2006 International Joint Conference on Neural Networks
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Factor Analysis for Geophysical Signal Processing with Seismic Profiles

Zhenhai Wang, *Student Member, IEEE*, and C.H. Chen, *Fellow, IEEE*

***Abstract*— In the petroleum industry, stacking, one of the principal steps of conventional seismic signal processing, plays an important role in enhancing events and cancelling random and coherent noises by utilizing the predesigned redundancy in the seismic data. This paper demonstrates that by applying an alternative technique, Factor Analysis, to the same dataset, better subsurface image of the earth can be obtained. Contrary to stacking, it takes into consideration the scaling of the latent signal and makes explicit use of the second order statistics, obtaining higher Signal-to-Noise Ratio. Moreover, Factor Analysis is compared with Principal Component Analysis and Independent Component Analysis, which can both be realized by neural networks, in processing the synthetic Marmousi dataset.**

I. Introduction to seismic signal processing

Formed millions of years ago from plants and animals that died and decomposed beneath soil and rock, fossil fuels, namely, coal and petroleum, due to their low cost availability, will remain the most important energy resource for at least another few decades. Ongoing petroleum research continues to focus on science and technology needs for increased petroleum exploration and production. The petroleum industry relies heavily on subsurface imaging techniques for the location of these hydrocarbons.

Due to their target-oriented capability, generally good imaging results, and computational efficiency, seismic reflection profiling becomes the principal method by which the petroleum industry explores for hydrocarbon-trapping structures. It works by processing echoes of seismic waves from boundaries between different earth subsurfaces that characterize different acoustic impedances. Depending upon the geometry of surface observation points and source locations, the survey is called 2D or 3D seismic survey. Figure 1 shows a typical 2D seismic survey, during which, a cable with attached receivers at regular intervals is dragged by a boat. The source moves along the predesigned seismic lines and generates seismic waves at regular intervals such that points in the subsurfaces are sampled several times by the receivers, producing a series of seismic traces. These seismic traces are saved on magnetic tape or hard disks in the recording boat for future processing.

There is a well-established sequence for standard seismic data processing. Deconvolution, stacking, and migration are the three principle processes, among which common-midpoint stacking is the most robust of all. Utilizing redundancy in CMP recording, stacking can significantly suppress uncorrelated noise, thereby increasing the Signal-to-Noise Ratio (SNR). It also can attenuate a large part of the coherent noise in the data, such as guided waves and multiples.

Zhenhai Wang and C.H. Chen are both with the Department of Electrical and Computer Engineering, University of Massachusetts, North Dartmouth, MA 02747-2300, USA (phone: 508-999-8475; fax: 508-999-8489 ; email: cchen@umassd.edu).

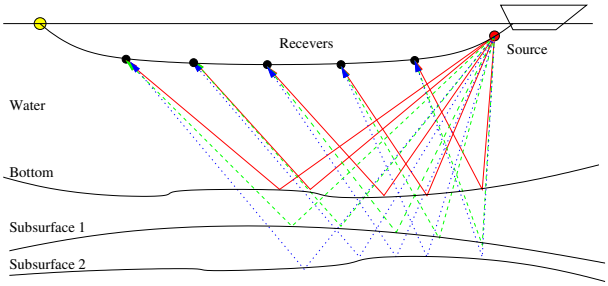

Fig. 1. A typical 2D seismic survey

II. Factor Analysis framework

Being a branch of multivariate analysis, Factor Analysis (FA) is concerned with the internal relationships of a set of variates [1]. It offers a conceptual framework within which many disparate methods can be unified and a base from which new methods can be developed.

A. General FA model

In Factor Analysis the basic model is

$$\mathbf{x} = A\mathbf{s} + \mathbf{n} \tag{1}$$

where

$\mathbf{x} = (x_1, x_2, ..., x_p)^T$: test scores,
$\mathbf{s} = (s_1, s_2, ..., s_r)^T$: $r < p$ common factor scores,
A : factor loadings,
$\mathbf{n} = (n_1, n_2, ..., n_p)^T$: order p unique factor scores.

The following assumptions are usually made for the factor model [2]:

(1) rank $(A) = r < p$.
(2) $E(\mathbf{x}|\mathbf{s}) = A\mathbf{s}$.
(3) $E(\mathbf{x}\mathbf{x}^T) = \Sigma$, $E(\mathbf{s}\mathbf{s}^T) = \Omega$ and

$$\Psi = E(\mathbf{n}\mathbf{n}^T) = \begin{bmatrix} \sigma_1^2 & & & \mathbf{0} \\ & \sigma_2^2 & & \\ & & \ddots & \\ \mathbf{0} & & & \sigma_p^2 \end{bmatrix}.$$

That is, the errors are assumed to be uncorrelated. The common factors however are generally correlated, and

Ω, the covariance matrix of the factors, is therefore not necessarily diagonal.

(4) $E\left(\mathbf{s}\mathbf{n}^T\right) = \mathbf{0}$ so that the errors and common factors are uncorrelated.

From the above assumptions, we have,

$$\begin{aligned} E\left(\mathbf{x}\mathbf{x}^T\right) &= E\left[\left(A\mathbf{s}+\mathbf{n}\right)\left(A\mathbf{s}+\mathbf{n}\right)^T\right] \\ &= E\left(A\mathbf{s}\mathbf{s}^TA^T + A\mathbf{s}\mathbf{n}^T + \mathbf{n}\mathbf{s}^TA^T + \mathbf{n}\mathbf{n}^T\right) \\ &= A\Omega A^T + E\left(\mathbf{n}\mathbf{n}^T\right) \\ &= \Gamma + \Psi \end{aligned} \quad (2)$$

where $\Gamma = A\Omega A^T$ and $\Psi = E\left(\mathbf{n}\mathbf{n}^T\right)$ are the true and error covariance matrices, respectively.

In addition, if postmultiplying Equation (1) by $\mathbf{s}^T$, taking the expectation, and using assumptions (3) and (4), we have

$$\begin{aligned} E\left(\mathbf{x}\mathbf{s}^T\right) &= E\left(A\mathbf{s}\mathbf{s}^T + \mathbf{n}\mathbf{s}^T\right) \\ &= AE\left(\mathbf{s}\mathbf{s}^T\right) + E\left(\mathbf{n}\mathbf{s}^T\right) \\ &= A\Omega. \end{aligned}$$

For the special case of $\Omega = I$, the covariance between the observation and the latent variables simplifies to $E\left(\mathbf{x}\mathbf{s}^T\right) = A$.

When $\mathbf{x}$ is multivariate Gaussian, the second moments of Equation (2) will contain all the information concerning the factor model. The factor model Equation (1) will be linear, and given the factors $\mathbf{s}$, the variables $\mathbf{x}$ are conditionally independent. Let $\mathbf{s} \sim \mathcal{N}\left(\mathbf{0}, I\right)$, the conditional distribution of $\mathbf{x}$ is

$$\mathbf{x}|\mathbf{s} \sim \mathcal{N}\left(A\mathbf{s}, \Psi\right) \quad (3)$$

with conditional independence following from the diagonality of Ψ. The common factors $\mathbf{s}$ therefore reproduce all covariances (or correlations) between the variables, but account for only a portion of the variance.

B. FA algorithms

Many methods have been developed for estimating the model parameters. Unweighted Least Squares (ULS) [3] algorithm is based on minimizing the sum of squared differences between observed and estimated correlation matrices, not counting the diagonal. Generalized least squares (GLS) [3] algorithm is adjusting ULS by weighting the correlations inversely according to their uniqueness. Another method, Maximum Likelihood (ML) algorithm [4], uses a linear combination of variables to form factors, where the parameter estimates are those most likely to have resulted in the observed correlation matrix. More details on Maximum Likelihood algorithm can be found in Appendix B.

Those methods are all second-order methods which find the representation using only the information contained in the covariance matrix of the test scores. In most cases, the mean is also used in the initial centering. The reason for the popularity of the second-order methods is that they are computationally simple, often requiring only classical matrix manipulations.

In contrast to second-order methods, most higher-order methods try to find a meaningful representation. Higher-order methods use information on the distribution of $\mathbf{x}$ that is not contained in the covariance matrix. The distribution of $\mathbf{x}$ must not be assumed to be Gaussian, because all the information of Gaussian variables is contained in the first two order statistics from which all the high statistics can be generated. However, for more general families of density functions, the representation problem has more degrees of freedom, and much more sophisticated techniques may be constructed for non-Gaussian random variables.

C. Within the framework

Principal component analysis (PCA), Independent Component Analysis (ICA) and Independent Factor Analysis (IFA) can be considered within the Factor Analysis framework.

1) Principal component analysis: Principal component analysis (PCA) also known as the Hotelling transform or the Karhunen-Loève transform. It is widely used in signal processing, statistics, and neural computing to find the most important directions in the data in the mean-square sense. It is the solution of the FA problem with minimum mean square error and an orthogonal weight matrix.

The basic idea of PCA is to find the $r \leq p$ linearly transformed components that explain the maximum amount of variance possible, which, accordingly, may then, be used to reduce the dimensionality of the original data for further analysis. However, all components are needed to reproduce accurately the correlation coefficients within $\mathbf{x}$.

The primary task in PCA is to reduce the dimension of the data. In fact, it can be proven that the representation given by PCA is an optimal linear dimension reduction technique in the mean-square sense [5][6]. The kind of reduction in dimension has important benefits [7]. First, the computational complexity of the further processing stages is reduced. Second, noise may be reduced, as the data not contained in the components may be mostly due to noise. Third, projecting into a subspace of low dimension is useful for visualizing the data.

2) Independent Component Analysis: The Independent Component Analysis (ICA) originates from the multi-input and multi-output (MIMO) channel equalization and several algorithms are derived from a neural networks viewpoint [8]. Its two most important applications are Blind Source Separation (BSS) and feature extraction. The mixing model of ICA is similar to that of the FA, but in the basic case without the noise term. Such an A is searched for that the components $\mathbf{s} = A^{-1}\mathbf{x}$ would be as independent as possible. The concept of ICA may be seen as an extension of PCA, which can only impose independence up to the second order and, therefore, defines directions that are orthogonal. In practice, the independence can be maximized e.g. by maximizing non-Gaussianity of the components or minimizing mutual information [9]. In some extensions the number of independent components can exceed the number of dimensions of the observations making the basis overcomplete [9] [10]. ICA can also be viewed as a generative model when the one

dimensional distributions for the components are modelled with for example mixtures of Gaussians (MoG).

The problem with ICA is that it has the ambiguities of scaling and permutation, i.e., the indetermination of the variances and order of the independent components.

3) Independent Factor Analysis: Independent Factor Analysis (IFA) aims to describe p generally correlated observed variables $\mathbf{x}$ in terms of $r < p$ independent latent variables $\mathbf{s}$ and an additive noise term $\mathbf{n}$. The proposed algorithm [11] derives from the Maximum Likelihood (ML) and more specifically from the Expectation-Maximization (EM) algorithm.

IFA model differs from the classic FA model in that the properties of the latent variables it involves are different. The noise variables $\mathbf{n}$ are assumed to be normally distributed, not necessarily uncorrelated. The latent variables in $\mathbf{s}$ are assumed to be mutually independent but not necessarily normally distributed; their densities are indeed modeled as mixtures of Gaussians. The independence assumption allows to model the density of each s_i in the latent space separately.

There are some problems with the EM-MoG algorithm. First, approximating source densities with MoGs is not so straightforward because the number of Gaussians has to be decided and the parameters have to be adjusted. Second, EM-MoG is computationally demanding where the complexity of computation grows exponentially with the number of sources [11]. Given a small number of sources the EM algorithm is exact and all the required calculations can be done analytically, whereas, it becomes intractable as the number of sources in the model increases.

III. Simulation

Now we suggest an alternative way of obtaining the subsurface image by using Factor Analysis instead of stacking, based on the redundancy concept that all the traces in one CMP gather, after preprocessing, correspond to the same signal embedded in different random noises. Factor Analysis is able to extract one unique common factor from the traces with maximum correlation among them. It fits well with the goal of stacking.

To illustrate the idea, $\mathbf{x(t)}$ are generated using the following equation:

$$\begin{aligned} \mathbf{x(t)} &= As(t) + \mathbf{n(t)} \\ &= A\cos(2\pi t) + \mathbf{n(t)}. \end{aligned}$$

where $s(t)$ is the sinusoidal signal, $\mathbf{n(t)}$ are 10 independent noise terms with Gaussian distribution. The matrix of factor loadings A is also generated randomly. Figure 2 shows the results of stacking and Factor Analysis. The top plot is one of the 10 observations $\mathbf{x(t)}$. The middle plot is the result of stacking and the bottom plot is the result of Factor Analysis using Maximum Likelihood algorithm. Comparing the two plots suggests that Factor Analysis outperforms stacking in improving the SNR of the component extracted.

There are two reasons that Factor Analysis works better than stacking. There are two reasons for this. First, Factor Analysis model considers scaling factor A while stacking assumes no scaling as shown below in Equation (4) and (5).

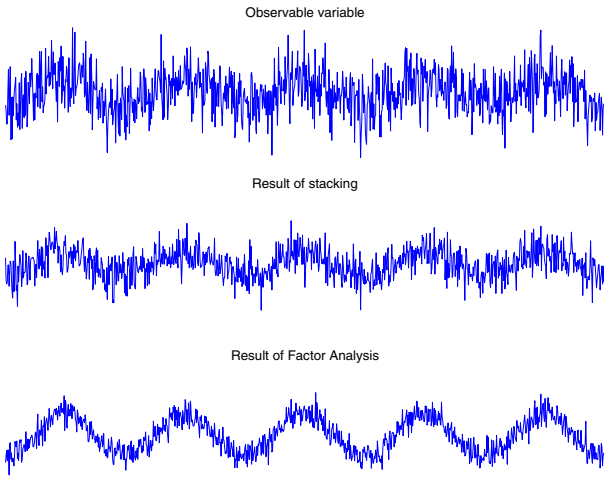

Fig. 2. Comparison of stacking and Factor Analysis

$$\text{Factor Analysis:} \quad \mathbf{x} = As + \mathbf{n} \tag{4}$$

$$\text{Stacking:} \quad \mathbf{x} = s + \mathbf{n}. \tag{5}$$

When the scaling information is lost, simple summation does not necessarily increase SNR. For example, if one scaling factor is 1 and the other is -1, summation will simply cancel out the signal component completely, leaving only the noise component. Second, Factor Analysis makes use of the second order statistics explicitly (See Appendix B) as the criterion to extract the signal while stacking does not. Therefore, SNR will improve more in the case of Factor Analysis than in the case in stacking.

IV. Main Results

A. Factor Analysis vs. stacking

The dataset used here is the Marmousi dataset, which is a 2-D synthetic dataset generated at the Institut Françis du Pétrole [12] [13]. The geometry and velocity model were created to produce a complex seismic data which requires advanced processing techniques to obtain a correct earth image.

Simulation result in the above section suggests that Factor Analysis be applied to the pre-stack seismic data. However, in Factor Analysis, while the choice of the number of factor scores is subjected to an upper bound (Refer to Appendix A for more details), the actual choice is not straightforward and subject to controversy [2]. Using the Latent Root criterion or the Scree Test criterion will give us an upper bound of 4. In this work, the number of factors is automatically determined as 1 based on the CMP redundancy concept. Besides, in the chi-squared test, the Right-tail significance level for the null hypothesis of 1 single common factor is very low, which means the test fails to reject the null hypothesis of 1 common factors, suggesting that this model provides a satisfactory explanation of the covariation in these data.

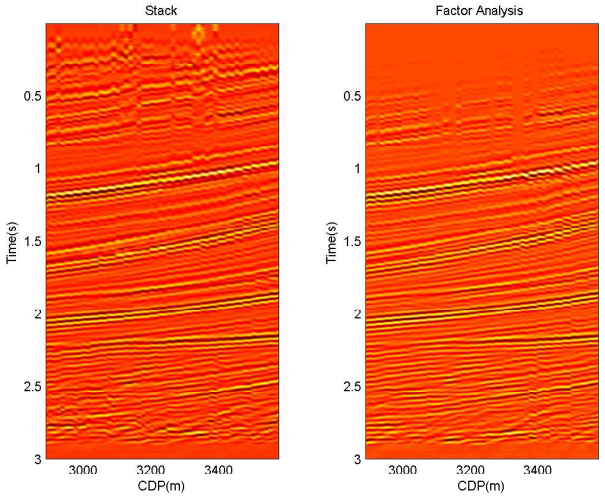

Fig. 3. Comparison of stacking and FA result

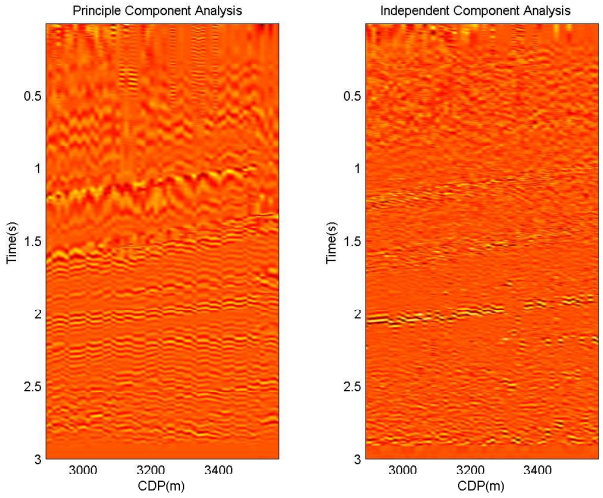

Fig. 4. Results of PCA and ICA

In conventional stacking, because of the approximation distortion, almost all the traces will have a segment set to zero (muting). In order to fully utilize all the data points available, we suggest applying Factor Analysis to the pre-stack traces without muting. All those traces that have the non-distorted segments is included in the Factor Analysis. For example, all 48 traces in a CMP gather have the non-distorted 2.3s to 3s segment. Factor analysis will be applied to those 48 traces and the segment 2.3s to 3s will be extracted from the result. While only half of the traces have the non-distorted 1.35s to 1.4s segment, Factor Analysis will be only applied to those 24 traces. Therefore, all the segments are extracted from where the pre-stack traces have virtually no approximation distortion. After putting all the segments together in series, the result of Factor Analysis is obtained, as shown on the right plot in Figure 3.

The stacking result is placed on the left for comparison purpose. Compare the results of Factor Analysis and stacking, we can see that events at around 1 second and 1.5 second are strengthened. Events from 2.2s to 3.0s are more smoothly presented instead of the broken dash-like events in the stacked result. Overall, the SNR of the image is improved.

It needs to be pointed out that the plots are after the automatic amplitude adjustment which is to stress the vague events so that both the vague events and strong events in the image are shown with approximately the same amplitude. The algorithm includes 3 easy steps:

(1) Compute Hilbert envelope of a trace.
(2) Convolve the envelope with a triangular smoother to produce the smoothed envelop.
(3) Divide the trace by the smoothed envelope to produce the amplitude-adjusted trace.

It is also noted that due to lack of data at small offset after pre-stack processing, events before 0.2s are shown as distorted and do not provide useful information.

B. Factor Analysis vs. PCA and ICA

The results of PCA and ICA are put side by side in Figure 4. As we can see in both plots important events are missing, the subsurface images are distorted. The reason is that the criteria used in PCA and ICA to extract the signals are improper to this particular scenario as discussed in sections II-C.1 and II-C.2. In PCA, traces are transformed linearly and orthogonally into an equal number of new traces that have the property of being uncorrelated, where the first component having the maximum variance instead of covariance is used to produce the image. In ICA, the algorithm tries to extract components that are as independent to each other as possible, where the obtained components suffer from the problems of scaling and permutation.

V. Conclusions

Stacking is one of the three most important and robust processing steps in seismic signal processing. By utilizing the redundancy of the CMP gathers, stacking can effectively remove noise and increase the Signal-to-Noise Ratio. In this paper we propose to use Factor Analysis to replace stacking to obtain better subsurface images after applying Factor Analysis algorithm to the synthetic Marmousi dataset. Comparisons with PCA and ICA show that Factor Analysis indeed have advantages over these two techniques in this scenario.

It is noted that the conventional seismic processing steps adopted here are very basic and for illustrative purposes only. Better result may be obtained in velocity analysis and stacking if careful examination and iterative procedures are incorporated as is often the case in real situations.

Appendix

A. Upper bound of the number of common factors

In equation $E(\mathbf{x}\mathbf{x}^T) = \Sigma = \Gamma + \Psi$, if Ψ is unique, matrix $\Sigma - \Psi$ must be of rank r. This is the covariance matrix for $\mathbf{x}$ where each diagonal element represents that part of the

variance which is due to the r common factors instead of the total variance of the corresponding variate. This is know as communality of the variate.

When $r = 1$, A reduces to a column vector of p elements. It is unique, apart from a possible change of sign of all its elements.

With $1 < r < p$ common factors, it is not generally possible to determine A and $\mathbf{s}$ uniquely, even in the case of a normal distribution. Although every factor model specified by Equation (3) leads to a multivariate normal, the converse is not necessarily true when $1 < r < p$. The difficulty is known as the factor identification or factor rotation problem.

Let H be any $(r \times r)$ orthogonal matrix, so that $HH^T = H^T H = I$, then

$$\begin{aligned} \mathbf{x} &= AHH^T\mathbf{s} + \mathbf{n} \\ &= A^\star \mathbf{s}^\star + \mathbf{n}. \end{aligned}$$

Thus, $\mathbf{s}$ and $\mathbf{s}^\star$ have the same statistical properties since

$$\begin{aligned} E(\mathbf{s}^\star) &= H^T E(\mathbf{s}) \\ cov(\mathbf{s}^\star) &= H^T cov(\mathbf{s}) H = H^T H = I. \end{aligned}$$

Suppose there exist $1 < r < p$ common factors such that $\Gamma = A\Omega A^T$ and Ψ is Grammian and diagonal. The covariance matrix Σ has $C\begin{pmatrix} p \\ 2 \end{pmatrix} + p = 1/2p(p+1)$ distinct elements, which equals the total number of normal equations to be solved. However, the number of solutions is infinite, as can be seen from the following derivation. Since Ω is Grammian, its Cholesky decomposition exists. That is, there exists a non-singular $(r \times r)$ matrix U, such that $\Omega = U^T U$ and

$$\begin{aligned} \Sigma &= A\Omega A^T + \Psi \\ &= AU^T U A^T + \Psi \\ &= \left(AU^T\right)\left(AU^T\right)^T + \Psi \\ &= A^\star A^{\star T} + \Psi. \end{aligned} \quad \text{(A-1)}$$

Apparently both factorizations Equation (2) and Equation (A-1) result in the same residual error Ψ and therefore must represent equally valid factor solutions. Also, we can substitute $A^\star = AB$ and $\Omega^\star = B^{-1}\Omega\left(B^T\right)^{-1}$, which again yields a factor model that is indistinguishable from Equation (2). Therefore, no sample estimator can distinguish between such an infinite number of transformations. Consequently, the coefficients A equals to $A^\star$ statistically and cannot be distinguished from it. Both the transformed and untransformed coefficients, plus Ψ, generate Σ in exactly the same way and cannot be differentiated by any estimation procedure without the introduction of additional restrictions.

To solve the rotational indeterminacy of the factor model restrictions on Ω are required, with $\Omega = I$ being the most straightforward and common one. The number m of free parameters implied by the equation

$$\Sigma = AA^T + \Psi \quad \text{(A-2)}$$

is then equal to the total number $pr + p$ for unknown parameters in A and Ψ, minus the number of zeros restrictions placed on the off-diagonal elements of Ω, which is equal to $1/2\left(r^2 - r\right)$ with Ω being symmetric. We then have

$$\begin{aligned} m &= (pr + p) - 1/2\left(r^2 - r\right) \\ &= p(r+1) - 1/2\left(r^2 - r\right) \end{aligned} \quad \text{(A-3)}$$

where the columns of A are assumed to be orthogonal. The number of degrees of freedom d is then given by the number of equations implied by Equation (A-2), i.e., the number of distinct elements in Σ minus the number of free parameters m.

$$\begin{aligned} d &= 1/2p(p+1) - \left[p(r+1) - 1/2\left(r^2 - r\right)\right] \\ &= 1/2\left[(p-r)^2 - (p-r)\right] \end{aligned} \quad \text{(A-4)}$$

which for a nontrivial empirical solution must be strictly positive, places an upper bound on the number of common factor r which may be obtained in practice, a number which is generally smaller than the number of variables p.

B. Maximum Likelihood algorithm

The Maximum Likelihood (ML) algorithm presented here is proposed by Jöreskog [4]. The algorithm uses a iterative procedure to compute a linear combination of variables to form factors. Assume that the random vector $\mathbf{x}$ has a multivariate normal distribution as defined in Equation (3). The elements of A, Ω and Ψ are the parameters of the model to be estimated from the data. Suppose that from a random sample of N observations of $\mathbf{x}$ we find the matrix the estimated covariance matrix $\tilde{\Sigma}$ whose elements are the usual estimates of variances and covariances of the components of $\mathbf{x}$.

$$\begin{aligned} \tilde{\mathbf{m}}_\mathbf{x} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \\ \tilde{\Sigma} &= \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x} - \tilde{\mathbf{m}}_\mathbf{x})(\mathbf{x} - \tilde{\mathbf{m}}_\mathbf{x})^T \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N} \mathbf{x}\mathbf{x}^T - N\tilde{\mathbf{m}}_\mathbf{x}\tilde{\mathbf{m}}_\mathbf{x}^T\right). \end{aligned} \quad \text{(B-1)}$$

The distribution of $\tilde{\Sigma}$ is the Wishart distribution [1]. The log-likelihood function is given by

$$\log L = -\frac{1}{2}(N-1)\left[\log|\Sigma| + tr\left(\tilde{\Sigma}\Sigma^{-1}\right)\right].$$

However, it is more convenient to minimize

$$F(A, \Omega, \Psi) = \log|\Sigma| + tr\left(\tilde{\Sigma}\Sigma^{-1}\right) - \log\left|\tilde{\Sigma}\right| - p$$

instead of maximizing $\log L$ [4]. They are equivalent because $\log L$ is a constant minus $\frac{1}{2}(N-1)$ times F. The function F is regarded as a function of A and Ψ. Note that if H is any nonsingular $(k \times k)$ matrix, then

$$F\left(AH^{-1}, H\Omega H^T, \Psi\right) = F(A, \Omega, \Psi)$$

which means that the parameters in A and Ω are not independent of another, and in order to make the Maximum

Likelihood estimates of A and Ω unique, k^2 independent restrictions must be imposed on A and Ω.

To find the minimum of F we shall first find the conditional minimum for a given Ψ and then find the overall minimum. The partial derivative of F with respect to A is

$$\frac{\partial F}{\partial A} = 2\Sigma^{-1}\left(\Sigma - \tilde{\Sigma}\right)\Sigma^{-1}A.$$

See details in [4]. For a given Ψ the minimization of A is to be found among the solution of

$$\Sigma^{-1}\left(\Sigma - \tilde{\Sigma}\right)\Sigma^{-1}A = 0.$$

Premultiplying with Σ gives

$$\left(\Sigma - \tilde{\Sigma}\right)\Sigma^{-1}A = 0.$$

Using the following expression for the inverse Σ^{-1} [1]

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1}A\left(I + A^T\Psi^{-1}A\right)^{-1}A^T\Psi^{-1}. \quad \text{(B-2)}$$

whose left side may be further simplified [4] so that

$$\left(\Sigma - \tilde{\Sigma}\right)\Psi^{-1}A\left(I + A^T\Psi^{-1}A\right)^{-1} = 0.$$

Postmultplying by $I + A^T\Psi^{-1}A$ gives

$$\left(\Sigma - \tilde{\Sigma}\right)\Psi^{-1}A = 0 \quad \text{(B-3)}$$

which after substitution of Σ from Equation (A-2) and rearrangement of terms gives

$$\tilde{\Sigma}\Psi^{-1}A = A\left(I + A^T\Psi^{-1}A\right).$$

Premultiplying by $\Psi^{-1/2}$ finally gives

$$\begin{aligned}&\left(\Psi^{-1/2}\tilde{\Sigma}\Psi^{-1/2}\right)\left(\Psi^{-1/2}A\right)\\ &= \left(\Psi^{-1/2}A\right)\left(I + A^T\Psi^{-1}A\right). \quad \text{(B-4)}\end{aligned}$$

From Equation (B-4), we can see that it is convenient to take $A^T\Psi^{-1}A$ to be diagonal, since F is unaffected by postmultipication of A by an orthogonal matrix and $A^T\Psi^{-1}A$ can be reduced to diagonal form by orthogonal transformations [14]. In this case, Equation (B-4) is an standard Eigen Decomposition form. The columns of $\Psi^{-1/2}A$ are latent vectors of $\Psi^{-1/2}\tilde{\Sigma}\Psi^{-1/2}$, and the diagonal elements of $I + A^T\Psi^{-1}A$ are the corresponding latent roots. Let $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_p$ be the latent roots of $\Psi^{-1/2}\tilde{\Sigma}\Psi^{-1/2}$ and let $\tilde{e}_1, \tilde{e}_2, \cdots, \tilde{e}_k$ be a set of latent vectors corresponding to the k largest roots. Let $\tilde{\Lambda}_k$ be the diagonal matrix with $\tilde{\lambda}_1, \tilde{\lambda}_2, \cdots, \tilde{\lambda}_k$ as diagonal elements and let $\tilde{E}_k$ be the matrix with $\tilde{e}_1, \tilde{e}_2, \cdots, \tilde{e}_k$ as columns. Then

$$\Psi^{-1/2}\tilde{A} = \tilde{E}_k\left(\tilde{\Lambda}_k - I\right)^{1/2}.$$

Premultiplying by $\Psi^{1/2}$ gives the conditional Maximum Likelihood estimate of A as

$$\tilde{A} = \Psi^{1/2}\tilde{E}_k\left(\tilde{\Lambda}_k - I\right)^{1/2}. \quad \text{(B-5)}$$

Up to now, we have considered the minimization of F with respect to A for a given Ψ. Now let's examine the partial derivative of F with respect to Ψ [1],

$$\frac{\partial F}{\partial \Psi} = diag\left[\Sigma^{-1}\left(\Sigma - \tilde{\Sigma}\right)\right]\Sigma^{-1}.$$

Substituting $\tilde{\Sigma}^{-1}$ with Equation (B-2) and use Equation (B-3) gives

$$\frac{\partial F}{\partial \Psi} = diag\left[\Psi^{-1}\left(\Sigma - \tilde{\Sigma}\right)\right]\Psi^{-1}$$

which by Equation (2) becomes

$$\frac{\partial F}{\partial \Psi} = diag\left[\Psi^{-1}\left(\tilde{A}\tilde{A}^T + \Psi - \tilde{\Sigma}\right)\right]\Psi^{-1}.$$

Minimzing it, we will get,

$$\tilde{\Psi} = diag\left(\tilde{\Sigma} - \tilde{A}\tilde{A}^T\right). \quad \text{(B-6)}$$

By iterating Equation (B-5) and Equation (B-6), the Maximum Likelihood Estimation of Factor Analysis model of Equation (1) can be obtained.

REFERENCES

[1] D. N. Lawley and A. E. Maxwell. *Factor analysis as a stastical method*. London: Butterworths, 1963.

[2] Alexander T. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley-Interscience, 1st edition, June 1994.

[3] H. Harman. *Modern Factor Analysis*. University of Chicago Press, 2nd edition, 1967.

[4] K. G. Jöreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32:443–482, 1967.

[5] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[6] M. Kendall. *Multivariate Analysis*. Charles Griffin & Co., 1975.

[7] Aapo Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[8] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[9] Juha Karhunen Hyvärinen Aapo and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

[10] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6:87–90, 1999.

[11] H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1998.

[12] R. J. Versteeg. Sensitivity of prestack depth migration to the velocity model. *Geophysics*, 58(6):873–882, 1993.

[13] R. J. Versteeg. The marmousi experience: Velocity model determination on a synthetic complex data set. *The Leading Edge*, (13):927–936, 1994.

[14] R. Bellman. *Introduction to matrix analysis*. New York: McGraw-Hill, 1960.