Journal of Data Science **7**(2009), 285-311

# Bayesian Computation of the Intrinsic Structure of Factor Analytic Models

Ernest Fokoué
*Rochester Institute of Technology*

*Abstract*: The study of factor analytic models often has to address two important issues: (a) the determination of the "optimum" number of factors and (b) the derivation of a unique simple structure whose interpretation is easy and straightforward. The classical approach deals with these two tasks separately, and sometimes resorts to ad-hoc methods. This paper proposes a Bayesian approach to these two important issues, and adapts ideas from stochastic geometry and Bayesian finite mixture modelling to construct an ergodic Markov chain having the posterior distribution of the complete collection of parameters (including the number of factors) as its equilibrium distribution. The proposed method uses an *Automatic Relevance Determination (ARD)* prior as the device of achieving the desired simple structure. A Gibbs sampler updating scheme is then combined with the simulation of a continuous-time birth-and-death point process to produce a sampling scheme that efficiently explores the posterior distribution of interest. The MCMC sample path obtained from the simulated posterior then provides a flexible ingredient for most of the inferential tasks of interest. Illustrations on both artificial and real tasks are provided, while major difficulties and challenges are discussed, along with ideas for future improvements.

*Key words:* Birth-and-death process, factor analysis, interpretability, intrinsic dimensionality, simple structure, sparsity, posterior simulation.

## 1. Introduction

### 1.1 The factor analysis model

Factor Analysis (FA) assumes that a $p$-dimensional manifest random vector is made up of highly correlated variables that can be grouped by their correlations. Under this assumption, variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables belonging to a different group. Each group of variables can therefore be thought of as the representation of a single underlying construct also known as a *factor* or more precisely a *common factor*. Central to factor analysis is the assumption that factors are responsible for the observed correlations, the consequence of such an assumption being that the observed variables are essentially independent given the factors.

From a modeling standpoint, the above factor analysis assumption implies that each observable random vector $\mathbf{X}_i$ can be expressed as a linear combination of $q < p$ latent random variables $(F_{i1}, F_{i2}, \cdots, F_{iq})^\top = \boldsymbol{F}_i$, called *common factors*, plus a mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^\top$, plus $p$ additional sources of variation $(\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{ip})^\top = \boldsymbol{\epsilon}_i$ referred to as idiosyncratic *disturbances.*

$$\underset{(p\times 1)}{\mathbf{X}_i} = \underset{(p\times 1)}{\boldsymbol{\mu}} + \underset{(p\times q)}{\boldsymbol{\Lambda}} \underset{(q\times 1)}{\boldsymbol{F}_i} + \underset{(p\times 1)}{\boldsymbol{\epsilon}_i.}$$

For simplicity, it will be assumed that $\boldsymbol{\mu} = \mathbf{0}$. As a result,

$$\underset{(p\times 1)}{\mathbf{X}_i} = \underset{(p\times q)}{\boldsymbol{\Lambda}} \underset{(q\times 1)}{\boldsymbol{F}_i} + \underset{(p\times 1)}{\boldsymbol{\epsilon}_i.} \tag{1.1}$$

The matrix $\boldsymbol{\Lambda} \in I\!R^{p\times q}$ is referred to as the *matrix of factor loadings*. Each element $\lambda_{ij}$ of $\boldsymbol{\Lambda}$ is called the *loading* of the $i$th variable on the $j$th factor. The orthogonal FA model assumes that $\boldsymbol{\epsilon}_i$ and $F_i$ are independent, so that with $\boldsymbol{\epsilon}_i \sim \mathcal{N}_p(0, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \text{diag}(\psi_1^2, \cdots, \psi_p^2)$. It is easy to see that $\text{cov}(\mathbf{X}_i, \boldsymbol{F}_i) = \boldsymbol{\Lambda}$, and that $\text{cov}(\boldsymbol{\epsilon}_i, \boldsymbol{F}_i) = \boldsymbol{E}\left[\boldsymbol{\epsilon}_i \boldsymbol{F}_i^\top\right] = 0$. All these assumptions imply that the conditional density of the data given realizations of factor scores is

$$\begin{aligned}\boldsymbol{p}(\mathbf{X}_i|\boldsymbol{F}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) &= (2\pi)^{-p/2}|\boldsymbol{\Psi}|^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\Lambda}\boldsymbol{F}_i)^\top \boldsymbol{\Psi}^{-1}(\mathbf{X}_i - \boldsymbol{\Lambda}\boldsymbol{F}_i)\right].\end{aligned} \tag{1.2}$$

From the above (1.2), it is easy to see that the posterior distribution of $\boldsymbol{F}$ is

$$[\boldsymbol{F}_i|\mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}] \sim \mathcal{N}_q\left([\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}_i, [\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\right). \tag{1.3}$$

By integrating $\boldsymbol{F}$ out from $\boldsymbol{p}(\mathbf{X}_i, \boldsymbol{F}|\boldsymbol{\Lambda}, \boldsymbol{\Psi})$, the marginal density of the data is given by

$$\boldsymbol{p}(\mathbf{X}_i|\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = (2\pi)^{-p/2}|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{X}_i^\top\left[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right]^{-1}\mathbf{X}_i\right] \tag{1.4}$$

For convenience, we will use the notation of (1.5) to refer to the marginal and the conditional distributions of $\mathbf{X}_i$, respectively.

$$\mathbf{X}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \quad \text{and} \quad [\mathbf{X}_i|\boldsymbol{F}_i] \sim \mathcal{N}_p(\boldsymbol{\Lambda}\boldsymbol{F}_i, \boldsymbol{\Psi}). \tag{1.5}$$

The description of the FA model so far has used $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Let $\boldsymbol{\theta} \equiv (\mathbf{\Lambda}, \mathbf{\Psi})$ denote the collection of all the parameters of the model. The observed-data likelihood is then given by

$$\boldsymbol{L}(\boldsymbol{\theta}; \mathbf{X}) \propto |\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathbf{X}_i^\top \left[\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}\right]^{-1} \mathbf{X}_i\right]. \tag{1.6}$$

On the other hand, treating the factor scores are unknowns in the same way as parameters, the complete-data likelihood is given by

$$\boldsymbol{L}(\boldsymbol{\theta}; \mathbf{X}, \boldsymbol{F}) \propto |\mathbf{\Psi}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{\Lambda}\boldsymbol{F}_i)^\top \mathbf{\Psi}^{-1} (\mathbf{X}_i - \mathbf{\Lambda}\boldsymbol{F}_i)\right]. \tag{1.7}$$

## 1.2 Position of the problem

The factor analysis model, as we have defined it, has been extensively studied by statisticians, economists, scientists, machine learning specialists, pattern recognition engineers and psychometricians. The estimation of the factor loading matrix $\mathbf{\Lambda}$ in particular has been explored both theoreticians and practitioners. Although the vast literature on the topic mainly covered the frequentist treatment for decades, it is encouraging to notice that the Bayesian perspective has recently received the long awaited attention from various authors, amongst whom Press (1972), Martin and McDonald (1981), Press and Shigemasu (1989), Ihara and Kano (1995), Press and Shigemasu (1998), Lopes and West (1999), Fokoué and Titterington (2003), Rowe (2003). This paper proposes a contribution to the Bayesian perspective on some important issues that arise in factor analytic modeling. Indeed, besides the important issue of parameter estimation that has been widely studied from both the frequentist and Bayesian perspectives, two of the other most commonly studied issues in factor analysis are:

- The search for a unique simple structure: In many applications of factor analysis, the investigator seeks a unique simple structure for which a straightforward interpretation can be provided. This is clearly an ill-defined problem because of the lack of identifiability inherent to the FA model. Well-posedness is usually achieved by imposing constraints. This paper explores two ways of obtaining a factor solution that is easy to interpret.
- The determination of the intrinsic number of factors: The expression *intrinsic dimensionality* is used throughout this paper to mean the optimum number of factors. Clearly, for any given factor analysis problem (task), this definition of intrinsic dimensional naturally raises three important questions,

namely : (a) Does such a number of exist? (b) If such a number exists, is it unique? (c) What is the method that helps find such a number as efficiently as possible? Questions (a) and (b) are addressed theoretically in a later paper. Both ad-hoc and principled methods have been developed to answer question (c). This paper proposes a Bayesian approach implemented through the simulation of a stochastic birth-and-death process.

The remainder of this paper is organized as follows: section 2 provides a brief description of parameter estimation in Bayesian Factor Analysis when $q$ is known and fixed. *In our derivation of full conditional posterior distribution, we will use the notation* $[\theta_*|\cdots]$ *to denote the conditional distribution of* $\theta_*$ *given all the quantities on which it depends. For instance, since the posterior distribution of* $\mathbf{\Lambda}$ *depends on* $\mathbf{\Psi}$*,* $\boldsymbol{F}$ *and* $\mathbf{X}$*, it is more complete to write* $[\mathbf{\Lambda}|\mathbf{\Psi}, \boldsymbol{F}, \mathbf{X}]$*. For simplicity however, we will simply write* $[\mathbf{\Lambda}|\cdots]$*, with the implied meaning given earlier.* Section 3 deals with the search for as a simple factor structure. The section first touches on some ways to address the issue of identifiability in the Bayesian setting, and concludes with the specification of an *Automatic Relevance Determination* prior that achieves a simple factor structure by putting a *sparsity* pressure on the space of factor loadings. Section 4 addresses the determination of the number of factors, beginning with an overview of existing methods, and concluding with the details of the proposed approach. Section 5 presents numerical results on both artificial and real data. The last section provides the conclusion and ideas for future improvements.

## 2. Bayesian Factor Analysis via Posterior Simulation

In this section, we assume that the number of factors $q$ in known and fixed. As we shall see later, this is generally either set by the experimenter or estimates via ad-hoc techniques that will be mentioned later. Now, given a random sample of size $n$, the maximum likelihood estimate $\hat{\mathbf{\Lambda}}$ of the $p \times q$ matrix of factor loadings $\mathbf{\Lambda}$ is given by

$$\hat{\mathbf{\Lambda}}_{\mathsf{MLE}} = \mathsf{arg}\max_{\mathbf{\Lambda}} \left\{ |\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}|^{-\frac{n}{2}} \exp\left[ -\frac{1}{2}\sum_{i=1}^{n} \mathbf{X}_i^\top \left[\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}\right]^{-1} \mathbf{X}_i \right] \right\}. \quad (2.1)$$

Unfortunately, it is crucial to note that the form of the variance-covariance matrix in the observed data likelihood, namely $\mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$, makes it hard to derive analytical expressions for estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Besides, the numerical derivation of estimates based on the observed data likelihood runs into a variety of difficulties.

Recall that, from a Bayesian perspective, the estimation of $\mathbf{\Lambda}$ is based on the posterior distribution of $\mathbf{\Lambda}$ which itself is obtained by combining a prior on/about

$\mathbf{\Lambda}$ with the likelihood. More precisely, if $\boldsymbol{p}(\mathbf{\Lambda})$ is our prior on $\mathbf{\Lambda}$, then we need to derive the posterior $\boldsymbol{p}(\mathbf{\Lambda}|\mathbf{X})$ via Bayes rule

$$\boldsymbol{p}(\mathbf{\Lambda}|\mathbf{X}) = \frac{\boldsymbol{p}(\mathbf{\Lambda})\boldsymbol{L}(\mathbf{\Lambda};\mathbf{X})}{\boldsymbol{p}(\mathbf{X})} \propto \boldsymbol{p}(\mathbf{\Lambda})\boldsymbol{L}(\mathbf{\Lambda};\mathbf{X}).$$

The Bayesian estimate of $\mathbf{\Lambda}$ is the conditional (posterior) expectation of $\mathbf{\Lambda}$ given the data, i.e.,

$$\hat{\mathbf{\Lambda}}_{\mathsf{Bayes}} = \mathbb{E}[\mathbf{\Lambda}|\mathbf{X}] = \int \mathbf{\Lambda}\boldsymbol{p}(\mathbf{\Lambda}|\mathbf{X})d\mathbf{\Lambda}. \tag{2.2}$$

Unfortunately, the estimation of $\mathbf{\Lambda}$ via (2.2) runs into even more problems than with (2.1). Indeed, because of the complicated expression of the variance-covariance matrix, no closed-form expression can be derived for $\boldsymbol{p}(\mathbf{\Lambda}|\mathbf{X})$, making it impossible to compute the needed expectations. This reasoning for $\mathbf{\Lambda}$ is valid for $\boldsymbol{\Psi}$, therefore valid for our parameter collection $\boldsymbol{\theta} = (\mathbf{\Lambda}, \boldsymbol{\Psi})$. In other words, the expression of the likelihood $L(\boldsymbol{\theta};\mathbf{X})$ complicates the estimation from both the frequentist (MLE) and Bayesian perspective. Fortunately, it turns out that the complete-data likelihood makes it possible to circumvent some of the above problems. Indeed, methods such as the Expectation-Maximization (EM) algorithm and its Bayesian counterpart, the Imputation-Posterior algorithm, use the expression of the complete-data likelihood to derive parameter estimates. Details of the EM algorithm for Maximum Likelihood estimation can be found in the standard literature. As for the Imputation-Posterior algorithm, the idea is very similar to the EM algorithm idea: *while the E-step of the EM algorithm uses expected values of the factor scores to compute an expected likelihood, the I-step of the Imputation-Posterior algorithm draws $n$ realizations of the factor scores to form the conditional density of the parameters given the factor scores. In the same way, while the M-step of the EM algorithm computes the current estimates of the parameters based on the current expected likelihood, the P-step of the IP algorithm draws the current set of parameter values based on the current conditional distribution of the parameters given the factor scores.* The IP algorithm so defined is sometimes referred to as the Data Augmentation algorithm for the obvious reason that the imputed factor scores can be viewed as data augmented to simplify the estimation task.

The key to the derivation of the data augmentation algorithm for factor analysis is that, instead of working with the intractable expression of the observed data posterior

$$\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{X}) \propto \boldsymbol{L}(\boldsymbol{\theta};\mathbf{X})\boldsymbol{p}(\boldsymbol{\theta}), \tag{2.3}$$

one resorts to the complete-data posterior

$$\boldsymbol{p}(\boldsymbol{\theta}, \boldsymbol{F}|\mathbf{X}) \propto \boldsymbol{L}(\boldsymbol{\theta};\mathbf{X}, \boldsymbol{F})\boldsymbol{p}(\boldsymbol{\theta}). \tag{2.4}$$

One of the greatest appeals of equation (2.4) lies in the fact that suitable choices of the prior density $\boldsymbol{p}(\boldsymbol{\theta})$ (such as conjugate priors) lead to nice tractable expressions of the conditional posterior needed to simulated the true posterior.

- The **I-step** consists in drawing samples from the conditional distribution of $\boldsymbol{F}$ given $\mathbf{X}$ and the current set of parameter values $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\Lambda}^{(t)}, \boldsymbol{\Psi}^{(t)})$.

$$\left[\boldsymbol{F}_i^{(t+1)}|\mathbf{X}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}\right] \sim \mathcal{N}_q\left([\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}_i, [\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\right).$$

- The **P-step** combines the prior density $\boldsymbol{p}(\boldsymbol{\theta})$ with the expression for the complete-data likelihood $\boldsymbol{L}(\boldsymbol{\theta}; \mathbf{X}, \boldsymbol{F})$ to derive the corresponding full posteriors $\boldsymbol{p}(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{F})$ needed in the Gibbs sampling, namely:

$$\boldsymbol{\Psi}^{(t+1)} \sim \boldsymbol{p}(\boldsymbol{\Psi}|\boldsymbol{F}^{(t+1)}, \boldsymbol{\Lambda}^{(t)}, \mathbf{X}) \text{ and } \boldsymbol{\Lambda}^{(t+1)} \sim \boldsymbol{p}(\boldsymbol{\Lambda}|\boldsymbol{F}^{(t+1)}, \boldsymbol{\Psi}^{(t+1)}, \mathbf{X}).$$

The I-step and the P-step are repeated until a large number of draws is collected to form the sample path. The theoretical study of the convergence of the IP algorithm and the properties of the estimates is beyond the scope of this paper. Suffices it to note that after throwing away the initial draws (many thousands of them), the remaining draws obtained from the IP algorithm are used for estimation and inference about both $\boldsymbol{\theta}$ and $\boldsymbol{F}$. Specifically, we have

$$\hat{\boldsymbol{\Lambda}}_{\mathsf{IP}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Lambda}^{(t)} \text{ and } \hat{\boldsymbol{\Psi}}_{\mathsf{IP}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Psi}^{(t)}.$$

Now, the I-step in the above algorithm does not need the prior distributions of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. However, the P-step cannot be done without the prior. It is therefore important to provide some guidance as to how the prior is specified. The complete-data likelihood (1.7) belongs to the regular exponential family of distributions, and therefore allows a straightforward derivation of conjugate priors. While this choice is made for mathematical convenience, it also turns out to be the only computationally viable choice in this context. Martin and McDonald (1981) and Ihara and Kano (1995) have shown that the use of standard improper reference priors leads to the Bayesian analogue of what is known in factor analysis as *Heywood cases*. In the classical maximum likelihood estimation of the FA model, it is often convenient to minimise the negative log-likelihood or some extensions of it. It often happens that the objective function used has a relative minimum corresponding to negative values for some variances. Such solutions are clearly inadmissible and are referred to as improper solutions or Heywood cases. Treated as a function of the variance parameter, the negative likelihood of the FA model is bounded below away from zero as $\boldsymbol{\Psi}_i^2$ tends to zero. For the above reasons, this paper will use conjugate priors in various forms.

### 2.1 Prior specification and posterior derivation

Treating equation (1.7) as a function of $\boldsymbol{\Psi}$, one can write the likelihood function as

$$\boldsymbol{L}(\boldsymbol{\Psi}^{-1}) \propto |\boldsymbol{\Psi}^{-1}|^{n/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}^{-1}\boldsymbol{W})\right], \tag{2.5}$$

where $\boldsymbol{W} = \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\Lambda}\boldsymbol{F}_i)(\mathbf{X}_i - \boldsymbol{\Lambda}\boldsymbol{F}_i)^\top = (\mathbf{X} - \boldsymbol{F}\boldsymbol{\Lambda}^\top)^\top(\mathbf{X} - \boldsymbol{F}\boldsymbol{\Lambda}^\top)$. The form of (2.5) suggests that a natural conjugate prior for $\boldsymbol{\Psi}^{-1}$ would be a Wishart distribution. However, since $\boldsymbol{\Psi}^{-1}$ is diagonal, (2.5) can be rewritten as

$$\boldsymbol{L}(\boldsymbol{\Psi}^{-1}) \propto \prod_{i=1}^{p} \left[\psi_i^{-2}\right]^{n/2} \exp\left[-\frac{1}{2}w_{ii}\psi_i^{-2}\right], \tag{2.6}$$

which has the form of a product of Gamma densities, suggesting the use of a product of Gamma prior densities $\boldsymbol{p}(\psi_i^{-2})$ for each $\psi_i^{-2}$. To write the likelihood as a function of $\boldsymbol{\Lambda}$,

$$\begin{aligned}(\mathbf{X} - \boldsymbol{F}\boldsymbol{\Lambda}^\top)^\top(\mathbf{X} - \boldsymbol{F}\boldsymbol{\Lambda}^\top) &= (\mathbf{X} - \boldsymbol{F}\hat{\boldsymbol{\Lambda}}^\top)^\top(\mathbf{X} - \boldsymbol{F}\hat{\boldsymbol{\Lambda}}^\top) \\ &\quad + (\boldsymbol{\Lambda}^\top - \hat{\boldsymbol{\Lambda}}^\top)^\top\boldsymbol{F}^\top\boldsymbol{F}(\boldsymbol{\Lambda}^\top - \hat{\boldsymbol{\Lambda}}^\top).\end{aligned}$$

Since $(\mathbf{X} - \boldsymbol{F}\hat{\boldsymbol{\Lambda}}^\top)^\top(\mathbf{X} - \boldsymbol{F}\hat{\boldsymbol{\Lambda}}^\top)$ does not depend on $\boldsymbol{\Lambda}$, one can then write

$$\boldsymbol{L}(\boldsymbol{\Lambda}) \propto \exp\left[-\frac{1}{2}\mathrm{tr}\boldsymbol{\Psi}^{-1}(\boldsymbol{\Lambda}^\top - \hat{\boldsymbol{\Lambda}}^\top)^\top(\boldsymbol{F}^\top\boldsymbol{F})(\boldsymbol{\Lambda}^\top - \hat{\boldsymbol{\Lambda}}^\top)\right]. \tag{2.7}$$

A slightly elaborate algebraic manipulation of (2.7) suggests that a Gaussian distribution would be a natural conjugate prior for each row $\boldsymbol{\Lambda}_i$ of $\boldsymbol{\Lambda}$. See Fokoué (2004) for more details. From the form suggested by the expression of the likelihood function, the prior density can be specified either with $\boldsymbol{\Lambda}$ dependent on $\boldsymbol{\Psi}$ as

$$\boldsymbol{p}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \boldsymbol{p}(\boldsymbol{\Psi})\boldsymbol{p}(\Delta)\boldsymbol{p}(\boldsymbol{\Lambda} \,|\, \boldsymbol{\Psi}, \Delta), \tag{2.8}$$

or simply with $\boldsymbol{\Lambda}$ independent of $\boldsymbol{\Psi}$, ie

$$\boldsymbol{p}(\boldsymbol{\Lambda}, \boldsymbol{\Psi}) = \boldsymbol{p}(\boldsymbol{\Psi})\boldsymbol{p}(\Delta)\boldsymbol{p}(\boldsymbol{\Lambda} \,|\, \Delta). \tag{2.9}$$

It turns out in practice that both specifications perform equally well. For simplicity, the prior as defined by (2.9) will be used throughout this paper. The matrix $\Delta$ in the above prior specifications is the matrix of hyperparameters.

As mentioned earlier, the Wishart distribution for $\boldsymbol{\Psi}^{-1}$ reduces to a product of Gamma distributions because of the diagonality of $\boldsymbol{\Psi}$. In other words, with $\psi_i^{-2} \sim \mathsf{Ga}(\alpha/2, \tau/2)$, the prior density for $\boldsymbol{\Psi}^{-1}$ becomes

$$\boldsymbol{p}(\psi^{-1}|\alpha,\tau) = \prod_{i=1}^{p} \boldsymbol{p}(\psi_i^{-2}|\alpha,\tau) \propto \prod_{i=1}^{p} [\psi_i^{-2}]^{\frac{1}{2}\alpha-1} \exp\left[-\frac{1}{2}\tau\psi_i^{-2}\right]. \qquad (2.10)$$

If (2.10) and (2.6) are combined, and $w_{ii}$ is used to denote the $i$th diagonal entry of the matrix $\boldsymbol{W}$, then it is easy to derive a Gamma full conditional distribution for each $\psi_i^{-2}$, that is,

$$[\psi_i^{-2}|\cdots] \sim \mathsf{Ga}\left(\frac{n+\alpha}{2}, \frac{w_{ii}+\tau}{2}\right), \quad \text{for} \quad i = 1, \cdots, p$$

The first assumption made for the distribution of $\boldsymbol{\Lambda}$ is that its rows are *a priori* independent. From the previous section, conjugacy suggests that each row $\boldsymbol{\Lambda}_i$ is normally distributed. Specifically, a zero mean Gaussian prior with covariance matrix $\Delta_0$ will be used for each $\boldsymbol{\Lambda}_i$, ie

$$\boldsymbol{p}(\boldsymbol{\Lambda}_i \,|\, \Delta_0) = (2\pi)^{-q/2}|\Delta_0|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\Lambda}_i^\top \Delta_0^{-1} \boldsymbol{\Lambda}_i\right] \qquad (2.11)$$

where $\Delta_0 \in I\!R^{q\times q}$ is the prior covariance matrix common to all the rows $\boldsymbol{\Lambda}_i$ of $\boldsymbol{\Lambda}$. With a little bit of algebra, the full conditional posterior for each row $\boldsymbol{\Lambda}_i$ of $\boldsymbol{\Lambda}$ is found to be Gaussian with mean $\mathsf{m}_i$ and covariance matrix $\mathsf{K}_i$ given by

$$\mathsf{K}_i^{-1} = \Delta_0^{-1} + \psi_i^{-2}(\boldsymbol{F}^\top\boldsymbol{F}), \qquad \mathsf{m}_i = \left[\psi_i^2\Delta_0^{-1} + (\boldsymbol{F}^\top\boldsymbol{F})\right]^{-1} \boldsymbol{F}^\top \mathbf{X}^i \qquad (2.12)$$

where $\mathbf{X}^i$ is the $i$-th column of the data matrix $\mathbf{X}$.

## 3. Simple Factor Structure and Interpretability

The FA model is inherently a non-identified model: for a given set of data, there exists an infinity of orthogonal transformations of the matrix of factor loadings that would produce the same covariance structure. To see this more clearly, let us assume $q > 1$, and let $\boldsymbol{H}$ be any $q \times q$ orthogonal matrix, so that $\boldsymbol{H}\boldsymbol{H}^\top = \boldsymbol{H}^\top\boldsymbol{H} = \mathbf{I}_q$. Equation (1.1) can be written

$$\mathbf{X} = \boldsymbol{\Lambda}\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}\boldsymbol{H}\boldsymbol{H}^\top\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}^*\boldsymbol{F}^* + \boldsymbol{\epsilon} \qquad (3.1)$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{H}$ and $\boldsymbol{F}^* = \boldsymbol{H}^\top\boldsymbol{F}$. It is easy to see that $\mathbb{E}(\boldsymbol{F}^*) = \boldsymbol{H}^\top\mathbb{E}(\boldsymbol{F}) = \mathsf{0}$, and that $\mathrm{cov}(\boldsymbol{F}^*) = \boldsymbol{H}^\top\mathrm{cov}(\boldsymbol{F})\boldsymbol{H} = \boldsymbol{H}^\top\boldsymbol{H} = \mathbf{I}_q$. In other words, the factors $\boldsymbol{F}$ and $\boldsymbol{F}^* = \boldsymbol{H}^\top\boldsymbol{F}$ have the same statistical properties. Looking at equations

(1.1) and (3.1), it is therefore impossible on the basis of observations on $\mathbf{X}$, to distinguish the matrices of factor loadings $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$. Moreover, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{H}\boldsymbol{H}^\top\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} = (\boldsymbol{\Lambda}^*)(\boldsymbol{\Lambda}^*)^\top + \boldsymbol{\Psi}$, which means that although different in general, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^*$ both generate the same covariance matrix $\Sigma$, and therefore the same representation of the data. One can therefore obtain an infinite number of equivalent matrices of factor loadings by simply applying successive orthogonal transformations to an initial one.

The search for a unique and easily interpretable factor solution is generally addressed by imposing constraints to identify a unique set of model parameters. In this vein, one of the most widely used approaches to identifiability and simple structure search consists in setting some of the elements of the matrix of factor loadings to some pre-assigned values, usually *zero*. Besides the pre-assignment approach just described, rotation techniques like Kaiser's varimax rotation of the initial solution are used to find a simple structure, but details of such are not provided in this paper. Equation (3.2) provides an example of a constrained structure.

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{q-1,1} & \lambda_{q-1,2} & \lambda_{q-1,3} & \cdots & \lambda_{q-1,q-1} & 0 \\ \lambda_{q,1} & \lambda_{q,2} & \lambda_{q,3} & \cdots & \lambda_{q,q-1} & \lambda_{q,q} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{p,1} & \lambda_{p,2} & \lambda_{p,3} & \cdots & \lambda_{p,q-1} & \lambda_{p,q} \end{pmatrix}. \tag{3.2}$$

Lopes and West (1999) have satisfactorily used this particular constrained structure of the factor loadings matrix in their application of factor analysis to portfolio management. This approach is widely used by psychometricians and other factor analysts who greatly value interpretability. It is important to note that this approach requires the need for the investigator to make use of his/her subjectivity and prior knowledge about the problem under consideration.

It turns out that the implementation of this constrained FA model is rather straightforward in the Bayesian posterior simulation context. In fact, such a restriction requires only a very minor modification in the derivation of the full conditional distribution of $\boldsymbol{\Lambda}$.

In this case, a univariate Gaussian prior is assumed for each of the non-preassigned $\lambda_{ij}$, that is

$$\lambda_{ij} \sim \mathcal{N}(0, \delta_0^{-1})$$

$\boldsymbol{F}_{(i)} \in I\!R^{n\times i}$ is used to denote the $n \times i$ matrix containing the first $i$ columns of $\boldsymbol{F}$. The mean vector and covariance matrix of the full conditional distribution of $\boldsymbol{\Lambda}_i$ for the first $q$ rows ($i = 1, \cdots, q$) are determined as follows:

$$\mathsf{K}_i^{-1} = \delta_0 \mathbf{I}_i + \psi_i^{-2}(\boldsymbol{F}_{(i)}^\top \boldsymbol{F}_{(i)}) \quad \text{and} \quad \mathsf{m}_i = \left[\psi_i^2 \delta_0 \mathbf{I}_i + (\boldsymbol{F}_{(i)}^\top \boldsymbol{F}_{(i)})\right]^{-1} \boldsymbol{F}_{(i)}^\top \mathbf{X}^i \tag{3.3}$$

For $i = (q+1), \cdots, p$, one gets

$$\mathsf{K}_i^{-1} = \delta_0 \mathbf{I}_q + \psi_i^{-2}(\boldsymbol{F}^\top \boldsymbol{F}) \quad \text{and} \quad \mathsf{m}_i = \left[\psi_i^2 \delta_0 \mathbf{I}_q + (\boldsymbol{F}^\top \boldsymbol{F})\right]^{-1} \boldsymbol{F}^\top \mathbf{X}^i \tag{3.4}$$

MacKay (1992) and Neal (1996) first proposed and applied the idea of *Automatic Relevance Determination (ARD)* in their Bayesian Analysis of Neural Networks and related models. The ARD idea has been adapted by Tipping (2001) in his derivation of the Relevance Vector Machine as a tool for obtaining a sparse function representation despite the use of a Gaussian prior. This paper adapts the ARD prior idea to Bayesian Factor Analysis. The key idea is that the task of searching for a simple factor structure that is easy and straightforward to interpret is in a sense equivalent to searching for a *sparse representation* of the factor model. With the embedded *sparsity* introduced through the prior, the proposed approach produces an estimate of $\boldsymbol{\Lambda}$ that has a simple structure with many *zeros*, making interpretation easy and straightforward. Following from Tipping (2001)'s development of the *Relevance Vector Machine*, sparsity pressure in the space of $\boldsymbol{\Lambda}$ may be achieved by specifying an independent Gaussian prior for each element $\lambda_{ij}$ of $\boldsymbol{\Lambda}$. Conditional on a precision hyperparameter $\delta_{ij}$, the prior density for each $\lambda_{ij}$ is given by

$$\boldsymbol{p}(\lambda_{ij} \,|\, \delta_{ij}) = \mathcal{N}(\lambda_{ij} \,|\, 0, \delta_{ij}^{-1}) \tag{3.5}$$

Each row $\boldsymbol{\Lambda}_i$ of $\boldsymbol{\Lambda}$ therefore has conditional prior density

$$\boldsymbol{p}(\boldsymbol{\Lambda}_i \,|\, \Delta_i) = \mathcal{N}_q(\boldsymbol{\Lambda}_i \,|\, 0, \Delta_i) \tag{3.6}$$

where $\Delta_i = \text{diag}(\delta_{i1}^{-1}, \delta_{i2}^{-1}, \cdots, \delta_{iq}^{-1})$. For each $\delta_{ij}$, the following *Gamma* prior will be used:

$$\boldsymbol{p}(\delta_{ij} \,|\, a, b) = \mathsf{Ga}(\delta_{ij} \,|\, a, b) \tag{3.7}$$

As shown in Tipping (2001), the use of a *Gamma* prior for each $\delta_{ij}$ leads to a marginal prior for $\boldsymbol{\Lambda}_i$ that is a product of independent Student's $t$-distributions. With the degrees of freedom being small, the product of independent Student-t distributions implies that the distribution of $\boldsymbol{\Lambda}_i$ is concentrated along the axes or at the origin, which is precisely the type of structure perceived as simple and interpretable in factor analysis. This is illustrated using a two-dimensional case to show that with such a prior the probability mass is concentrated both at the origin and along the "spines" where one of the coefficients $\lambda_{ij}$ is zero.

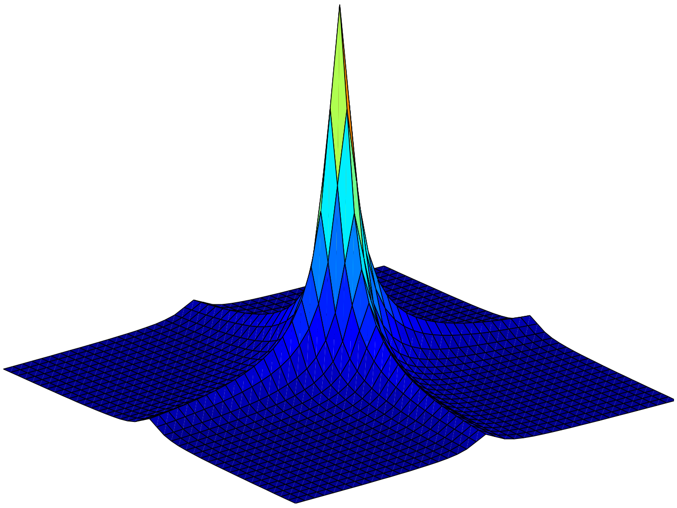

Figure 1: The 2-dimensional marginal prior for a row $\mathbf{\Lambda}_i$

As Figure 1 shows, it is "surprisingly" possible to achieve a *sparse representation* using a Gaussian prior. From a practical standpoint however, a bit of thresholding might be needed, and this is done in this case by setting to *zero* (declared *irrelevant*) factor loadings $\lambda_{ij}$ whose precision tends to "infinity". (ie gets too large beyond a pre-specified threshold). This approach to sparsity has been used extensively, and has produced many satisfactory results in a variety of settings. The present use of it is the first such adaptation to the factor analysis context, and has produced satisfactory results as well.

It is easy to see that the full conditional posterior for each row $\mathbf{\Lambda}_i$ of $\mathbf{\Lambda}$ is found to be Gaussian with mean $\mathsf{m}_i$ and covariance matrix $\mathsf{K}_i$ given by

$$\mathsf{K}_i^{-1} = \Delta_i^{-1} + \psi_i^{-2}(\boldsymbol{F}^\top \boldsymbol{F}), \qquad \mathsf{m}_i = \left[\psi_i^2 \Delta_i^{-1} + (\boldsymbol{F}^\top \boldsymbol{F})\right]^{-1} \boldsymbol{F}^\top \mathbf{X}^i \tag{3.8}$$

where $\mathbf{X}^i$ is the $i$-th column of the data matrix $\mathbf{X}$.

The conjugate Gamma prior for $\delta_{ij}$ makes the derivation of its posterior straightforward.

$$\boldsymbol{p}(\Delta_i \,|\, \cdots) \propto \boldsymbol{p}(\Delta_i)\boldsymbol{p}(\mathbf{\Lambda}_i \,|\, \Delta_i)$$

Since both the $\lambda_{ij}$'s and the $\delta_{ij}$'s are assumed to be *a priori* independent,

$$\boldsymbol{p}(\delta_{ij} \,|\, \cdots) \propto \mathcal{N}(\lambda_{ij} \,|\, 0, \delta_{ij})\mathsf{Ga}(\delta_{ij} \,|\, a, b)$$

As a result,

$$[\delta_{ij} \mid \cdots] \sim \mathsf{Ga}\left(a + \frac{1}{2}, b + \frac{1}{2}\lambda_{ij}^2\right).$$

In a sense, this can be viewed as the Bayesian alternative to Kaiser's varimax rotation, since the sparse representation implies a very high communality. However, unlike the Kaiser's varimax that requires two stages and some subjectivity about when to stop rotating, the proposed method achieves both the esimation and the simple structure simultaneously.

From all above, a Data Augmentation scheme for Factor Analysis can be summarized as:

Data Augmentation for Factor Analysis

- **I-step -**
$$[\boldsymbol{F}|\mathbf{X}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}] \sim \mathcal{N}_q\left([\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}, [\mathbf{I}_q + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}]^{-1}\right)$$
- **P-step -**

$$\begin{aligned}
[\boldsymbol{\Lambda}_i \mid \cdots] &\sim \mathcal{N}_q(\mathsf{m}_i, \mathsf{K}_i), \quad i = 1, \cdots, p \\
[\delta_{ij} \mid \cdots] &\sim \mathsf{Ga}\left(a + \frac{1}{2}, b + \frac{1}{2}\lambda_{ij}^2\right), \quad i = 1, \cdots, p \text{ and } j = 1, \cdots, q \\
\left[\psi_i^{-2} \mid \cdots\right] &\sim \mathsf{Ga}\left(\frac{n+\alpha}{2}, \frac{w_{ii}+\tau}{2}\right), \quad i = 1, \cdots, p
\end{aligned}$$

## 4. Estimating the Intrinsic Dimensionality

While there are many cases in practice where the number of factors $q$ is known and/or fixed, as it has been assumed so far, it must be said that this value is very often unknown in real-life applications, and the study of the FA model therefore needs to address its uncertainty. At the root of model determination in Factor Analysis lies the difficult issue of finding and/or defining principled methods to decide what makes a particular factor important. In fact, for FA, this difficult problem has been one of the burning issues over the years, captivating the interests of researchers from both the likelihood-based and Bayesian perspectives. Some of the references on this topic include Krzanowski and Marriott (1994), Krzanowski and Marriott (1995) on the frequentist side, and Press (1972), Press and Shigemasu (1998), Lopes and West (1999) and Fokoué and Titterington (2003) on the Bayesian side.

A crude upper bound for the number of factors is $p$, the original dimensionality of the input space, while a crude lower bound is simply 1, the simplest factor model one can have. These crude bounds are clearly not very helpful, it is interesting to derive more useful ones. Recall that the ideal is to find a factor structure that is: (a) *unique*; (b) *simple* and (c) *intrinsic.* The marginal distribution of the observed random variable $X \in \mathbb{R}^p$ has covariance $\Sigma = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$. Since $\Sigma$ is symmetric, it has $p(p+1)/2$ free parameters. If a sparse representation or a structure constrained as in (3.2) has $q(q-1)/2$ zero values, then to guarantee a unique solution, it is necessary to determine $q$ such that $p(q+1) - \frac{1}{2}q(q-1) \le \frac{1}{2}p(p+1)$, which means

$$(p+q) \le (p-q)^2 \tag{4.1}$$

From (4.1) an upper bound on the number of factors to be included in a model is given by

$$q \le \frac{1}{2}(2p+1-\sqrt{8p+1}). \tag{4.2}$$

It is important to note that there are situations where solutions satisfying constraint (4.1) might not provide an adequate fit for the data. In fact, given a data set, a fundamental question (without an obvious answer) is whether there exists a matrix of factor loadings $\mathbf{\Lambda}$ such that the model in equation (1.1) adequately fits the data. An exploration of this issue and many other related topics of FA can be found in such references as Bartholomew (1987), Everitt (1984) and Press (1972) amongst others.

Many of the most widely used methods are based on various functions of the eigenvalues of the sample correlation matrix. While such methods produce satisfactory results, the fact of focusing only on the eigenvalues could lead to the neglect of vital information: almost all the criteria used to decide on the number of factors to retain are essentially ad hoc (eigenvalues less than 1) and often subjective (elbow of the screeplot) criteria that in some special cases would either overestimate or underestimate the adequate number of factors. For instance, if one variable is virtually independent of all the rest, it will appear as a separate component with variance slightly less than 1, but there is no reason to suppose that such a variable is uninformative. Thus, while this method may provide rough estimates of the number of factors, there is a clear need for more principled and objective methods for estimating the intrinsic dimensionality of factor analytic models.

With the normality assumption for the manifest variable $\mathbf{X}$, maximum likelihood via the EM algorithm is straightforward in factor analysis. On the other hand, the goodness-of-fit of the resulting $q$-factor model can be judged using a classical likelihood ratio test, with the null hypothesis stating the covariance matrix of $\mathbf{X}$ has the structure $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$, and the alternative saying the covariance matrix is unconstrained. Under the normal assumption, it is easy to see that the test statistic for the test is

$$\boldsymbol{\omega} = n(\mathrm{tr}(\hat{\mathbf{\Sigma}}^{-1}\boldsymbol{S}) - \log|\hat{\mathbf{\Sigma}}^{-1}\boldsymbol{S}| - p), \tag{4.3}$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top + \hat{\mathbf{\Psi}}$ is the estimate of $\Sigma$, and $S$ is the sample covariance matrix. A standard result in the literature shows that if $\mathbf{\Psi} > 0$, then $\boldsymbol{\omega}$ is asymptotically $\chi^2$ distributed with $\nu = \frac{1}{2}\left[(p-q)^2 - (p+q)\right]$ degrees of freedom under the null hypothesis. An alternative setting proposed by Bartlett (1954) suggests to replace $n$ in (4.3) by $n-1-\frac{1}{6}(2p+5)-\frac{2}{3}q$. It must be said that the value of $\nu$ used above presupposes that one has efficiently fitted the model, and therefore that instead of the $p(q+1)$ parameters of the unrestricted FA model, only $p(q+1)-\frac{1}{2}q(q-1)$ parameters have to be estimated.

Likelihood-based approaches to the determination of intrinsic factor model dimensionality mainly consist in sequentially applying a series of likelihood ratio tests. In practice, one starts with $q = 1$ (single factor model), then fits successive values and tests the goodness-of-fit, until the test produces a non-significant result indicating in a sense that the fit of the model is adequate. However, while this method appears as an objective procedure for estimating $q$, it is not strictly valid as a hypothesis test as argued by Krzanowski and Marriott (1995), since *no adjustment is made to the significance level to allow for its sequential nature.* On the other hand, the fact of having a non-significant p-value cannot be taken to indicate that the optimum value of $q$ has been found, since large values of $q$ correspond to more parameters and therefore better fits, obviously at the expense of more complex models and risks of overfitting. For the “best” model to be determined, there needs to be a trade-off between the number of parameters and the goodness-of-fit. In this likelihood-based framework, one way to determine the “best” model is to use Akaike’s Information Criterion, which consists of selecting the model that minimises AIC as defined in (4.4).

$$\mathsf{AIC} = -2\log(\mathsf{maximised\ likelihood}) + 2(\mathsf{number\ of\ parameters\ fitted}). \tag{4.4}$$

In the factor analysis context, the above criterion (4.4) is equivalent to choosing $q$ that minimises $\omega - 2\nu$, as suggested by Akaike (1987), where $\omega$ and $\nu$ are respectively the test statistic and the number of degrees of freedom. It has been noticed in practice that AIC tends to overfit models. In the analysis of mixtures for instance, AIC tends to overestimate the correct number of components. The Bayesian Information Criterion (BIC)of equation (4.5) is often used as an alternative to AIC.

$$\mathsf{BIC} = -2\log(\mathsf{maximised\ likelihood}) + \log n(\mathsf{number\ of\ parameters\ fitted}) \tag{4.5}$$

The reason why BIC performs better than AIC can be explained simply as follows: the penalty term of BIC penalises complex models more heavily than AIC, whose penalty term does not depend on the sample size. BIC therefore reduces the tendency of the AIC criterion to overfit models.

The determination of the optimum number of factors has been studied before in the Bayesian setting. Press and Shigemasu (1998) approached the problem by deriving an "approximate" posterior probability mass function $\boldsymbol{P}(q \mid \mathbf{X})$ for the number of factors. A potential drawback to this approach lies in the approximate nature of the posterior mass function: it would seem that in some very simply problems as shown in the next section, the approximation error can lead to rather poor estimations of the number of factors. The approach proposed in this paper avoids this approximation pitfall by constructing an ergodic Markov chain whose final sample path provided ingredients for the exact probability mass function $\boldsymbol{P}(q \mid \mathbf{X})$.

### 4.1 Elements of stochastic model selection for FA

The approach used here is based on the construction of a Markov chain having the posterior distribution of the complete collection of all the unknowns (parameters and $q$) as its equilibrium distribution. From a Bayesian perspective, Green (1995)'s Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is one such algorithm. Lopes and West (1999) applied an adaptation of RJMCMC to the factor analysis model with an unknown number of common factors, and obtained good results on both synthetic and real-life problems. More recently, Stephens (2000), using ideas from stochastic geometry and spatial statistics, developed an alternative to RJMCMC, based on the simulation of a continuous-time birth-and-death Markov marked point process. Stephens (2000) applied the derived Birth-and-Death MCMC (BDMCMC) method to mixtures of univariate and bivariate Gaussians with unknown numbers of components, and obtained promising results. BDMCMC was later successfully adapted by Fokoué and Titterington (2003) in the study of Mixtures of Factor Analyzers. Despite the fact that RJMCMC is based on a discrete-time Markov process while BDMCMC is based on a continuous time Markov process, the two methods are essentially equivalent in that they both successfully construct ergodic Markov chains in spaces of varying dimensions. In fact, BDMCMC can be thought of as a limit of RJMCMC. However, for practical reasons and to a certain extent for computational convenience, this paper adopts an approach closer to BDMCMC.

The central idea behind this approach is to view and treat each parameter that directly affects the dimensionality of the model as a point in the parameter space, and adapt the methodology of point process simulation to help construct a Markov chain with the posterior distribution of the parameters as its equilibrium distribution.

Geometrically speaking, the columns of $\mathbf{\Lambda}$ can be viewed as defining the axes of the lower-dimensional latent space (coordinate system) of factors. Since a rotation is a non-singular orthogonal transformation, and a permutation of columns is particular type of rotation, the factor solution is therefore invariant to permutations of axes. Equivalently, it can be said that FA has a posterior distribution that is invariant to permutations of the order of their parameters. From a stochastic simulation perspective, the collection of parameters can therefore be viewed as a random configuration or point process. This complete collection of our model parameters is now given by $\boldsymbol{\theta} = \{q, \mathbf{\Lambda}, \mathbf{\Psi}\}$. If one assumes that $q$ is unknown a priori, the aim from a posterior simulation perspective now extends to the construction of an ergodic Markov chain with the joint posterior distribution $\boldsymbol{p}(q, \mathbf{\Lambda}, \mathbf{\Psi}|\mathbf{X})$ as its equilibrium distribution. In one of the previous sections, Data Augmentation was used to construct a Markov chain with $\boldsymbol{p}(\mathbf{\Lambda}, \mathbf{\Psi}|q, \mathbf{X})$ as its equilibrium distribution. With $q$ unknown, there is the need to accommodate the new *counting* random variable $q$. Intuitively, the overall sampling scheme takes on a Gibbs sampler-like form, with each iteration consisting of two steps:

$$\begin{array}{ll} \textsf{Step 1: Birth-and-death:} & q^{(t+1)} \sim \boldsymbol{p}(q|\mathbf{\Lambda}^{(t)}, \mathbf{\Psi}^{(t)}, \mathbf{X}) \\ \textsf{Step 2: Gibbs sampling:} & (\mathbf{\Lambda}^{(t+1)}, \mathbf{\Psi}^{(t+1)}) \sim \boldsymbol{p}(\mathbf{\Lambda}, \mathbf{\Psi}|q^{(t+1)}, \mathbf{X}) \end{array}$$

In the above scheme, Step 1 allows us to draw a new value of $q = q^{(t+1)}$ by simulating a birth-and-death Markov point process, the main difference with a classical algorithm of this type being that the dimension of the parameter vector is allowed to vary at each iteration. Step 2 draws a new set of model parameters via Data Augmentation, using the value of $q$ obtained from the run of the birth-and-death process.

The simulation of the type of birth-and-death process used in this paper has been extensively studied and applied in recent years, and the reader is referred to references like Stoyan *et al.* (1995) and Barndorff-Nielsen *et al.* (1999) for comprehensive coverage of applications of such sampling schemes in stochastic geometry and spatial statistics. Baddeley (1994) and van Lieshout (1994) also provide very useful insights into other aspects of such sampling schemes. Stephens (2000) provides a detailed account of his application of BDMCMC to mixtures.

### 4.2 Birth-and-death point process for factor analysis

Let $\mathsf{M}^{(t)} = \{q^{(t)}, \mathbf{\Lambda}_1^{(t)}, \mathbf{\Lambda}_2^{(t)}, \cdots, \mathbf{\Lambda}_i^{(t)}, \cdots, \mathbf{\Lambda}_{q^{(t)}}^{(t)}, \mathbf{\Psi}^{(t)}\}$ define a random configuration at time $t$. Additionally, let $\mathsf{M}_{-i}^{(t)} = \mathsf{M}^{(t)} \backslash \{\mathbf{\Lambda}_i^{(t)}\}$ be the random configuration $\mathsf{M}^{(t)}$ without $\mathbf{\Lambda}_i^{(t)}$. By virtue of the fact that $\boldsymbol{p}(q, \mathbf{\Lambda}, \mathbf{\Psi}|\mathbf{X}) \equiv \boldsymbol{p}(\mathsf{M}|\mathbf{X})$ is invariant under permutations of the $\mathbf{\Lambda}_i$'s, the sequence $\{\mathsf{M}^{(t)} : t > 0\}$ defines a point process.

**Note:** For notational simplicity, $\boldsymbol{h}(\mathsf{M})$ will be used to denote the posterior $\boldsymbol{p}(\mathsf{M} \,|\, \mathbf{X})$.

It turns out that one can efficiently construct the desired ergodic Markov chain by simulating a sampling scheme comprising a birth-and-death point process step and a Data Augmentation step, both jointly converging to $\boldsymbol{p}(q, \mathbf{\Lambda}, \mathbf{\Psi}|\mathbf{X})$ as the stationary distribution. The key idea behind the simulation of the birth-and-death process is that each birth increases the number of points in the configuration by one, while each death decreases this number by one. Furthermore, both the birth and the death processes are constructed in such a way that they are inverse operations to each other in the equilibrium state of the chain. One way to construct such a process is to define births and deaths as follows:

- Define a **birth** density $b(\mathsf{M}; \boldsymbol{\lambda})$ according to which new points are added to the current configuration of the point process.
- When the current configuration of the chain is $\mathsf{M} = \{\mathbf{\Lambda}_1, \mathbf{\Lambda}_2, \cdots\}$, each point $\mathbf{\Lambda}_i$ dies independently of the others as a Poisson process with rate $\zeta_i(\mathsf{M}) = d(\mathsf{M}; \mathbf{\Lambda}_i)$, where $d(\mathsf{M}; \boldsymbol{\lambda})$ is the **death** density function, so that the overall death rate is given by
$$\zeta(\mathsf{M}) = \sum \zeta_i(\mathsf{M}).$$

The general practice consists of imposing suitable constraints on the birth and death functions $b$ and $d$ to ensure that the process does not jump to an area with zero density.

For simplicity, one can restrict the process to cases where births are assumed to be occurring at an overall constant rate $\beta(\mathsf{M}) = \beta$. Such a simplification has the clear disadvantage that many different birth rates have to be tried empirically before the "appropriate" one is found. These trials can however be avoided by randomly perturbing the birth rate during the simulation.

With $\beta(\mathsf{M})$ and $\zeta(\mathsf{M})$ defined, the following general results (stated without proof) on Poisson processes will be used. These results are used to obtain the distribution of the time to the next event in the simulation of the birth-and-death process and the distribution of the next event.

**Theorem 1.** The birth and the death being independent Poisson processes, the time to the next event (birth or death) is *exponentially* distributed with mean $1/(\beta(\mathsf{M}) + \zeta(\mathsf{M}))$.

**Fact**

Since the overall rate of the birth-and-death process is equal to $\beta(\mathsf{M}) + \zeta(\mathsf{M})$, the next event will be a birth with probability $\beta(\mathsf{M})/(\beta(\mathsf{M}) + \zeta(\mathsf{M}))$, while the death of $\mathbf{\Lambda}_i$ will occur with probability $\zeta_i(\mathsf{M})/(\beta(\mathsf{M}) + \zeta(\mathsf{M}))$.

One is therefore in the presence of a continuous-time process since the time to the next event is a continuous random variable, and, by virtue of the *memorylessness* property of the exponential distribution, one has a continuous time Markov process. In order to simulate such a continuous time process, a fixed unit of time, $\rho$, is defined, and a discrete-time Markov chain $\{\mathsf{M}^{(\rho)}, \mathsf{M}^{(2\rho)}, \mathsf{M}^{(3\rho)}, \cdots\}$ is constructed, and used as an approximation to the continuous-time chain $\{\mathsf{M}^{(\rho+s)} : s > 0\}$. This simply means that, at each discrete iteration $(t = 1, \cdots, T)$, the birth-and-death process is run for a duration of $\rho$.

Preston (1976) stated sufficient conditions that the above densities $b$ and $d$ must satisfy for the above birth-and-death process to define an ergodic Markov chain with the desired equilibrium distribution. Preston (1976)'s work was later extended and applied by Ripley (1977), and recently adapted to the analysis of finite mixtures by Stephens (2000). The following theorem, which states the sufficient conditions that $b$ and $d$ must satisfy, is from Preston (1976) and Ripley (1977). A proof of its extended version as applied to finite mixtures can be found in Stephens (2000).

**Theorem 2.** If the birth density $b$ and the death density $d$ satisfy

$$(q+1)d(\mathsf{M} \cup \{\boldsymbol{\lambda}\}; \boldsymbol{\lambda})\boldsymbol{h}(\mathsf{M} \cup \{\boldsymbol{\lambda}\}) = \beta(\mathsf{M})b(\mathsf{M}; \boldsymbol{\lambda})\boldsymbol{h}(\mathsf{M}) \tag{4.6}$$

for all configurations $\mathsf{M}$ and all points $\boldsymbol{\lambda}$, then the birth-and-death process defined above has $\boldsymbol{p}(q, \mathbf{\Lambda}, \mathbf{\Psi}|\mathbf{X})$ as its stationary distribution.

**Remark:** In the above theorem, $\boldsymbol{h}(\mathsf{M} \cup \{\boldsymbol{\lambda}\})$ represents the posterior density of a configuration with $q+1$ points. Intuitively, equation (4.6) means that, under the equilibrium distribution $\boldsymbol{p}(\cdot|\mathbf{X})$, transitions from $\mathsf{M}$ into $\mathsf{M} \cup \{\boldsymbol{\lambda}\}$ are exactly matched by transitions from $\mathsf{M} \cup \{\boldsymbol{\lambda}\}$ into $\mathsf{M}$. From equation (4.6), it is easy to see that

$$d(\mathsf{M}; \boldsymbol{\lambda}) = b(\mathsf{M}; \boldsymbol{\lambda}) \left[\frac{\beta(\mathsf{M})}{(q+1)}\right] \left[\frac{\boldsymbol{h}(\mathsf{M})}{\boldsymbol{h}(\mathsf{M} \cup \{\boldsymbol{\lambda}\})}\right], \tag{4.7}$$

which is equivalent to

$$d(\mathsf{M}; \boldsymbol{\lambda}) = b(\mathsf{M}; \boldsymbol{\lambda}) \left[\frac{\beta(\mathsf{M})}{q}\right] \left[\frac{\boldsymbol{h}(\mathsf{M}\backslash\{\boldsymbol{\lambda}\})}{\boldsymbol{h}(\mathsf{M})}\right], \tag{4.8}$$

where $\mathsf{M}\backslash\{\boldsymbol{\lambda}\}$ represents the current configuration $\mathsf{M}$ less the element $\boldsymbol{\lambda}$. From (4.8), it is easy to see that the appropriate death rate for element $\mathbf{\Lambda}_i$ $(i = 1, \cdots, q)$

is given by

$$\zeta_i(\mathsf{M}) = \left[\frac{\beta}{q}\right]\left[\frac{b(\mathsf{M};\boldsymbol{\Lambda}_i)}{\boldsymbol{p}(\boldsymbol{\Lambda}_i)}\right]\left[\frac{\boldsymbol{L}(\mathsf{M}\backslash\boldsymbol{\Lambda}_i)}{\boldsymbol{L}(\mathsf{M})}\right]\left[\frac{\boldsymbol{p}(q-1)}{\boldsymbol{p}(q)}\right] \tag{4.9}$$

where $\boldsymbol{L}(\mathsf{M})$ is the likelihood associated with the current configuration $\mathsf{M}$. The prior for $q$ can be chosen to be either a $\mathsf{uniform}$ prior or a Poisson prior truncated at the right end by a predetermined value $q_{max}$, ie

$$\boldsymbol{p}(q|\nu) \propto \frac{\nu^q}{q!}\exp(-\nu) \quad \text{for} \quad q = 1, \cdots, q_{max} \tag{4.10}$$

If the birth density is chosen to be the prior density of a candidate element $\boldsymbol{\lambda}$ to be added to the current configuration, then $b(\mathsf{M};\boldsymbol{\lambda}) = \boldsymbol{p}(\boldsymbol{\lambda})$ and

$$\zeta_i(\mathsf{M}) = \left[\frac{\beta}{q}\right]\left[\frac{\boldsymbol{L}(\mathsf{M}\backslash\boldsymbol{\Lambda}_i)}{\boldsymbol{L}(\mathsf{M})}\right] \tag{4.11}$$

Based on all the above ingredients, a pseudocode of the birth-and-death process is

**Algorithm A**: Birth-and-Death Process for Factor Analysis

```
Repeat
  $\zeta_j := \left[\frac{\beta}{q}\right]\left[\frac{\boldsymbol{L}(\mathsf{M}\backslash\boldsymbol{\lambda}_j)}{\boldsymbol{L}(\mathsf{M})}\right]$ for $j = 1, \cdots, q$
  $\zeta := \sum_{j=1}^{q} \zeta_j$
  $s :=$ Exponential $\left(\frac{1}{\beta+\zeta}\right)$
  $t := t + s$
  birth := Bernoulli $\left(\frac{\beta}{\beta+\zeta}\right)$
  If birth $= 1$ /* It is a birth */
      $\boldsymbol{\lambda}^{\mathsf{new}} := b(\mathsf{M};\boldsymbol{\lambda})$
      $\mathsf{M} := \mathsf{M} \cup \{\boldsymbol{\lambda}^{\mathsf{new}}\}$
      $q := q + 1$
  Else /* It is a death */
      $\pi_j = \zeta_j/\zeta$ for $j = 1, \cdots, q$
      out := Multinomial $(\pi_1, \pi_2, \cdots, \pi_q)$
      $\mathsf{M} := \mathsf{M}\backslash\{\boldsymbol{\lambda}_{\mathsf{out}}\}$
      $q := q - 1$
```

```
        End;
    Until $(t \geq \rho)$
```

### 4.3 Maximum a posteriori estimate for $q$

The mode of $\boldsymbol{p}(q|\mathbf{X})$ provides the maximum a posteriori estimate for $q$, namely

$$q^{\mathsf{opt}} \quad = \quad \mathop{\mathsf{arg\,max}}_{q}\, \boldsymbol{p}(q|\mathbf{X}) \tag{4.12}$$

Once the Markov chain $\{\mathsf{M}^{(t)} = \{q^{(t)}, \boldsymbol{\Lambda}_1^{(t)}, \boldsymbol{\Lambda}_2^{(t)}, \cdots, \boldsymbol{\Lambda}_i^{(t)}, \cdots, \boldsymbol{\Lambda}_{q^{(t)}}^{(t)}, \boldsymbol{\Psi}^{(t)}\} : t = 1, \cdots, T\}$ has converged to the desired equilibrium distribution, the sequence $\{q^{(t)} : t = 1, \cdots, T\}$ is essentially a sequence of draws from the marginal distribution $\boldsymbol{p}(q|\mathbf{X})$. Inference for $q$ can be based on an estimate of this marginal posterior distribution obtained from the MCMC sample path as follows: Let $N_m$ be the number of time the birth and death process yielded $m$ as the number of factors after burn-in. Clearly,

$$N_m = \sum_{j=1}^{M} \mathbb{I}(q^{(t)} = m) \tag{4.13}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Rigorously,

$$\Pr[q = m|\mathbf{X}] = \lim_{M \to \infty} \frac{N_m}{M} \approx \frac{N_m}{M} \tag{4.14}$$

Using (4.12), (4.13) and (4.14), the maximum a posteriori estimate for $q$ is obtained by simply choosing the value of $q$ having the highest frequency in the sample path of the Markov chain.

$$\hat{q}^{\mathsf{opt}} \quad = \quad \mathop{\mathsf{arg\,max}}_{m}\, N_m \tag{4.15}$$

Since the sample path after burning is a realization from the true posterior $\boldsymbol{p}(q|\mathbf{X})$, the estimate provided by this method is a more accurate estimate of the "true" $q$. In this sense, this estimate may be seen as "better" than estimated obtained through an approximate posterior as proposed by Press and Shigemasu (1998).

## 5. Numerical Results

This section presents numerical performance of the proposed method on three problems. All the simulations are written in Matlab 6.5 Release 13 for Unix.

$\beta = (\sqrt{5}-1)/2 = 0.61803$ (golden ratio) is used as the overall constant birth rate throughout the computations.

### 5.1 Artificial dataset from Press and Shigemasu (1998)

The dimensionality of the input space here is $p = 10$, and the sample size is $n = 200$. All the datasets are generated using $\epsilon \sim \mathcal{N}(0, 0.36\,\mathbf{I}_p)$ and

$$\boldsymbol{\Lambda}^\top = \begin{pmatrix} .8 & .8 & .8 & .8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .8 & .8 & .8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .8 & .8 & .8 \end{pmatrix}.$$

The number of factors is known to be $q = 3$, and the aim here is to compare the performance of different methods on the task of estimating $q$.

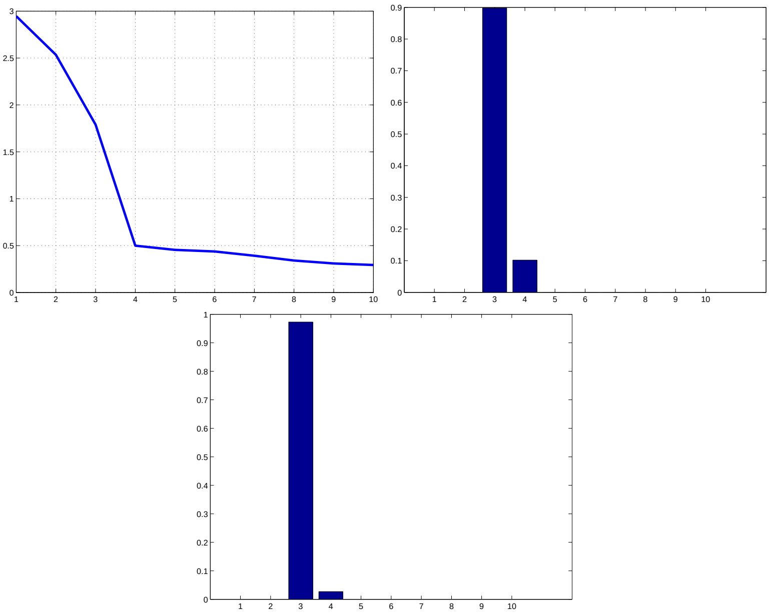

Figure 2: (Left) Screeplot (Center) Histogram of the number of factors when $q^{(0)} = 1$. (Right) Histogram of the number of factors when $q^{(0)} = 5$.

### Results from the screeplot method

The screeplot method is usually the quickest and easiest way to obtain a rough ad-hoc estimate of the number of factors. For this problem, the elbow of the screeplot (see figure 2-[left]) seems to be suggesting $q = 4$ as the "appropriate" number of factors, which in this case is the wrong answer.

### Results from the stochastic simulation method

The burn-in here is $T_o = 2500$, and the final sample path length $M = 1000$. Both initializations lead to roughly the same equilibrium distribution (see figure 3-[center] and figure 2-[right]), and produce exactly the same maximum a posteriori inference for $q$. This, and many other examples revealed the insensitivity of the proposed method to initial conditions. The convergence is always achieved whatever the initial values of $q$.

### Results from the method used in Press and Shigemasu (1998)

The values of the log-likelihood obtained by Press and Shigemasu (1998) seem too close for $q = 3$ and $q = 4$ as revealed by figure 3, which makes the evidence in favor of $q = 3$ not as compellingly clear as the evidence provided by the histograms from the birth-and-death process. This suggests that the birth-and-death process is better in this setting.

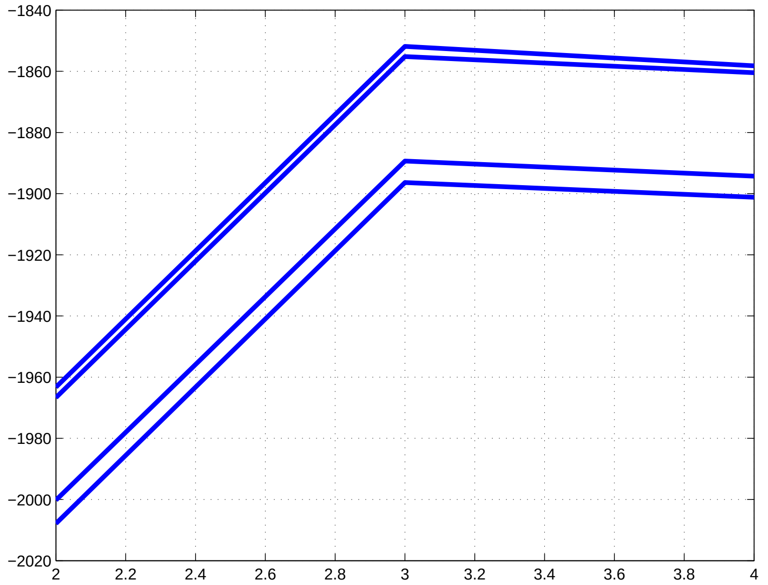

Figure 3: Each line in the plot uses a different random dataset $\mathbf{X}$ to compute the approximate value of $\log(\boldsymbol{P}(Q = q|\mathbf{X}))$ (within an additive constant) for $q = 2, 3, 4$.

## 5.2 Analysis of the wine data set

This dataset is available at the Machine Learning repository of the University of California, Irvine Blake and Merz (1998). It is ranked as a moderately difficult dataset, and has been widely used to test classification methods and algorithms. McLachlan and Peel (2000) used it in the mixture of factor analyzers chapter of his book, and perform sequential likelihood ratio tests to determine the number of factors underlying the $p = 13$ input space variables.

### Results from the screeplot method

Unlike the relatively obvious position of the elbow seen earlier on the artificial dataset, it is unclear in this wine dataset where the "right" position of the elbow would be (see figure 4-left). In any case, it doesn't seem to be anywhere near the value $q = 6$ found by many principled methods used on this dataset. The subjectivity of the position of the elbow on the screeplot clearly reveals one of the main drawbacks of this ad hoc method.

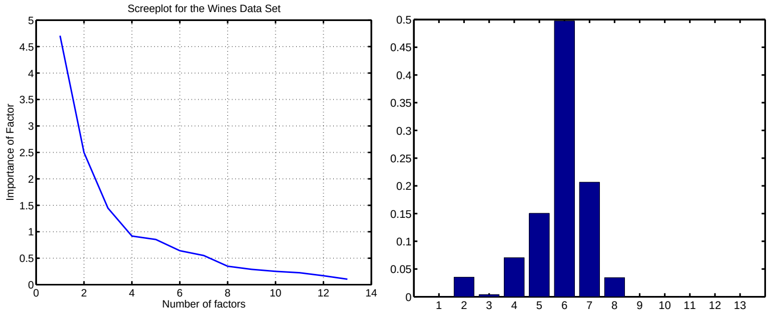

Figure 4: Screeplot and Histogram from stochastic simulation for the wine data

### Results from the stochastic simulation method

The stochastic simulation method is applied here using $T_o = 9500$ burn-in iterations. $\nu = 0.618$ is used as the overall constant birth-rate, and $M = 2500$ is the number of final MCMC samples retained. Figure (4-[right]) strongly suggests that $q = 6$ would be the intrinsic dimensionality of the wine data. The result obtained here by stochastic simulation is the same obtained by McLachlan and Peel (2000) through the use of sequential likelihood ratio tests.

### 5.3 The job application dataset

There are 48 applicants for a certain job, and they have been scored on $p = 15$ variables regarding their acceptability. The observed variables are the following:

(1) Form of letter application
(2) Appearance
(3) Academic ability
(4) Likeabiliy
(5) Self-confidence
(6) Lucidity
(7) Honesty
(8) Salesmanship
(9) Experience
(10) Drive
(11) Ambition
(12) Grasp
(13) Potential
(14) Keenness to join
(15) Suitability

The dataset can be downloaded from Daniel Rowe's website.

http://www.biophysics.mcw.edu/BRI-people/Rowe/BFA.html

Both Press and Shigemasu (1989, 1997), and Rowe and Press (2001) have studied this dataset.

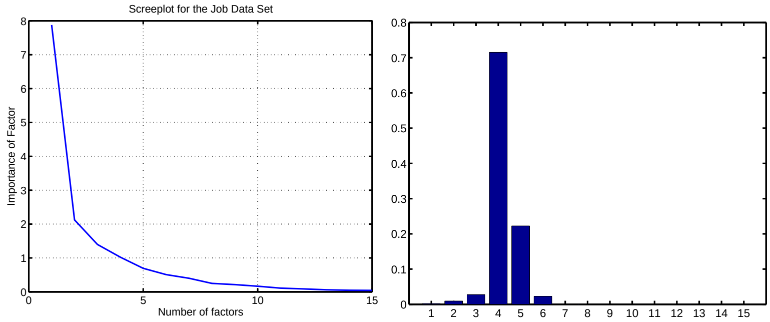

Figure 5: Job application dataset. (Left) Screeplot (Right) Histogram for the number of factors.

### Results from the screeplot method

The position of the elbow in the screeplot (see Figure 5-[Left]) is rather unclear in this case, which makes room for very subjective estimations. This underlines once again another weakness of this approach to the estimation of the number of factors.

### Results from the stochastic simulation method

The histogram of the number of factors (see Figure 5-[Right]) provides an overwhelming evidence in favor of $q = 4$. This estimate is strongly backed by the history behind the dataset. In fact, it would seem that this was a case of confirmatory factor analysis whereby the questionnaire was constructed with these four factors in view, so that the factor analysis actually only served to find the factor loadings. Note: It would be interesting to see the performance of Press and Shigemasu (1998)'s method on this dataset, especially considering the fact that this dataset came from them.

## 6. Conclusion and Discussion

This paper has presented a method to simultaneously extract a simple structure and also determine of the number of factors in orthogonal factor analysis. The technique has the advantage of been straightforward, principled and very easy to interpret. Besides, unlike many other methods before it, the estimates derived do not rely on any form of approximation or ad-hoc scheme.

Although the method has been derived specifically for orthogonal factor analysis, extending it to oblique factor analysis is straightforward, and can therefore be done with very little extra effort.

The computation required is very light, making the scheme useful for practical applications.

It is anticipated that the present computational efficiency would be maintained for $p \leq 100$, which is a reasonable input space dimensionality for many practical factor analysis applications where interpretability and uniqueness are of interest.

When it comes to deriving a point process formulation of the FA model, the constrained structure defined in (3.2) poses a great difficulty due to the fact that it is not rotation invariant. This constrained structure cannot be easily incorporated in the stochastic scheme for determining $q$, since the scheme requires invariance to axes permutation. Despite this drawback however, constraints of this type are very good when $q$ is known because of its ease of implementation and the fact it allows a clear isolation of a unique solution.

On the other hand, the estimate of $\mathbf{\Lambda}$ obtained via the ARD prior approach although constructively simple by virtue of its inherent sparseness, is not guaranteed to be unique. In fact, it may well happen that the prior induces a good number of zeros in the final estimate of $\mathbf{\Lambda}$. ARD is therefore good when one wants to extract a unique simple structure and determine the intrinsic dimensionality simultaneously.

Factor analysis is also heavily used in engineering and pattern recognition as a dimensionality reduction tool. Although an exact number of factors is not as crucial there as it is in confirmatory and exploratory factor analysis, contexts like mixtures of factor analyzers for density estimation would make great use of good intrinsic dimensionality determination as a way to control overfitting.

In this paper, the time interval $(0, T]$ of simulation of the continuous-time birth-and-death point process was divided into intervals of equal length $\rho$. Although, this way of simulating continuous-time processes has been widely used, it turns that the interval $(0, T]$ need not be discretized. The details of such a continuous time simulation along with its advantages are described in a future paper.

## References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika* **52**, 317-332.

Baddeley, A. (1994). Discussion representation of knowledge in complex systems by Grenan- der and Miller. *Journal of the Royal Statistical Society*, Series B **56**, 584-585.

Barndorff-Nielsen, O., Kendall, D. and van Lieshout, M (1999). *Stochastic Geometry: Like- lihood and Computation.* Monographs on Statistics and Applied Probability. Chapman and Hall.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis.* Griffin's Statistical Monographs and Courses. Charles Griffin.

Bartlett, M. (1954). A note on the multiplying factors for various $\chi^2$ approximations. *Journal of the Royal Statistical Society*, Series B **16**, 296-298.

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases[1].

Everitt, B. S. (1984). *An Introduction to Latent Variable Models* (First ed.). Monographs on Statistics and Applied Probability. Chapman and Hall.

Fokoué, E. (2004). Stochastic determination of the intrinsic structure in Bayesian factor analysis. Technical Report 2004-17, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA.

Fokoué, E. and Titterington, D. M. (2003). Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation. *Machine Learning* **50**, 73-94.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Ihara, M. and Kano, Y. (1995). Identifiability of full, marginal and conditional factor analysis model. *Statistics and Probability Letters* **23**, 343-350.

Krzanowski, W. and Marriott, F. (1994). *Multivariate Analysis* (First ed.). Kendall's Library of Statistics 1. Edward Arnold.

Krzanowski, W. and Marriott, F. (1995). *Multivariate Analysis* (First ed.). Kendall's Library of Statistics 2. Arnold.

Lopes, H. F. and West, M. (1999). Model uncertainty in factor analysis. Technical Report ISDS, Institute of Statistics and Decision Sciences, Duke University.

MacKay, D. J. C. (1992). Bayesian Methods for Adaptive Models. Ph. D. thesis, California Institute of Technology, Pasadena, California, USA.

Martin, J. and McDonald, R. (1981). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika* **40**, 505-517.

[1] http://www.ics.uci.edu/m̃learn/MLRepository.html

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Probability and Mathematical Statistics. John Wiley.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks.* Springer.

Press, S. J. (1972). *Applied Multivariate Analysis* (First ed.). Holt, Rinehart and Winston.

Press, S. J. and K. Shigemasu (1989). Bayesian Inference in Factor Analysis. In *Contributions to Probability and Statistics* (Edited by L. J. Glesser ), Chapter 15. Springer Verlag.

Press, S. J. and K. Shigemasu (1998). A note on choosing the number of factors. *Communications in Statistics: Theory and Methods* **29**, 1653-1670.

Preston, C. (1976). Spatial birth-and-death process. *Bull. Inst. Internat. Statist.* **46**, 371-391.

Ripley, B. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc.* Ser. B **39**, 172-212.

Rowe, D. B. (2003). *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing.* CRC Press.

Stephens, M. (2000). Bayesian analysis of mixtures models with an unknown number of components – an alternative to reversibe jump methods. *Annals of Statistics* **28**, 40-74.

Stoyan, D., Kendall, W. and Mecke, J. (1995). *Stochastic Geometry and its Applications* (second ed.). Wiley Series in Probability and Statistics. John Wiley.

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research* **1**, 211-244.

van Lieshout, M. (1994). Discussion on representation of knowledge in complex systems by Grenander and Miller. *Journal of the Royal Statistical Society*, Series B **56**, 585.

Received October 8, 2007; accepted September 2, 2008.

Ernest Fokoué
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive
Rochester, NY 14623, USA
ernest.fokoue@gmail.com