_0 + n\bar{\mathbf{y}})}{\sigma_0^{-2}+n}\right)'\mathcal{J}\left(\mu - \frac{\sigma_0^{-2}\mu_0 + n\bar{\mathbf{y}})}{\sigma_0^{-2}+n}\right)$$
$$+ (\sigma_0^2 + n^{-1})^{-1}(\mu_0-\bar{\mathbf{y}})'\mathcal{J}(\mu_0-\bar{\mathbf{y}})$$

Integrating out μ gives that the posterior of $\mathcal{J}$ is proportional (up to a constant function of

the data and parameters) to

$$|\mathcal{J}|^{n/2-p}\text{etr}\left\{-\frac{A(\mathbf{Y})\mathcal{J}}{2}\right\}\pi_1(\text{diag}(\mathcal{J}^{-1}))\times$$
$$\pi_2((\text{diag}\mathcal{J}^{-1})^{-1/2}\mathcal{J}^{-1}(\text{diag}\mathcal{J}^{-1})^{-1/2})|\text{diag}(\mathcal{J}^{-1})|^{-(p-1)/2}$$

where $A(\mathbf{Y}) = (n-1)\mathbf{S} + (\sigma_0^2 + 1/n)^{-1}(\bar{\mathbf{y}} - \mu_0)(\bar{\mathbf{y}} - \mu_0)'$.

Proof of Corollary 3. When $\delta_i \sim$ Inverse Gamma $(\alpha_{0i}, \beta_{0i})$, and putting $\sigma_{ii} = \delta_i$, then $k(\mathcal{J}) = \prod_{i=1}^{p}(1/\delta_i)^{\alpha_{0i}+1}\exp(-\beta_{0i}/\delta_i)(\delta_i)^{-(p-1)/2} =$
$\prod_{i=1}^{p}(\sigma_{ii})^{-\alpha_{0i}-(p+1)/2}\exp(-\beta_i/\sigma_{ii})$. A similar calculation applies to the log-normal case.

References

Bartholomew, D.J. (1987) Latent Variable Models and Factor Analysis. Charles Griffin, London.

Bernard, J., McCulloch, R. and Meng, X. L. (2000) Modeling covariance matrices in terms of standard deviations and correlations with application to shrinkage. Statistica Sinica 10, 1281-1311.

Cao, Y. (2010) A Bayesian Approach to Factor Analysis via Comparing Posterior and Prior Concentration. Ph.D. Thesis, Dept. of Statistics, University of Toronto.

Evans, M., Gilula, Z. and Guttman, I. (1993) Computational issues in the Bayesian analysis of categorical data : loglinear and Goodman's RC model. Statistica Sinica, 3, 391-406.

Evans, M., Gilula, Z., Guttman, I., and Swartz, T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. JASA, Vol. 92, No. 437, 208-214.

Evans, M. and Shakhatreh, M. (2008). Optimal properties of some Bayesian inferences. Electronic Journal of Statistics, Vol. 2, 1268-1280.

Ghosh, S. and Henderson, S. G. (2003), Behavior of the NORTA method for correlated random vector generation as the dimension increases. ACM Transactions on Modeling and Computer Simulation (TOMACS), Vol. 13, 276-294.

Ghosh, J.K., Delampady, M. and Samanta, T. (2006) An Introduction to Bayesian Analysis: Theory and Methods. Springer, New York.

Kaufman, G. and Press, S.J. (1973) Bayesian factor analysis. Working paper 662-73, Sloan School of Management, MIT.

Lee, S. E., and Press, S. J. (1998) Robustness of Bayesian factor analysis estimates. Communications in. Statistics—Theory and Methods, 27, 1871–1893.,

Lee, S-Y. (2007) Structural Equation Modeling, A Bayesian Approach. John Wiley & Sons, New York.

Lopes, H.F. and West, M. (2004) Bayesian model assessment in factor analysis. Statistica Sinica 14, 41-67.

Press, S.J. and Shigemasu, K. (1989) Bayesian inference in factor analysis. Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin, edited by L.J. Gleser, M.D. Perlman, S.J. Press and A.R. Sampson. Springer-Verlag, New York.

West, M. and Harrison, J. (1997) Bayesian Forecasting and Dynamic Models (2nd edition). Springer-Verlag, New York.