Journal of Interdisciplinary Mathematics

Index of Volume 5 (2002)

Number 1, February

A Bayesian approach to blind source separation

Daniel B. Rowe
Biophysics Research Institute
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI 53226
U.S.A.

ABSTRACT

This paper presents a Bayesian statistical approach to the blind source separation problem. The blind source separation model is described; the source distribution is discussed; other approaches such as Principal Components, Independent Components, and Factor Analysis are detailed; prior distributions are introduced to incorporate available prior knowledge; the posterior distribution for the model parameters (including the number of sources) is derived; and the parameter estimation procedure is outlined. Finally Bayesian blind source separation is applied in a simulated example and its advantages over the other methods are stated.

1. INTRODUCTION AND MODEL

The problem addressed by blind source separation is that of separating unobservable or latent source signals when mixed signals are observed. In other words, to take a set of observed mixed signal vectors and unmix or separate them into a set of true unobservable source signal vectors. This paper adopts a Bayesian statistical approach and a linear synthesis model.

In the Bayesian approach to statistical inference, available prior information either from subjective expert experience or prior experiments is incorporated into the inferences. This prior information yields progressively less influence in the final results as the sample size increases, thus allowing the data to "speak the truth." It should also be noted that forcing the components of the source vectors to be independent, as is done in independent component analysis, is too constraining. Here the components of the source vectors are allowed to be correlated, as is frequently the case.

Journal of Interdisciplinary Mathematics
Vol. 5 (2002), No. 1, pp. 49-76

To motivate the blind separation of sources model, the context of the "cocktail party problem" is adopted. At a cocktail party, there are p microphones that record or observe m partygoers or speakers at n time increments. The observed conversations consist of mixtures of true conversations. In other words, p-dimensional mixed signal vectors $x_i = (x_{i1}, \ldots, x_{ip})'$ are observed and the goal is to separate these observed signal vectors into m-dimensional true underlying source signal vectors, $s_i = (s_{i1}, \ldots, s_{im})'$ where $i = 1, \ldots, n$.

The general blind separation of sources model is

$$\underset{(p \times 1)}{(x_i \mid s_i, m)} = \underset{(p \times 1)}{f(s_i \mid m)} + \underset{(p \times 1)}{\varepsilon_i}, \tag{1}$$

where $f(s_i \mid m)$ is a function that mixes the source signals and ε_i is the measurement error. Using a Taylor series expansion, the function f, with appropriate assumptions can be expanded about the vector c and written as

$$f(s_i \mid m) = f(c) + f'(c)(s_i - c) + \ldots$$

and approximated by taking the first two terms

$$\begin{aligned} f(s_i \mid m) &= f(c) + f'(c)(s_i - c) \\ &= [f(c) - f'(c)c] + f'(c)s_i \\ &= \mu + \Lambda s_i \end{aligned} \tag{2}$$

where $f'(c)$ and Λ are $p \times m$ matrices. This is the linear synthesis model. More formally the adopted model is

$$\underset{(p \times 1)}{(x_i \mid \mu, \Lambda, s_i, m)} = \underset{(p \times 1)}{\mu} + \underset{(p \times m)}{\Lambda} \underset{(m \times 1)}{s_i} + \underset{(p \times 1)}{\varepsilon_i}, \tag{3}$$

where

- μ = a p-dimensional unobserved population mean vector, $\mu = (\mu_1, \ldots, \mu_p)'$;
- Λ = a $p \times m$ matrix of unobserved mixing constants, $\Lambda = (\lambda_1', \ldots, \lambda_p')'$;
- s_i = the ith m-dimensional unobservable source vector, $s_i = (s_{i1}, \ldots, s_{1m})'$; and
- ε_i = the p-dimensional vector of errors or noise terms of the ith observed signal vector, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})'$.

The observed mixed signal x_{ij} is the jth element of the observed mixed signal vector i, which may be thought of as the recorded mixed conversation signal at time increment i, $i = 1, \ldots, n$ for microphone j, $j = 1, \ldots, p$. The observed signal x_{ij} is a mixture of the components or true unobserved source signal conversations s_i with error, at time increment i, $i = 1, \ldots, n$. The unobserved source signal s_{ik} is the kth component of the unobserved source vector i, which may be thought of as the unobserved source signal conversation of speaker k, $k = 1, \ldots, m$ at time increment i, $i = 1, \ldots, n$.

The model describes the mixing process by writing the observed signal x_{ij} as the sum of an overall mean part μ_j plus a linear combination of the unobserved source signal components s_{ik} and the observation error ε_{ij} as

$$(x_{ij} \mid \mu_j, \lambda_j, s_i, m) = \mu_j + \sum_{k=1}^{m} \lambda_{jk} s_{ik} + \varepsilon_{ij}$$

$$= \mu_j + \lambda_j' s_i + \varepsilon_{ij}. \tag{4}$$

2. LIKELIHOOD

Regarding the errors of the observations, it is assumed that they are independent normally distributed random vectors with mean zero and covariance matrix Ψ. This will be referred to as model assumption (a). Formally, it is assumed that

$$(\varepsilon_i \mid \Psi) \sim N(0, \Psi) \tag{a}$$

for all i, $i = 1, \ldots, n$. From (a), it is seen that the observation vector given the mean, the number of sources, the source vector, the mixing matrix, and the error covariance matrix is normally distributed as expressed by

$$(x_i \mid \mu, \Psi, m, \Lambda, s_i) \sim N(\mu + \Lambda s_i, \Psi), \tag{2.1}$$

and the likelihood is

$$p(x_i \mid \mu, \Psi, m, \Lambda, s_i)$$

$$= (2\pi)^{-\frac{p}{2}} \mid \Psi \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(x_i - \mu - \Lambda s_i)\Psi^{-1}(x_i - \mu - \Lambda s_i)}. \tag{2.2}$$

The joint likelihood of the observations is

$$p(x_1, \ldots, x_n \mid \mu, \Psi, m, \Lambda, s_1, \ldots, s_n)$$

$$= (2\pi)^{-\frac{np}{2}} |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu-\Lambda s_i)'\Psi^{-1}(x_i-\mu-\Lambda s_i)}. \tag{2.3}$$

The goal is to unmix the sources, s_i's by computing estimates of their values from probability distributions. It is further wished to gain knowledge as to the mixing process by estimating the overall mean μ, the mixing matrix Λ, and the error matrix Ψ.

3. SOURCE DISTRIBUTION

As Anderson and Rubin (1956) point out, two kinds of model can be described. In the first, the sources can be considered to be random vectors and in the second considered to be nonrandom (fixed) vectors which vary from one sample to another. In the first case, the distribution of the sample $x_1, \ldots, x_n$ is equivalent to that of any other sample of size n, $x_{n+1}, \ldots, x_{2n}$. In the second case, the distribution of the set of observations $x_1, \ldots, x_n$ is not equivalent to the distribution of $x_{n+1}, \ldots, x_{2n}$ because $s_1, \ldots, s_n$ is not equivalent to $s_{n+1}, \ldots, s_{2n}$ which enter as parameters.

The BSS model with nonrandom sources is inherently indeterminate. The values of the parameters are not uniquely determined by the likelihood. Additional information regarding the source vectors aid in remedying this problem. Since the source vectors are unknown, this uncertainty is quantified in the form of a prior distribution.

Anderson and Rubin further show that for the classical factor analysis model, random sources are asymptotically equivalent to fixed sources and that the estimates of Λ and Ψ for random sources can be used for nonrandom sources in large samples due to asymptotic convergence. For these reasons, the models include the sources as random quantities.

When the source vectors are viewed as random quantities, then they must have a (prior) distribution associated with them, say $p(s_i \mid m, R)$. The prior distribution is a model assumption and will be assigned the letter (b) with as appropriate subscript for each approach. In general, a given unobserved source signal vector s_i has zero expectation and variance R. This is general assumption (b_g) stated as

$$E(s_i \mid m, R) = 0 \quad \text{and} \quad var(s_i \mid m, R) = R. \tag{b_G}$$

The above leads to the following expectations and variances for the observed signals

$$E(x_i \mid \mu, \Psi, m, \Lambda) = E(\mu + \Lambda s_i + \varepsilon_i \mid \mu, \Psi, m)$$

$$= \mu + \Lambda E(s_i \mid m, R) + E(\varepsilon_i \mid \Psi)$$

$$= \mu \tag{3.1}$$

and

$$var(x_i \mid \mu, \Psi, m, \Lambda) = var(\Lambda s_i + \varepsilon_i \mid \Psi, m, \Lambda)$$

$$= \Lambda var(s_i \mid m, R)\Lambda' + var(\varepsilon_i \mid \Psi)$$

$$= \Lambda R \Lambda' + \Psi$$

$$= \Sigma. \tag{3.2}$$

It is assumed that the sources have a mean of zero. If the source mean were in fact non zero, say s_0, then $E(x_i \mid \mu, m, \Psi, \Lambda) = \mu^* + \Lambda s_0 = \mu$ and the sources have a mean of zero.

The joint distribution of the observed mixed signals and the unobserved source signals is

$$p(x, s \mid \mu, \Psi, m, \Lambda, R) = p(x \mid \mu, \Psi, m, \Lambda, s)p(s_i \mid m, R)$$

$$= (2\pi)^{-\frac{np}{2}} \mid \Psi \mid^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu-\Lambda s_i)'\Psi^{-1}(x_i-\mu-\Lambda s_i)}$$

$$\cdot \prod_{i=1}^{n} p(s_i \mid m, R) \tag{3.3}$$

where $x = (x_1, \ldots, x_n)'$ and $s = (s_1, \ldots, s_n)'$.

It is easily shown that the maximum likelihood estimator for the mean μ is $\bar{x}$, so without loss of generality and for simplicity, it is assumed that the observations have been centered about the sample mean.

The blind source separation model is now

$$\underset{(p \times 1)}{(x_i \mid \Lambda, s_i, m)} = \underset{(p \times m)(m \times 1)}{\Lambda \quad s_i} + \underset{(p \times 1)}{\varepsilon_i}, \tag{3.4}$$

and the likelihood is

$$p(x_i \mid \Psi, m, s_i, \Lambda) = (2\pi)^{-\frac{p}{2}} \mid \Psi \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(x_i-\Lambda s_i)'\Psi^{-1}(x_i-\Lambda s_i)} \tag{3.5}$$

The observations are sometimes scaled or normalized so that Σ is a matrix of correlations.

There is an important fact regarding the product of the mixing matrix and the source covariance matrix. The covariance between the observation vectors and the unobserved source vectors is

$$\begin{aligned} cov(x_i, s_i \mid m, \Lambda, \Psi, R) &= E(x_i s_i' \mid m, \Lambda, \Psi, R) \\ &= E(x_i \mid m, \Lambda, \Psi) E(s_i' \mid m, R) \\ &= E(\Lambda s_i s_i' \mid m, R) \\ &= \Lambda E(s_i s_i' \mid m, R) \\ &= \Lambda R. \end{aligned} \tag{3.6}$$

The matrix ΛR is interpreted as a matrix of covariances between the p elements of the observed mixed signal vectors and the m components of the unobserved source signal vectors. (If $R = I_m$ as in orthogonal factor analysis, then the mixing matrix is a matrix of covariances.) The element in the jth row and the kth column of ΛR is the covariance between the jth element of the observed mixed signal vectors and the kth component of the unobserved source signal vector.

Thus a large element of ΛR imply a strong relationship between the corresponding observed signal dimension and unobserved source signal dimension. In the "cocktail party problem", this implies that information on the location of partygoers with respect to microphones can be gained.

Analogous to regression, the blind source separation model can be written in terms of matrices as

$$\underset{(n \times p)}{(X \mid \Lambda, S, m)} = \underset{(n \times m)}{S} \underset{(m \times p)}{\Lambda'} + \underset{(n \times p)}{E}, \tag{3.7}$$

the likelihood as

$$p(X \mid \Psi, m, S, \Lambda) = (2\pi)^{-\frac{np}{2}} \mid \Psi \mid^{-\frac{n}{2}} e^{-\frac{1}{2} tr \Psi^{-1} (X - S\Lambda')'(X - S\Lambda')} \tag{3.8}$$

where $X' = (x_1, \ldots, x_n)$, $S' = (s_1, \ldots, s_n)$, and $E' = (\varepsilon_1, \ldots, \varepsilon_n)$.

Again, the objective is to unmix the sources by estimating S and to obtain knowledge about the mixing process by estimating Λ and Ψ.

4. PRINCIPAL COMPONENT ANALYSIS

For principal component analysis (PCA) it is assumed that the number of signal sources is equal to the number of observed signals

($m = p$) and the observation error (noise) is zero ($\Psi = 0$). Further, the population covariance matrix $\Sigma = \Lambda R \Lambda'$ is estimated by its sample value $\hat{\Sigma} = \dfrac{X'X}{n}$. Now, an estimate of the mixing matrix is computed.

It is assumed that the columns of Λ are orthonormal and each sequentially maximizes the percent of variation, but no assumption is made as to the form of the distribution of the sources.

No distributional assumption is made for $p(s_i \mid m, R)$. (b_{PCA})

The matrix $W = \Lambda^{-1}$ is to be determined so that the sources can be unmixed as

$$\begin{aligned} s_i &= \Lambda^{-1} x_i \\ &= W x_i \\ &= \begin{pmatrix} w_1' \\ \vdots \\ w_p' \end{pmatrix} x_i. \end{aligned}$$

The first component of the unobserved source s is

$$s_{i1} = w_1' x_i$$

with variance given by

$$\begin{aligned} var(s_{i1} \mid R) &= w_1' var(x_i) w_1 \\ &= w_1' \Sigma w_1. \end{aligned}$$

The vector w_1 is now determined to be that value that maximizes the variance subject to $w_1' w_1 = 1$. The method of Lagrange multipliers is applied

$$\frac{\partial}{\partial w_1} [w_1' \Sigma w_1 - \theta_1 (w_1' w_1 - 1)] = 2\Sigma w_1 - 2\theta_1 w_1 = 0$$

which is reexpressed as

$$(\Sigma - \theta_1 I_p) w_1 = 0$$

and since $w_1 \neq 0$, there can only be a solution if

$$|\Sigma - \theta_1 I_p| = 0.$$

It is apparent that θ_1 must be a latent root of Σ and w_1 is a normalized latent vector of Σ. There are p such latent roots that satisfy the equation. The largest is selected. The other rows of W are found in a similar fashion.

The variance of the unobserved sources is

$$var(s_i \mid m, R) = W\, var(x \mid \mu, \Sigma) W'$$

$$R = W\Sigma W'$$

and because we assumed that the rows of W were orthogonal, R is a diagonal matrix. Thus the source components are uncorrelated, $R = diag(r_1, \ldots, r_p)$.

Now the factorization $\Sigma = W' D_\theta W = \Lambda R \Lambda'$ can be written because $WW' = W'W = I_p$ and $R = D_\theta = diag(\theta_1, \ldots, \theta_p)$. The sources can now be separated by $s_i = Wx_i$. For a more detailed account of the procedure refer to Press (1982).

5. INDEPENDENT COMPONENT ANALYSIS

If it is further assume that the components of the source signals are independent, which is a more stringent requirement than being uncorrelated, then the described model is the independent components analysis (ICA) model. This model also assumes that the number of source signal components is equal to the number of observed signal elements ($m = p$) and that the observation error (noise) is zero ($\Psi = 0$).

The probability distribution for the source signals with independent components is

$$p(s_i \mid m) = \prod_{k=1}^{m} p(s_{ik}). \qquad (b_{ICA})$$

Thus for no distributional assumptions have been made about the independent source signal components although it can be assumed that they are normal, as in the next section.

The likelihood for the observations is

$$p(x_i \mid s_i, m, \Lambda) = \delta(x_i - \Lambda s_i), \qquad (5.1)$$

a delta function due to the zero noise limit. The change of variable $u_i = \Lambda^{-1}x_i = Wx_i$ is made and thus

$$p(u_i \mid s_i, m, W) = \mid W \mid \delta(u_i - s_i). \tag{5.2}$$

The likelihood and the prior distribution for the source components are combined to form their joint distribution

$$p(u_i, s_i \mid m, W) = \mid W \mid \delta(u_i - s_i) = \prod_{j=1}^{m} p(s_{ij}), \tag{5.3}$$

and is integrated with respect to the source signal components, to obtain

$$p(u_i \mid m, W) = \int \mid W \mid \delta(u_i - s_i) \prod_{k=1}^{m} p(s_{ik}) ds_i$$

$$= \mid W \mid \prod_{k=1}^{m} p(u_{ik}). \tag{5.4}$$

As stated by MacKay (1996), upon taking the logarithm, differentiating, and changing Λ to ascend the gradient, the learning algorithm of Bell and Sejnowski (1995) is obtained.

6. FACTOR ANALYSIS

If it is assumed that observation errors (noise) are non zero ($\Psi \neq 0$), that there are at most as many unobserved source signal components as observed mixed signal elements ($m \leq p$), that R is a correlation matrix, and several other traditional assumptions that round out the psychologic model, then this describes the factor analysis (FA) model.

It is common to assume that the unobserved source signal components are uncorrelated, a weaker assumption than independence. In particular model assumption (b_{FA}) is made, that a priori, the source vectors are distributed as

$$(s_i \mid m) \sim N(0, R = I_m), \tag{b_{FA}}$$

and is in PCA and ICA it is assumed that (c) the error vectors and the source vectors are independent,

$$(\varepsilon_i \mid \Psi) \ \& \ (s_i \mid m, R) \text{ are independent.} \tag{c}$$

When $R = I_m$, this is called the orthogonal factor model and when R is a general correlation matrix, called the oblique factor model. The orthogonal model is most common due to interpretability. Recall that when $R = I_m$ the mixing matrix called the factor loading matrix in factor analysis has the interpretation of being a matrix of covariances (correlations) between the observed elements and the unobserved source components called factor scores. In assuming that the sources have an identity covariance, the distribution of the observations remains unchanged. For example, if it was assumed that

$$(s_i \mid m, R^*) \sim N(0, R^*), \tag{6.1}$$

where R^* is a general covariance matrix, then

$$(x_i \mid m, \Lambda, \Psi, R^*) \sim N(0, \Lambda R^* \Lambda' + \Psi) \tag{6.2}$$

which is the same as

$$(x_i \mid m, \Lambda, \Psi) \sim N(0, (\Lambda R^{*\frac{1}{2}}) R (\Lambda R^{*\frac{1}{2}})' + \Psi)$$

where R is the identity matrix or a correlation matrix.

The resulting probability distribution function is

$$p(s_i \mid m, R) = (2\pi)^{-\frac{m}{2}} \mid R \mid^{-\frac{n}{2}} e^{-\frac{1}{2} s_i' R^{-1} s_i} \tag{6.3}$$

and the joint prior probability distribution function for all the unobserved source signal vectors is

$$p(S \mid m, R) = (2\pi)^{-\frac{nm}{2}} \mid R \mid^{-\frac{n}{2}} e^{-\frac{1}{2} tr S R^{-1} S'}. \tag{6.4}$$

The likelihood for the observed mixed signals and the prior (model) distribution for the unobserved source signal vectors are combined to form their joint distribution

$$\begin{aligned} p(X, S \mid \Psi, m, \Lambda, R) = {} & (2\pi)^{-\frac{np}{2}} \mid \Psi \mid^{-\frac{n}{2}} \\ & \cdot e^{-\frac{1}{2} tr \Psi^{-1} (X - S\Lambda')'(X - S\Lambda')} \\ & \cdot (2\pi)^{-\frac{nm}{2}} \mid R \mid^{-\frac{n}{2}} e^{-\frac{1}{2} tr S R^{-1} S'}. \end{aligned} \tag{6.5}$$

There are several approaches to estimating the unobserved source vectors, the mixing matrix, and the errors. For classical factor analysis the estimation methods range from maximum likelihood by the method of Lawley (1940) to EM maximum likelihood by Rubin and Thayer (1982). More recently, the modern Bayesian approach of Press and Shigemasu (1989/1997) that incorporate prior knowledge about the parameter values with the result of eliminating indeterminacies has generated much activity. For details and brief overview of classical and Bayesian factor analysis see Rowe (1998). Unfortunately, factor analysis has the shortcoming that only normalized sources are obtained.

7. BAYESIAN BLIND SOURCE SEPARATION

The aforementioned models are expanded to the general Bayesian blind source separation (BBSS) model for several reasons. In the real "cocktail party problem" the number of speakers is not in general equal to the number of microphones as required by PCA and ICA, or less than or equal to the number of microphones as required by FA. The FA assumption that the sources have unit variance is inadequate because only normalized sources are separated. It is further believed by the current author that no observation is truly noiseless, as is required by ICA and PCA. There is always some non zero random variability as alluded to in MacKay (1996). BBSS also does not constrain the sources to be independent as ICA does. This is obviously not the case since speakers are not acting independently. BBSS allows the source components to be correlated.

7.1. *Likelihood*

Continuing, the blind source separation model is

$$\underset{(n\times p)}{(X \mid m, \Lambda, S)} = \underset{(n\times m)}{S}\,\underset{(m\times p)}{\Lambda'} + \underset{(n\times p)}{E}, \tag{7.1}$$

with assumptions

(a) $(\varepsilon_i \mid \Psi) \sim N(0, \Psi)$

(b) $(s_i \mid m, R) \sim N(0, R)$, R a general covariance matrix,

and it has implicitly been assumed that

(c) $(\varepsilon_i \mid \Psi)$ and $(s_i \mid m, R)$ are independent.

That is, the mixed signals are observed with error that is normally distributed with mean zero and variance-covariance matrix Ψ, the unobserved source signals are normally distributed with mean

zero and covariance matrix R, and the errors and the sources are independent.

From assumption (a), the likelihood for the observed signals is

$$p(X \mid \Psi, m, S, \Lambda) = (2\pi)^{-\frac{np}{2}} \mid \Psi \mid^{-\frac{n}{2}} e^{-\frac{1}{2} tr\Psi^{-1}(X-S\Lambda')(X-S\Lambda')} \tag{7.2}$$

where $X' = (x_1, \ldots, x_n)$, $S' = (s_1, \ldots, s_n)$, and $E' = (\varepsilon_1, \ldots, \varepsilon_n)$.

The advantage of the Bayesian statistical approach is that available prior information about parameter values can formally be brought to bear in the problem through prior distributions. As stated earlier, the prior parameter values will have decreasing influence in the posterior estimates with increasing sample size, and the sources are allowed to deviate from their a priori values, thus allowing the data to "speak the truth." This paper follows the Bayesian paradigm by assessing prior distributions for the remaining unknown parameters.

7.2. *Priors and Posterior*

Natural conjugate prior distributions are assessed for the model parameters. It is specified that the prior distributions for the source covariance matrix R, the error covariance matrix Ψ, and the mixing matrix Λ follow inverted Wishart, inverted Wishart, and normal distributions respectively

$$p(R \mid m) = c(\eta, m) \mid R \mid^{-\frac{\eta}{2}} e^{-\frac{1}{2} tr R^{-1} V}, \tag{7.3}$$

$$p(\Psi) = c(\nu, p) \mid \Psi \mid^{-\frac{\nu}{2}} e^{-\frac{1}{2} tr\Psi^{-1} B}, \tag{7.4}$$

$$p(\Lambda \mid \Psi, m) = (2\pi)^{-\frac{pm}{2}} \mid H \mid^{\frac{p}{2}} \mid \Psi \mid^{-\frac{m}{2}} \cdot e^{-\frac{1}{2} tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \tag{7.5}$$

where $\nu > 2p$, $\eta > 2m$; R, V, Ψ and H are positive definite matrices; $c(\eta, m)$ and $c(\nu, p)$ are constants depending only on their arguments. Further,

$$c(\eta, m)^{-1} = 2^{\frac{(\eta-m-1)m}{2}} \pi^{\frac{m-1}{4}} \prod_{k=1}^{m} \Gamma\left(\frac{n-m-k}{2}\right)$$

and $c(\nu, p)$ has the same form. The hyperparameters η, V, ν, B, Λ_0, and H are to be assessed. Hyperparameter assessment is discussed in an appendix.

For simplicity of assessment, it is specified that $B = b_0 I_p$, $V = v_0 I_m$ and $H = h_0 I_m$. The prior distribution $p(\Lambda | \Psi, m)$ comes from writing $\tau = vec(\Lambda')$, and specifying a multivariate normal distribution. As a consequence, $E(\tau) = vec(\Lambda_0')$ and $var(\tau) = I_m \otimes h_0^{-1} E(\Psi)$. Since B is a priori diagonal, or equivalently the errors are a priori uncorrelated, the elements of Λ are a priori uncorrelated.

A big question in the blind separation of sources problem has been neglected. How many unobserved source signals are to be unmixed? The number of sources is often unknown. The methods in this paper allow researchers to incorporate their prior beliefs as to the number of sources in the form of a prior distribution.

It allows the general specification of a discrete prior distribution $p(m)$ on the number of sources m. If the true number of sources were known, then the degenerate distribution

$$p(m) = \begin{cases} 1, & \text{if } m = m_0 \\ 0, & \text{if } m \neq m_0, \end{cases} \tag{7.6}$$

is assigned, and the posterior is only defined for $m = m_0$.

Upon using Bayes' rule the posterior distribution for the unknown parameters is

$$\begin{aligned} p(m, S, \Lambda, \Psi, R | X) \propto p(m)(2\pi)^{-\frac{(n+p)m}{2}} |H|^{\frac{p}{2}} |\Psi|^{-\frac{(n+m+v)}{2}} \\ \cdot c(\eta, m) |R|^{-\frac{n+\eta}{2}} e^{-\frac{1}{2} tr R^{-1}(S'S+V)} \\ \cdot e^{-\frac{1}{2} tr \Psi^{-1} U} \end{aligned} \tag{7.7}$$

where

$$U \equiv (X - S\Lambda')'(X - S\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B,$$

and $\propto$ denotes proportionality.

This posterior distribution must now be evaluated in order to obtain parameter estimates of the number of sources, the mixing matrix, the sources, the errors of the sources, and the errors of observation.

8. ESTIMATION

With the above posterior distribution, it is not possible to obtain marginal distributions and thus marginal estimates for any of the parameters in an analytic closed form. For this reason, the deterministic hill climbing Lindley/Smith optimization (LSO) is used. It is also possible to use stochastic Gibbs sampling (as outlined in the

appendix) but it is more computationally intensive and time consuming. For the LSO estimation procedure, the posterior conditional distributions are required.

From the joint posterior distribution we can obtain the posterior conditional distributions. The conditional posterior distributions for the mixing matrix is

$$\begin{aligned}
p(\Lambda \mid \Psi, m, S, R, X) &\propto p(\Psi, m, S, \Lambda, R)p(X \mid \Psi, m, F, \Lambda) \\
&\propto p(\Psi)p(m)p(S \mid m, R)p(\Lambda \mid \Psi, m)p(R) \\
&\quad \cdot p(X \mid \Psi, m, S, \Lambda) \\
&\propto p(\Lambda \mid \Psi, m)p(X \mid \Psi, m, S, \Lambda) \\
&\propto |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \\
&\quad \cdot |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-S\Lambda')'(X-S\Lambda')} \\
&\propto e^{-\frac{1}{2}tr\Psi^{-1}[(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'+(X-S\Lambda')'(X-S\Lambda')]} \\
&\propto e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\tilde{\Lambda})(H+S'S)(\Lambda-\tilde{\Lambda})'}
\end{aligned} \tag{8.1}$$

where the posterior conditional mean and mode is given by

$$\tilde{\Lambda} = [X'S + \Lambda_0 H](H + S'S)^{-1}.$$

The conditional distribution for the mixing matrix the observation error covariance matrix, the number of sources, the sources, the source covariance matrix, and the data is normally distributed.

The conditional posterior distribution of the observation error matrix is

$$\begin{aligned}
p(\Psi \mid m, S, \Lambda, R, X) &\propto p(\Psi, m, S, \Lambda, R)p(X \mid \Psi, m, S, \Lambda) \\
&\propto p(\Psi)p(m)p(S \mid m, R)p(\Lambda \mid \Psi, m)p(R) \\
&\quad \cdot p(X \mid \Psi, m, S, \Lambda) \\
&\propto p(\Psi)p(\Lambda \mid \Psi, m)p(X \mid \Psi, m, S, \Lambda) \\
&\propto |\Psi|^{-\frac{\nu}{2}} e^{-\frac{1}{2}tr\Psi^{-1}B} |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'} \\
&\quad \cdot |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-S\Lambda')'(X-S\Lambda')} \\
&\propto |\Psi|^{-\frac{(n+m+\nu)}{2}} e^{-\frac{1}{2}tr\Psi^{-1}U}
\end{aligned} \tag{8.2}$$

where

$$U = (X - S\Lambda')'(X - S\Lambda') + (\Lambda - \Lambda_0)H(\Lambda - \Lambda_0)' + B \tag{8.3}$$

with a mode given by

$$\Psi = \frac{U}{n + m + \nu}. \tag{8.4}$$

The distribution of the observation error covariance matrix given the number of sources, the mixing matrix, the sources, the source covariance matrix, and the data is an inverted Wishart.

The conditional posterior distribution for the sources is

$$\begin{aligned}
p(S \mid \Psi, m, \Lambda, R, X) &\propto p(\Psi, m, S, \Lambda, R)p(X \mid \Psi, m, S, \Lambda) \\
&\propto p(\Psi)p(m)p(S \mid m, R)p(\Lambda \mid \Psi, m)p(R) \\
&\quad \cdot p(X \mid \Psi, m, S, \Lambda) \\
&\propto p(S \mid m, R)p(X \mid \Psi, m, S, \Lambda) \\
&\propto |R|^{-\frac{n}{2}} e^{-\frac{1}{2}tr SR^{-1}S'} \\
&\quad \cdot |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-S\Lambda')'(X-S\Lambda')} \\
&\propto |\Psi|^{-\frac{n}{2}} |R|^{-\frac{n}{2}} e^{-\frac{1}{2}tr[SR^{-1}S' + (X-S\Lambda')\Psi^{-1}(X-S\Lambda')']} \\
&\propto e^{-\frac{1}{2}tr(S-\tilde{S})(R^{-1}+\Lambda'\Psi^{-1}\Lambda)(S-\tilde{S})'}
\end{aligned}$$

where the posterior conditional mean and mode is given by

$$\tilde{S} = X\Psi^{-1}\Lambda(R^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}.$$

The conditional posterior distribution for the sources given the observation error covariance matrix, the number of sources, the mixing matrix, the sources, the source covariance matrix, and the data is normally distributed.

The conditional posterior distribution for the source covariance matrix is

$$p(R \mid \Psi, m, S, \Lambda, X) \propto p(\Psi, m, S, \Lambda, R)p(X \mid \Psi, m, S, \Lambda)$$

$$\propto p(\Psi)p(m)p(S \mid m, R)p(\Lambda \mid \Psi, m)p(R)$$

$$\cdot p(X \mid \Psi, m, S, \Lambda)$$

$$\propto p(S \mid m, R)p(R)$$

$$\propto |R|^{-\frac{n}{2}} e^{-\frac{1}{2}trSR^{-1}S'}$$

$$\cdot |R|^{-\frac{\eta}{2}} e^{-\frac{1}{2}trR^{-1}V}$$

$$\propto |R|^{-\frac{(n+\eta)}{2}} e^{-\frac{1}{2}trR^{-1}(S'S+V)}$$

with the posterior conditional mode given by

$$\widetilde{R} = \frac{S'S + V}{n + \eta}. \tag{8.5}$$

The conditional posterior distribution for the source covariance matrix given the observation error covariance matrix, the number of sources, the sources, the mixing matrix, and the data is normally distributed.

All of these conditional posterior densities are well known recognizable distributions that do not require rejection sampling. Standard random variable generation methods can be used.

However, the conditional distribution for the number of sources is not tractable and recognizable.

The conditional posterior distribution of the number of sources is

$$p(m \mid \Psi, S, \Lambda, R, X) \propto p(X \mid \Psi, m, S, \Lambda)p(\Psi, m, S, \Lambda, R)$$

$$\propto p(X \mid \Psi, m, S, \Lambda)p(\Psi)p(S \mid m, R)p(\Lambda \mid \Psi, m)p(R)p(m)$$

$$\propto (2\pi)^{-\frac{np}{2}} |\Psi|^{-\frac{n}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(X-S\Lambda')(X-S\Lambda')}$$

$$\cdot |\Psi|^{-\frac{\nu}{2}} e^{-\frac{1}{2}tr\Psi^{-1}B} (2\pi)^{-\frac{nm}{2}} |R|^{-\frac{n}{2}} e^{-\frac{1}{2}trSR^{-1}S'}$$

$$\cdot (2\pi)^{-\frac{pm}{2}} |H|^{\frac{p}{2}} |\Psi|^{-\frac{m}{2}} e^{-\frac{1}{2}tr\Psi^{-1}(\Lambda-\Lambda_0)H(\Lambda-\Lambda_0)'}$$

$$\cdot c(\eta, m) |R|^{-\frac{\eta}{2}} e^{-\frac{1}{2}trR^{-1}V} p(m)$$

$$\propto p(m)c(\eta, m)\,|R|^{-\frac{n+\eta}{2}} e^{-\frac{1}{2}\mathrm{tr}R^{-1}(S'S+V)} (2\pi)^{-\frac{(n+p)m}{2}} |H|^{\frac{p}{2}}$$

$$\cdot\,|\Psi|^{-\frac{(n+m+\nu)}{2}} e^{-\frac{1}{2}\mathrm{tr}\Psi^{-1}U}. \tag{8.6}$$

As previously stated, the conditional posterior density for the number of sources given the observation error covariance matrix, the sources, the source covariance matrix the mixing matrix, and the data does not have a tractable and recognizable form.

It is not necessary to compute modes from the above number of sources conditional posterior distribution for LSO (or generate random samples for Gibbs sampling). Parameter estimates can be computed for each number of sources and the value that maximizes the posterior conditional distribution is selected.

The Lindley/Smith optimization procedure consists of starting with initial values for m, S and R, say $\tilde{m}_{(0)}$, $\tilde{S}_{(0)}$, $\tilde{R}_{(0)}$ and for a given number of sources $m = \tilde{m}_{(l)}$ cycling through

$$\tilde{\Lambda}_{(l+1)} \equiv (X'\tilde{S}_{(l)} + \Lambda_0 H)(H + \tilde{S}'_{(l)}\tilde{S}_{(l)})^{-1},$$

$$\tilde{\Psi}_{(l+1)} \equiv \frac{(X - \tilde{S}_{(l)}\tilde{\Lambda}'_{(l+1)})'(X - \tilde{S}_{(l)}\tilde{\Lambda}'_{(l+1)}) + (\tilde{\Lambda}_{(l+1)} - \Lambda_0)H(\tilde{\Lambda}_{(l+1)} - \Lambda_0)' + B}{n + m + \nu}$$

$$\tilde{S}_{(l+1)} \equiv X\,\tilde{\Psi}^{-1}_{(l+1)}\tilde{\Lambda}_{(l+1)}(\tilde{R}^{-1}_{(l)} + \tilde{\Lambda}'_{(l+1)}\tilde{\Psi}^{-1}_{(l+1)}\tilde{\Lambda}_{(l+1)})^{-1}.$$

$$\tilde{R}_{(l+1)} \equiv \frac{\tilde{S}'_{(l+1)}\tilde{S}_{(l+1)} + V}{n + \eta}$$

until covergence is reached. The converged values $(\tilde{S}, \tilde{\Lambda}, \tilde{\Psi}, \tilde{R})$ are joint posterior modal estimators of the parameters for the given number of sources.

We carry out these procedures for each value of the number of sources m, then find the value of the number of sources that makes the posterior conditional distribution for the number of sources a maximum given the corresponding estimates of the other parameters. This is the same as selecting the number of sources to be that value which makes the conditional posterior odds ratio a maximum.

9. EXAMPLE

For an example, a simulation was carried out. The number of sources was fixed at $m = 4$ and $n = 100$ observations of dimension $p = 3$ were simulated with known true parameter values

$$\Lambda_T = \begin{pmatrix} 5 & 0 & 5 & 0 \\ 5 & 0 & 0 & 5 \\ 0 & 5 & 5 & 0 \end{pmatrix},$$

$$R_T = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix},$$

and

$$\Psi_T = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}.$$

The observations were formed by generating a random s_i from $N(0, R_T)$, premultiplying it by Λ_T, and adding an error term generated randomly from $N(0, \Psi_T)$.

The hyperparameters were assessed according to the methods in the appendix to be $\nu = 206$, $b_0 = 1386$, $h_0 = 20$,

$$\Lambda_0 = \begin{pmatrix} 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \end{pmatrix},$$

$\eta = 1114.5$ and $\upsilon_0 = 115970$.

Upon applying the LSO estimation procedure, the posterior parameter estimates were found to be

$$\tilde{\Psi} = \begin{pmatrix} 5.5878 & 0.2801 & 0.2943 \\ 0.2801 & 5.5096 & -0.0373 \\ 0.2943 & -0.0373 & 5.5302 \end{pmatrix}$$

$$\tilde{R} = \begin{pmatrix} 102.5938 & -1.1721 & 2.7926 & 3.1392 \\ -1.1721 & 100.4824 & 3.1910 & 0.6295 \\ 2.7926 & 3.1910 & 102.6598 & -1.1864 \\ 3.1392 & 0.6295 & -1.1864 & 100.4450 \end{pmatrix}$$

$$\tilde{\Lambda} = \begin{pmatrix} 4.8892 & -0.4958 & 4.9022 & -0.5088 \\ 4.4900 & 0.2165 & -0.4012 & 5.1077 \\ -0.3808 & 5.1323 & 4.5253 & 0.2263 \end{pmatrix},$$

and the true sources, the mixed sources, along with the unmixed (estimated) sources are displayed in Figure 1, while in the true and unmixed sources are displayed in more detail in Figure 2 as dashed and solid lines respectively.

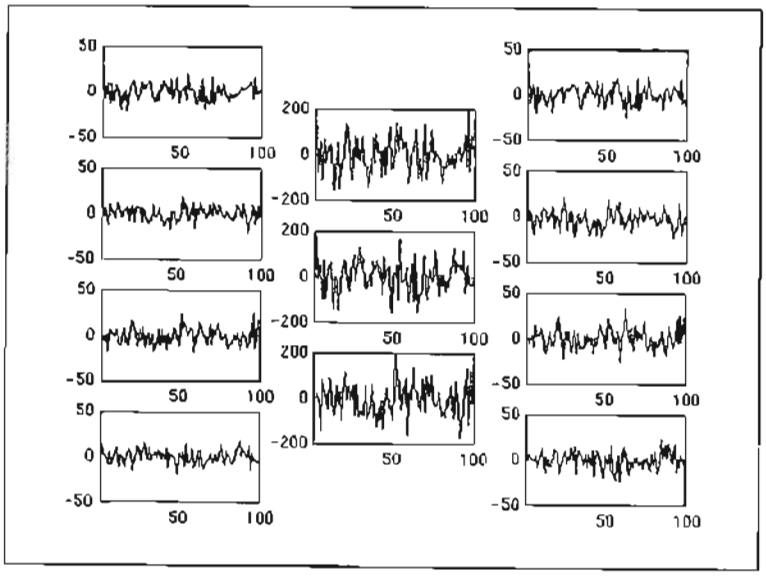

Figure 1. The Sources, Mixed Sources and Unmixed Sources

10. CONCLUSION

This paper has laid the foundation for a Bayesian blind source separation model and showed its relation to the PCA, ICA and FA models. The Bayesian blind separation of sources model, has several basic advantages over PCA, ICA, and FA.

Advantages over PCA:

(1) does not assume $\Psi = 0$
(2) does not assume $m = p$
(3) allows prior information to be incorporated
(4) does not assume that m is known.

Advantages over ICA:

(1) does not assume $\Psi = 0$
(2) does not assume $m = p$
(3) does not require Λ^{-1}

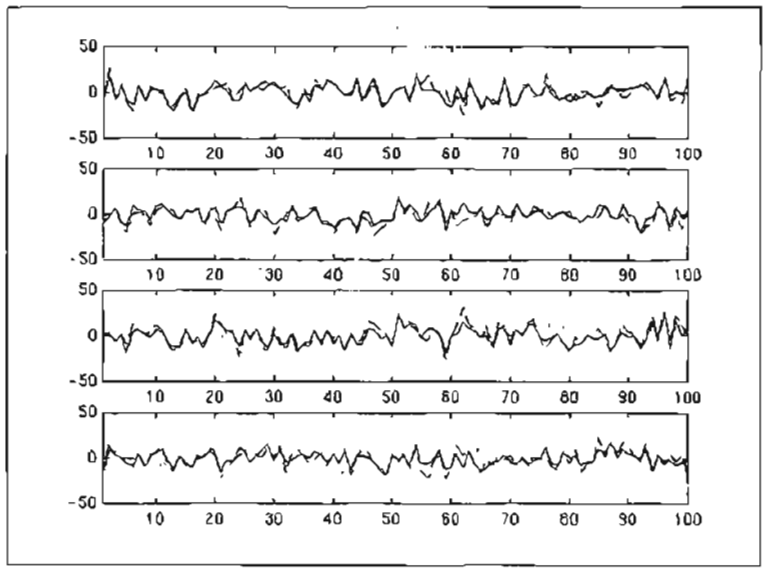

Figure 2. True Sources and Unmixed Sources

(4) does not constrain sources to be independent
(5) allows prior information to be incorporated
(6) does not assume that m is known.

Advantages over FA and Bayesian FA:

(1) does not assume $m \leq p$
(2) does not assume R is a correlation matrix
(3) does not require psychologic assumptions.

As shown in the example, Bayesian blind separation of sources appears to be very promising in solving the real "cocktail party problem".

Appendices

A. *Hyperparameter Assessment*

The hyperparameters of the prior distributions are assessed in this appendix. These hyperparameters are assessed in terms of questions asked to the substantive field expert.

m:

What range of values do you believe are possible for the number of sources?

$m_l, \ldots, m_u$:

What is your degree of belief for each of the possible values? If you had an urn with different colored balls in it corresponding to the different number of sources, how many of each color would be in the urn. For the example, the number of sources was assumed to be known.

For each m:

Ψ: ν, B

For simplicity of assessment, specify that $B = b_0 I_p$. With B diagonal, the mean and variance of any diagonal element of Ψ, Ψ_{jj} are

$$E(\Psi_{jj}) = \frac{b_0}{\nu - 2p - 2}$$

$$var(\Psi_{jj}) = \frac{2b_0^2}{(\nu - 2p - 2)^2(\nu - 2p - 4)}. \tag{A.1}$$

Solving for ν in the above system of equations,

$$\nu = \frac{[E(\Psi_{jj})]^2}{2[var(\Psi_{jj})]} + 2p + 4. \tag{A.2}$$

The unknowns are $E(\Psi_{jj})$ and $var(\Psi_{jj})$. Prior values for the mean and variance are to be elicited from the substantive field expert. In the example, the values assessed were assumed to be $E(\Psi_{jj}) = 7$ and $var(\Psi_{jj}) = 0.5$.

$(R \mid m)$: η, V

Similarly, assessment is simplified by specifying that $V = v_0 I_m$ and thus

$$\eta = \frac{[E(R_{kk})]^2}{2[var(R_{kk})]} + 2m + 4. \tag{A.3}$$

In the example, the values assessed were assumed to be $E(R_{kk}) = 105$ and $var(R_{kk}) = 20$.

Λ : Λ_0, $H = h_0 I_m$

As shown by Lee (1994) and Lee & Press (1998) in the area of Bayesian factor analysis, the posterior distribution is robust to values

of the hyperparameter h_0. Making an analogy for BBSS, it is assumed that the value $h_0 = 20$ was assessed from a substantive field expert.

The prior mean for the mixing matrix Λ_0 must be subjectively elicited from the substantive field expert or estimated from training data using another technique. In the example, it was assumed that a substantive field expert provided the prior mean for the mixing matrix.

B. *Bayesian Estimation Methods*

In this section we define some estimation procedures. The procedures are marginalization and conditional estimation, LSO and Gibbs sampling.

B.1. *Conditional Model Estimation*

Often we have a set of parameters, $\theta = (\theta_1, \ldots, \theta_J)$ in our posterior distribution $p(\theta \mid X)$. The marginal posterior distribution of any of the parameters, say θ_j can be obtained by integrating $p(\theta \mid X)$ with respect all parameters except θ_j. That is

$$p(\theta_j \mid X) = \int p(\theta_1, \ldots, \theta_J) d\theta_1 \ldots d\theta_{j-1}\, d\theta_{j+1} \ldots d\theta_J \tag{B.1}$$

where the integral is evaluated over the appropriate range of the parameters. It is possible to calculate the marginal posterior distribution for each of the parameters and calculate marginal posterior estimates such as the mean

$$\hat{\theta}_j = E(\theta_j \mid X) = \int \theta_j\, p(\theta_j \mid X) d\theta_j. \tag{B.2}$$

We may instead choose to compute conditional posterior distributions. If again $\theta = (\theta_1, \ldots, \theta_J)$, then the conditional distribution of any one of the parameters say θ_k given another say θ_j is given by

$$p(\theta_k \mid \theta_j, X) = \frac{p(\theta_k, \theta_j \mid X)}{p(\theta_j \mid X)} \tag{B.3}$$

where

$$p(\theta_k, \theta_j \mid X) = \int p(\theta_1, \ldots, \theta_J \mid X) d\theta_1 \ldots d\theta_{j-1}\, d\theta_{j+1} \ldots d\theta_{k-1}\, d\theta_{k+1} \ldots d\theta_J. \tag{B.4}$$

Now the conditional posterior mean (and mode) estimator may be computed such as

$$\hat{\theta}_k = E(\theta_k \mid \theta_j, X) = \int \theta_k \, p(\theta_k \mid \theta_j, X) d\theta_k, \tag{B.5}$$

or

$$\hat{\theta}_l = E(\theta_l \mid \theta_k, \theta_j, X) = \int \theta_l \, p(\theta_l \mid \theta_k, \theta_j, X) d\theta_l, \tag{B.6}$$

where

$$p(\theta_l \mid \theta_k, \theta_j, X) = \frac{P(\theta_l, \theta_k, \theta_j \mid X)}{p(\theta_k, \theta_j \mid X)}. \tag{B.7}$$

It is not always possible to obtain one or more marginal posterior distributions in an analytic closed form. For this reason Lindley/Smith optimization or Gibbs Sampling may be used.

B.2 *Lindley/Smith Optimization (LSO)*

Lindley/Smith optimization (Lindley and Smith, 1972) sometimes called iterated conditional modes (ICM) is a deterministic optimization method that finds the joint posterior modal estimators of $p(\theta \mid X)$ where θ denotes the vector of parameters, and X denotes the data.

Assume that $\theta = (\theta_1, \theta_2)$ where θ_1 and θ_2 are scalars and the posterior density of θ is $p(\theta_1, \theta_2 \mid X)$. We have a surface in 3-Dimensional space. We have θ_1 along one axis and θ_2 along the other with $p(\theta_1, \theta_2 \mid X)$ being the height of the surface or hill.

We want to find the top of the hill which is the same as finding the peak or maximum of the function $p(\theta_1, \theta_2 \mid X)$ with respect to both θ_1 and θ_2. Well we find the maximum of a surface by differentiating with respect to each variable (direction).

The maximum of the function $p(\theta_1, \theta_2 \mid X)$ satisfies

$$\frac{\partial}{\partial \theta_1} p(\theta_1, \theta_2 \mid X)\big|_{\theta_1=\tilde{\theta}_1} = \frac{\partial}{\partial \theta_2} p(\theta_1, \theta_2)\big|_{\theta_2=\tilde{\theta}_2} = 0, \tag{B.8}$$

which is the same as

$$\frac{\partial}{\partial \theta_1} p(\theta_1 \mid \theta_2, X) p(\theta_2 \mid X)\big|_{\theta_1=\tilde{\theta}_1} = \frac{\partial}{\partial \theta_2} p(\theta_2 \mid \theta_1, X) p(\theta_1 \mid X)\big|_{\theta_2=\tilde{\theta}_2} = 0 \tag{B.9}$$

or

$$p(\theta_2 \mid X) \frac{\partial}{\partial \theta_1} p(\theta_1 \mid \theta_2, X)\big|_{\theta_1=\tilde{\theta}_1} = p(\theta_1 \mid X) \frac{\partial}{\partial \theta_2} p(\theta_2 \mid \theta_1, X)\big|_{\theta_2=\tilde{\theta}_2} = 0 \tag{B.10}$$

assuming that $p(\theta_1 \mid X) \neq 0$ and $p(\theta_2 \mid X) \neq 0$.

We can obtain the posterior conditionals (functions) $p(\theta_1 | \theta_2, X)$ and $p(\theta_2 | \theta_1, X)$ along with their respective modes (maximum) $\tilde{\theta}_1 = \tilde{\theta}_1(\theta_2, X)$ and $\tilde{\theta}_2 = \tilde{\theta}_2(\theta_1, X)$.

We have the maximum of θ_1, $\tilde{\theta}_1$ for a given value of (conditional on) θ_2, and the maximum of θ_2, $\tilde{\theta}_2$ for a given value of (conditional on) θ_1.

The optimization procedure consists of

(1) Selecting an initial value for θ_2; call it $\tilde{\theta}_2^{(0)}$.
(2) Calculate the modal (maximal) value of $p(\theta_1 | \tilde{\theta}_2^{(0)}, X)$, $\tilde{\theta}_1^{(1)}$.
(3) Calculate the modal (maximal) value of $p(\theta_2 | \tilde{\theta}_1^{(1)}, X)$, $\tilde{\theta}_2^{(1)}$.
(4) Continue to calculate the remainder of the sequence $\tilde{\theta}_1^{(1)}, \tilde{\theta}_2^{(1)}, \tilde{\theta}_1^{(2)}, \tilde{\theta}_2^{(2)}, \ldots$ until convergence is reached.

If the posterior conditional distributions are not unimodal, we may converge to a local maximum and not the global maximum. If the posterior conditionals are unimodal, then we will always converge to a global maximum.

When convergence is reached, the point estimators $(\tilde{\theta}_1, \tilde{\theta}_2)$ are the maximum a posteori estimators.

This method can be generalized to more than two parameters. If θ is partitioned by $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ into J groups of parameters, we begin with a starting point $\tilde{\theta}^{(0)} = (\tilde{\theta}_1^{(0)}, \tilde{\theta}_2^{(0)}, \ldots, \tilde{\theta}_J^{(0)})$ and at the ith iteration define $\tilde{\theta}^{(i+1)}$ by

$$\tilde{\theta}_1^{(i+1)} = \tilde{\theta}_1(\tilde{\theta}_2^{(i)}, \tilde{\theta}_3^{(i)}, \ldots, \tilde{\theta}_J^{(i)}) \tag{B.11}$$

$$\tilde{\theta}_2^{(i+1)} = \tilde{\theta}_2(\tilde{\theta}_1^{(i+1)}, \tilde{\theta}_3^{(i)}, \ldots, \tilde{\theta}_J^{(i)}) \tag{B.12}$$

$$\vdots$$

$$\tilde{\theta}_J^{(i+1)} = \tilde{\theta}_1(\tilde{\theta}_2^{(i+1)}, \tilde{\theta}_3^{(i+1)}, \ldots, \tilde{\theta}_{J-1}^{(i+1)}) \tag{B.13}$$

at each step computing the maximum or mode. To apply this method we need to determine the functions $\tilde{\theta}_j$ which give the maximum of $p(\theta | X)$ with respect to $\tilde{\theta}_j$, conditional on the fixed values of all the other elements of θ. This is the general form of LSO.

B.3. *Gibbs Sampling*

Gibbs sampling is a stochastic method that draws random samples from the posterior conditional distribution for each of the parameters conditional on the fixed values of all the other parameters and the

data X. Let $p(\theta|X)$ be the posterior distribution of the parameters where θ is the set of parameters and X is the data. Let θ is partitioned by $\theta = (\theta_1, \theta_2, \ldots, \theta_J)$ into J groups of parameters. Ideally, we would like to perform the integration of the joint posterior distribution to obtain marginal posterior distributions

$$p(\theta_j|X) = \int p(\theta_1, \ldots, \theta_J) d\theta_1 \ldots d\theta_{j-1}\, d\theta_{j+1} \ldots d\theta_J \tag{B.14}$$

and marginal posterior mean estimates

$$E(\theta_j|X) = \int \theta_j\, p(\theta_j|X) d\theta_j. \tag{B.15}$$

Unfortunately, these integrations are usually of very high dimension and not available in a closed form. This is why we need the Gibbs sampling procedure. With the random samples drawn from the posterior conditional distributions, we can estimate the marginal posterior distributions and the marginal posterior means.

For the Gibbs sampling, we begin with an initial value

$$(\bar{\theta})^{(0)} = (\bar{\theta}_1^{(0)}, \bar{\theta}_2^{(0)}, \ldots, \bar{\theta}_J^{(0)})$$

and the ith iteration define

$$\bar{\theta}^{(i+1)} = (\bar{\theta}_1^{(i+1)}, \bar{\theta}_2^{(i+1)}, \ldots, \bar{\theta}_J^{(i+1)})$$

by the values from

$$\bar{\theta}_1^{(i+1)} = \text{a random sample from } p(\bar{\theta}_1|\bar{\theta}_2^{(i)}, \bar{\theta}_3^{(i)}, \ldots, \bar{\theta}_J^{(i)}, X) \tag{B.16}$$

$$\bar{\theta}_2^{(i+1)} = \text{a random sample from } p(\bar{\theta}_2|\bar{\theta}_1^{(i+1)}, \bar{\theta}_3^{(i)}, \ldots, \bar{\theta}_J^{(i)}, X) \tag{B.17}$$

$$\vdots$$

$$\bar{\theta}_J^{(i+1)} = \text{a random sample from } p(\bar{\theta}_J|\bar{\theta}_1^{(i+1)}, \bar{\theta}_2^{(i+1)}, \ldots, \bar{\theta}_{J-1}^{(i+1)}, X) \tag{B.18}$$

that is, at each step drawing a random sample from the conditional posterior distribution. To apply this method we need to determine the posterior conditionals of θ_j, conditional on the fixed values of all the other elements of θ and X from $p(\theta|X)$.

We will have $\bar{\theta}^{(1)}, \bar{\theta}^{(2)}, \ldots, \bar{\theta}^{(s+1)}, \ldots, \bar{\theta}^{(s+t)}$. The first s random samples called the "burn in" are discarded and the remaining t samples are kept.

The marginal posterior distributions (Equation B.14) are estimated by

$$\bar{p}(\theta_j) = \frac{1}{t}\sum_{k=1}^{t} p(\bar{\theta}_j^{(s+k)} \mid \bar{\theta}_1^{(s+k)}, \bar{\theta}_2^{(s+k)}, \ldots, \bar{\theta}_{j-1}^{(s+k)}, \bar{\theta}_{j+1}^{(s+k)}, \ldots, \bar{\theta}_J^{(s+k)}, X), \quad j = 1, \ldots, J \tag{B.19}$$

and the marginal posterior mean estimators of the parameters (Equation B.15) are estimated by $\bar{\theta} = (\bar{\theta}_1, \ldots, \bar{\theta}_J)$ where

$$\bar{\theta}_j = \frac{1}{t}\sum_{k=1}^{t} \bar{\theta}_j^{(s+k)}, \quad j = 1, \ldots, J. \tag{B.20}$$

C. *BBSS Gibbs Sampling*

The Gibbs sampling estimation procedure consists of starting with initial values for the parameters m, Λ, Ψ, S, and R say $\bar{m}_{(0)}$, $\bar{\Lambda}_{(0)}$, $\bar{\Psi}_{(0)}$, $\bar{S}_{(0)}$, and $\bar{R}_{(0)}$.

Then for a given number of sources $m = \bar{m}_{(l)}$ cycle through

$\bar{\Lambda}_{(l+1)} \equiv$ a random sample from $p(\Lambda \mid \bar{\Psi}_{(l)}, \bar{S}_{(l)}, \bar{R}_{(l)}, \bar{m}_{(l)}, X)$

$\bar{\Psi}_{(l+1)} \equiv$ a random sample from $p(\Psi \mid \bar{S}_{(l)}, \bar{\Lambda}_{(l+1)}, \bar{R}_{(l)}, \bar{m}_{(l)}, X)$

$\bar{S}_{(l+1)} \equiv$ a random sample from $p(S \mid \bar{\Psi}_{(l+1)}, \bar{\Lambda}_{(l+1)}, \bar{R}_{(l)}, \bar{m}_{(l)}, X)$

$\bar{R}_{(l+1)} \equiv$ a random sample from $p(R \mid \bar{\Psi}_{(l+1)}, \bar{S}_{(l+1)}, \bar{S}_{(l+1)}, \bar{m}_{(l)}, X)$

and for the given value for the number of sources $m = \bar{m}_{(l)}$ we have the sequence

$(\bar{\Lambda}_{(1)}, \bar{\Psi}_{(1)}, \bar{S}_{(1)}, \bar{R}_{(1)})$

$\vdots$

$(\bar{\Lambda}_{(u)}, \bar{\Psi}_{(u)}, \bar{S}_{(u)}, \bar{R}_{(u)})$

$(\bar{\Lambda}_{(u+1)}, \bar{\Psi}_{(u+1)}, \bar{S}_{(u+1)}, \bar{R}_{(u+1)})$

$\vdots$

$(\bar{\Lambda}_{(u+t)}, \bar{\Psi}_{(u+t)}, \bar{S}_{(u+t)}, \bar{R}_{(u+t)})$.

The first u random samples called the "burn in" are discarded and the remaining t samples are kept to be used for our estimates. We use the means of the remaining t random samples

$$\overline{\Lambda} = \frac{1}{t}\sum_{k=1}^{t} \overline{\Lambda}_{(u+k)}$$

$$\overline{\Psi} = \frac{1}{t}\sum_{k=1}^{t} \overline{\Psi}_{(u+k)}$$

$$\overline{S} = \frac{1}{t}\sum_{k=1}^{t} \overline{S}_{(u+k)}$$

$$\overline{R} = \frac{1}{t}\sum_{k=1}^{t} \overline{R}_{(u+k)}$$

as the sampling based marginal posterior mean estimates of the parameters for a given number of factors $m = \overline{m}_{(l)}$.

REFERENCES

1. T. W. Anderson and H. Rubin, Statistical inference in factor analysis, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1956, Vol. 5, pp. 345-357, University of California Press.
2. A. J. Bell and T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Computation*, Vol. 7, No. 6, pp. 1004-1034, 1995.
3. K. Knuth, Bayesian source separation and localization, in A. Mohammad-Djafari, editor, *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems*. San Diego, CA, pp. 147-158, July 1998.
4. K. Knuth, A Bayesian approach to source separation, in C. Jutten, J. -F. Cardoso and P. Loubaton, editors, *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA'99*, Aussios, France, pp. 283-288, 1999.
5. D. N. Lawley, The estimation of factor loadings by the method of maximum likelihood, *Proceedings of the Royal Society of Edinburgh*, Vol. 60, pp. 64-82, 1940.
6. S. E. Lee, *Robustness of Bayesian Factor Analysis Estimates*, Ph.D. thesis, Department of Statistics, University of California, Riverside, 1994.
7. S. E. Lee and S. J. Press, Robustness of Bayesian factor Analysis Estimates, *Communications in Statistics – Theory and Methods*, Vol. 27, No. 8, 1998.
8. D. J. C. MacKay, Maximum likelihood and covariant algorithms for independent component analysis, Working Paper, Draft 3.7, University of Cambridge, Cavindish Laboratory,-Cambridge, UK, 1996.
9. A. Mohammad-Djafari, A Bayesian estimation method for detection, localisation and estimation of superposed sources in remote sensing, in *SPIE 97 annual meeting*, (San Diego, CA, USA, July 27-Aug. 1, 1997), 1997.

10. A. Mohammad-Djafari, A Bayesian approach to source separation, in *Proceedings of The Nineteenth International Conference on Maximum Entropy and Bayesian Methods.* (August 2-6, 1999, Boise, ID.), 1999.

11. S. J. Press, *Applied Multivariate Analysis. Using Bayesian and Frequentist Methods of Inference,* Robert E. Krieger Publishing Company, Malabar, Florida, 1982.

12 S. J. Press, *Bayesian Statistics: Principles, Models and Applications,* John Wiley and Sons, New York, 1989.

13. S. J. Press and K. Shigemasu, Bayesian inference in factor analysis, in *Contributions to Probability and Statistics,* Chapter 15, Springer-Verlag, 1989.

14. S. J. Press and K. Shigemasu, Bayesian inference in factor analysis – Revised, Tech. Rep. No. 243, Department of Statistics, University of California, Riverside, May 1997.

15. S. J. Press and K. Shigemasu, Posterior distribution for the number of factors, Tech. Rep. No. 208, Department of Statistics, University of California, Riverside, April 1994.

16. D. B. Rowe and S. J. Press, Gibbs sampling and hill climbing in Bayesian factor analysis, Tech. Rep. No. 255, Department of Statistics, University of California, Riverside, May 1998.

17. D. B. Rowe, Correlated Bayesian Factor Analysis, Ph.D. thesis, Department of Statistics, University of California, Riverside, 1998.

Received April, 2000