

A Re-examination of some Popular Latent Factor Estimation Methods

Alvin L. Stroyny, PhD
ALS Consulting
PO Box 583
Grafton WI 53024
astroyny@ix.netcom.com
Phone: (414) 405-3717

and

Daniel B. Rowe, PhD
Biophysics Research Institute
The Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI 53226
Phone: (414) 456-4027
Fax: (414) 456-6512
Email: dbrowe@mcw.edu

Abstract

We present a brief overview of several popular approaches for estimating latent factor models of security returns, highlighting the similarities and differences of each. The various methods are recast in a Bayesian estimation framework within the context of the EM algorithm. The EM algorithm results in a simple, unified presentation of the various methods in which differences in the implicit underlying assumptions of each method are readily evident. These differences typically involve either 1) the assumed distributional form of the common factors or 2) the restrictions on the form of the variance of the unique factors. In addition, the use of a Bayesian-EM framework results in a solid theoretical foundation which allows for many extensions of the standard factor model to be easily incorporated.

Since the Arbitrage Pricing Theory, (APT), was introduced by Ross (1977), there has been a great deal of interest in estimating multi-factor models of security returns on the part of academics and practitioners alike. Three of the more popular approaches to building (linear) multi-factor models include 1) (time series) regressions of security returns on innovations in *economic* data series, 2) (cross-sectional) regressions on industry-standardized *fundamental* accounting data, as well as 3) several *statistical* (latent) factor estimation techniques.¹ While most practitioners are familiar with the basic regression methods used in the first two approaches, there is some apparent confusion, even in academic literature, regarding the different assumptions, features and requirements of various latent factor estimation methods.²

Recent work in the area of Bayesian factor analysis, Mayekawa (1985), Press and Shigemasu (1989,1997) and Rowe (2002) which followed the development of the *EM* algorithm of Dempster, Laird, and Rubin (1977) and Wu (1983) provide a solid common theoretical framework for examining the similarities and differences between these methods. We also bring attention to some apparently lesser-known earlier work in the area of latent factor modeling, that shed considerable light on these issues. We review three popular latent factor methods used in the financial literature, which are sometimes presented as distinct approaches to latent variable estimation. By formulating the various approaches in a Bayesian-*EM* framework, we are able to specify a common algorithm in which the similarities and differences between the methods become readily apparent.³

¹Connor (1995) reviews these three basic methods and examines the relative performance of each approach.

²While the basic methods used in the first two approaches generally involve ordinary and or weighted least squares regression to estimate the model parameters, the construction of the innovations in economic data series as well as the selection and transformation of the fundamental data variables often involve far more complex statistical methods.

³While the use of Bayesian priors apply to all variables in the linear factor models, we focus our attention solely on the form of the prior for the factor scores.

Section II introduces notation for the basic linear factor model and reviews some of the standard assumptions. In section III, we briefly review several of the more popular methods of estimating latent factor models that have appeared in the financial literature. In section IV we review the EM algorithm of Dempster, Laird, and Rubin (1977) and its application to factor analysis by Rubin and Thayer (1982). Section V demonstrates how the EM algorithm can easily be used to develop an algorithm for a variety of different assumptions regarding the “missing” factors. We develop a general *EM* factor analysis algorithm in which the specific form of key variables in the algorithm can be simply selected from a table entry corresponding to the specific assumptions implicit in each of the different methodologies. We extend the *EM* algorithm for factor analysis of Rubin and Thayer (1982) to the case of a (stochastic) vague prior for the factor scores.

II The Linear Factor Model

The general form of the linear common factor model is

$$y_{i,j} = \mathbf{z}_i \mathbf{b}_{i,j} + e_{i,j} \quad i = 1, \dots, n \quad j = 1, \dots, p \quad (1)$$

where

- $y_{i,j}$ is the i th return observation for security j ,
- \mathbf{z}_i is a $(1 \times q)$ vector of the i th observation of the factor scores,
- $\mathbf{b}_{i,j}$ is a $(q \times 1)$ vector of i th period factor loadings, for the j th security, and
- $e_{i,j}$ is the i th residual error for security j .

Each of the three approaches to building factor models of security returns described above make different assumptions as to which variables are observed and which must

be estimated as well as placing restrictions on the estimated values. The approach taken in building a factor model depends on the end objective of the model. Portfolio managers may be most concerned with identifying a limited number of macro-economic variables that give them an “edge” in forecasting. A derivatives trader may be more interested in models that capture changes in forecasted volatility, estimating a GARCH model for the residual and/or factor variance process. A hedge fund manager running a long-short market neutral equity portfolio may care little about the “names” of the factors or changes in volatility as long as he has captured all common factors and constructed his portfolio with the same amount of factor exposure on both the long and short positions.

The fundamental variable approach popularized by Barr Rosenberg, treats the $\mathbf{b}_{i,j}$ as being observed data and estimates the factor scores, \mathbf{z}_i as parameters. In its simplest form, the economic variable approach treats the factors, \mathbf{z}_i as being observed, the factor loadings or betas as parameters to be estimated, and due to the limited number of degrees of freedom available, usually imposes the restriction that $\mathbf{b}_{i,j} = \mathbf{b}_j \forall j$. Latent factor models must estimate both the factor scores and factor loadings, again typically with the same restriction that the estimates of the factor loadings be constant across all observations.⁴ Some methods treat the factor scores as parameters while others treat the factors as (missing) stochastic variables. In the simplest case, all three approaches typically assume that the $e_{i,j}$ are i.i.d. normal and that the appropriate number of factors have been identified such that $cov(e_j, e_k) = var(e_j)$ if $j = k$, and equals 0 otherwise.

We focus our attention on the simplest case of regression and latent factor models since the algorithm for the maximum likelihood estimate (MLE) of the

⁴The stationarity assumption on the factor loadings in 1) and 3) is not required. For instance, a linear trend in the factor loadings can be estimated by simply including a dummy variable in which the factor series are incremented linearly each period.

parameters to the former is the basis of *EM* algorithms for the later. With the restriction on $\mathbf{b}_{i,j} = \mathbf{b}_j$, we can write the model in more compact matrix notation as follows;⁵

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E} \tag{2}$$

where

- \mathbf{Y} is an $(n \times p)$ array of security returns,
- \mathbf{Z} is an $(n \times q)$ array of factors (scores),
- \mathbf{B} is a $(q \times p)$ array of factor loadings (betas), and
- \mathbf{E} is an $(n \times p)$ array of disturbances.

To simplify notation, we assume that the return data, \mathbf{Y} , have no missing observations, and have been de-meanned.⁶ If the observed data are demeaned, then the factor means, μ_Z , will equal zero. In the case where the factors, \mathbf{Z} , are observed, and given our assumptions on $e_{i,j}$, the MLE estimates of the model parameters, \mathbf{B} and residual variances, $\boldsymbol{\tau}_j^2$, can be estimated by OLS regression of \mathbf{Y} on \mathbf{Z} .

$$\begin{aligned} \mathbf{B} &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} \\ \boldsymbol{\tau}^2 &= n^{-1} \text{diag}(\mathbf{E}^T \mathbf{E}) = n^{-1} \text{diag}(\mathbf{Y}^T \mathbf{E}) \end{aligned}$$

where

$$\mathbf{E} = \mathbf{Y} - \mathbf{Z}\mathbf{B} \tag{3}$$

Equations (3) typically yields maximum likelihood estimates of \mathbf{B} and $\boldsymbol{\tau}^2$ regardless

⁵We attempt to keep our notation consistent with Rubin and Thayer (1982) with a few minor extensions.

⁶Rubin and Thayer (1982) point out that with no missing returns, the MLE estimate of the mean is the sample mean.

of the distributional form of the factor scores, \mathbf{Z} .⁷ We will see in section IV that equation (3) will form the basis of a general *EM* algorithm for factor analysis under a variety of assumptions concerning subsequently “missing” factor scores, \mathbf{Z} .

III Some Popular Latent Factor Models

We briefly review three popular approaches for estimating factor models that appear early on in the financial literature. While we recognize that many new techniques have subsequently been developed (Rowe 2002), these methods continue to be widely used in modeling security returns.⁸

Roll and Ross (1980) were the first to use factor analysis in an attempt to test the assumptions of the APT of Ross (1977). Roll and Ross used the algorithm of Jöreskog, which is often referred to as *maximum likelihood factor analysis*. This convention has been the source of some apparent confusion. While the model assumes the factor scores are multivariate *normal* stochastic variables, maximum likelihood estimation can be done for a variety of different distributional forms of \mathbf{Z} . We will distinguish between model assumptions and estimation algorithms, referring to this model as the traditional normal-linear model, and Jöreskog’s algorithm as one approach for obtaining maximum likelihood estimators for this model.

One caveat to note here is that Jöreskog’s algorithm includes another term in addition to likelihood equation in the objective function of his algorithm, $\log(\det(\mathbf{S}))$, where \mathbf{S} is defined as the $(p \times p)$ sample covariance matrix.⁹ As noted in Rubin and Thayer (1982), for a given data set, \mathbf{Y} , this term is a constant, unaffected by the parameter estimates. Nonetheless, the inclusion of the term in the objective function

⁷The main issue when assuming stochastic regressors concerns the independence of the now random regressors and disturbances.

⁸Robert Korajczyk maintains an extensive list of papers on estimating APT models and factor methodology at: www.kellogg.nwu.edu/faculty/korajczy/htm/aptlist.htm

⁹The term is included for subsequent calculation of a likelihood ratio test.

requires that the number of observations, n , exceed the number of variables, p . If $n < p$, then \mathbf{S} will be singular, $\det(\mathbf{S})$ will equal zero, and since $\log(0) = -\infty$, the objective function cannot be computed, causing an error message. Apparently as a result of computer programs generating this error message, many people have erroneously concluded that maximum likelihood estimates of the normal linear factor model are not possible when $n < p$. In fact, the model is perfectly well-defined for the case of $q < n < p$.¹⁰

Another feature to note is that Jöreskog's algorithm computes the Hessian matrix of all parameters of the model. While this leads to very rapid convergence when the number of parameters is small, it quickly becomes computationally burdensome for very large p , as the computational requirements are of the order p^2 . Roll and Ross (1980) note that the maximum number of stocks in any one group were limited to 30 due to computer resource limitations. While computer power has increased significantly in the last two decades, this can still be a problem with data sets requiring an extremely large number of variables.

Brown and Weinstein (1983) examine the "bi-linear" factor model in which the factor scores are treated as additional parameters of the model in the same sense as the factor loadings (betas). This method is also known as the least squares method of factor analysis (LSMFA) and was originally proposed by Young (1942) and Lawley (1943). The LSMFA is intuitively pleasing in its simplicity and produces directly observable estimated residuals just as in regression analysis.¹¹ The LSMFA model parameters can also be easily estimated via a simple conditional maximization (CM) algorithm. CM results in an iterative algorithm in which the steps alternate between

¹⁰This form of the likelihood ratio test, however, is not defined for $n < p$.

¹¹Note that in the standard normal-linear model, the differences between the observed and predicted values involve the sum of two random variables, \mathbf{Z} and \mathbf{E} . Thus estimates of \mathbf{E} by itself are not possible.

OLS regression of stock returns on the estimated factor scores from the previous iteration, and WLS regression of returns on the estimated factor loadings from the previous iteration, where the estimated factor loadings are weighted by the inverse of the estimated residual variance from the previous iteration.

$$\begin{aligned}\mathbf{Z} &= \mathbf{Y}\boldsymbol{\tau}^{-2}\mathbf{B}^T(\mathbf{B}\boldsymbol{\tau}^{-2}\mathbf{B}^T)^{-1} \\ \mathbf{B} &= (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y} \\ \boldsymbol{\tau}^2 &= n^{-1}\text{diag}(\mathbf{E}^T\mathbf{E}) = n^{-1}\text{diag}(\mathbf{Y}^T\mathbf{E})\end{aligned}$$

where

$$\mathbf{E} = \mathbf{Y} - \mathbf{Z}\mathbf{B} \tag{4}$$

Note that (4) is identical to the OLS regressions in (3) with the addition of the estimation of \mathbf{Z} . Brown and Weinstein note that under appropriate assumptions, the algorithm produces maximum likelihood parameter estimates.

Another widely used technique for constructing factor models is principal components analysis (PCA). There are three basic variations on PCA of security returns, 1) PCA of the covariance matrix of returns, 2) PCA of the correlation matrix of security returns, and 3) principal factor analysis (PFA). PCA of either the covariance or correlation matrix of security returns, results in a set of eigenvectors associated with the q -largest eigenvalues of the matrix. Connor and Korajczyk (1986) point out that PCA can easily be applied to very large data sets by noting that non-zero eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ are identical to those of $\mathbf{Y}\mathbf{Y}^T$. In the first case, the q -largest eigenvectors of $\mathbf{Y}^T\mathbf{Y}$ are effectively the “portfolio weights”, $(p \times q)$, which multiplied by the $(n \times p)$ returns matrix, $\mathbf{Y}(n \times p)$, result in the estimated factor

scores, $\mathbf{Z}(n \times q)$. In the second case, the q -largest eigenvectors of $\mathbf{Y}\mathbf{Y}^T(n \times q)$ are now direct estimates of the factor scores. Once the factor score estimates have been obtained, the factor loadings, \mathbf{B} , are then estimated by OLS regression of returns, \mathbf{Y} on \mathbf{Z} .

In PFA, the return series, \mathbf{Y} , are first standardized by dividing each security's returns by the estimated residual variance from the previous OLS regressions. This results in an updated estimate of the covariance matrix and resulting eigenvectors. The process is repeated until convergence is obtained. For security returns, the number of securities, p , is typically an order of magnitude greater than the number of available return periods, n . Thus working with the eigenvalues of $\mathbf{Y}\mathbf{Y}^T$ will be computationally more efficient.¹²

Whittle (1952) shows that the LSMFA models are equivalent to PCA/PFA in the sense that a solution to one approach will also be a solution to the other, up to a rotation of the factor space.¹³ The equivalence of the methods are as follows;

$$\begin{aligned} \text{PCA Covariance Matrix} &\iff \text{LSMFA with } \tau_k^2 = \tau_j^2 \quad \forall k, j \\ \text{PCA Correlation Matrix} &\iff \text{LSMFA with } \tau_k^2/\text{var}(\mathbf{y}_k) = \tau_j^2/\text{var}(\mathbf{y}_j) \quad \forall k, j \\ \text{Principal Factor Analysis} &\iff \text{LSMFA with unique } \tau_j^2 \end{aligned}$$

While PCA of the correlation and covariance matrix are one-step estimates both PFA and LSMFA are iterative procedures. Despite the computational differences in the two approaches, we can collapse the variations of PCA into the LSMFA and consider only the later models for simplicity.¹⁴ One drawback of these models is that by

¹²Connor and Korajczyk (198x) estimate principal factor models on data sets in excess of 10,000 variables.

¹³The estimates of the unique factor variances are unaffected by rotation and will thus be identical up to the tolerance of the computations.

¹⁴While we focus on the LSMFA for reasons of exposition, practical experience suggests that PFA

treating the factor scores as parameters of the model, nq degrees of freedom are used in fitting the factor scores alone versus the $(q^2 + q)/2$ used in estimating \mathbf{R} for the normal model.

The key feature of these models is that the factor scores are treated as unknown but non-stochastic parameters. This becomes a critical issue in obtaining parameter estimates. Anderson and Rubin (1957) show that for non-stochastic factor models with unrestricted unique variance estimates, the derivative of the likelihood function is negative in the neighborhood of zero for some τ_j^2 . This implies that it is not possible to obtain true maximum likelihood estimates for this class of models because the global maximum of the likelihood function will always involve a zero estimate of a unique variance for some variable.¹⁵ Any (unrestricted) estimate using all non-zero unique variance estimates can at best be a local stable point. Anderson and Rubin thus prove Whittle’s lament that the “LSMFA was too unstable to be useful”.¹⁶

IV The EM Algorithm

The *EM* algorithm of Dempster, Laird, and Rubin (1977) (DLR-77) and Wu (1983) often results in simple iterative algorithms for maximum likelihood parameter estimates for missing data problems. DLR-77 note that many statistical models can be recast in the context of the *EM* algorithm by viewing the problem as one of “conceptually missing” data, citing factor analysis as one example. The central feature of the *EM* algorithm when working with conceptually missing-data problems,

exhibits more stability and converges in fewer iterations.

¹⁵In the case where one of the unique variance estimates go to zero, the model effectively treats that particular variable by itself as one of the factors. While selecting a single variable as a factor will typically result in higher estimates of unique variance for all the other variables in the data, these increases in variance are “overwhelmed” by the one zero variance estimate since the determinant of the (diagonal) residual variance matrix is simply the product of the individual estimates.

¹⁶We would suggest that anyone using either LSMFA or principal factor analysis should include an explicit check on the lower bound of each unique variance estimate as some computer languages will continue on even after encountering divide by (nearly) zero conditions.

is that the algorithm utilizes the often simple-solution to the analogous complete-data problem with some *minor* adjustments.

Rubin and Thayer (1982) derive the *EM* algorithm for factor analysis for the traditional normal-factor model, noting that the MLE solution to the analogous complete-data is an OLS regression as in equation (3) above. It is important to note that the key requirement in the complete-data linear factor model that results in MLE-OLS regression algorithm, is that $e_{i,j} \sim N(0, \tau_j^2) \forall i, j$. With i.i.d. normal residuals, the conditional distribution of \mathbf{y}_j given the factors, \mathbf{Z} , and model parameters, $\boldsymbol{\Omega}_j \equiv (\mathbf{b}_j, \tau_j^2)$, is normal with mean $\mathbf{Z}\mathbf{b}_j$ and variance τ_j^2 .

$$(\mathbf{y}_j | \mathbf{Z}, \boldsymbol{\Omega}_j) \sim N(\mathbf{Z}\mathbf{b}_j, \tau_j^2) \quad (5)$$

As noted in Rubin and Thayer, the sufficient statistics for the complete-data MLE estimators are given by $C_{ZZ} \equiv n^{-1} \mathbf{Z}^T \mathbf{Z}$, $C_{ZY} \equiv n^{-1} \mathbf{Z}^T \mathbf{Y}$, and $C_{YY} \equiv n^{-1} \mathbf{Y}^T \mathbf{Y}$. Note that up to this point, *no assumption* is required regarding the distributional form of the factor scores, \mathbf{Z} . While the complete-data sufficient statistics are independent of the distribution of \mathbf{Z} , the *conditional expectation of the sufficient statistics*, given the observed data, \mathbf{Y} , and current estimate of the parameters, $\boldsymbol{\Omega}$ in the E-step of the *EM* algorithm requires a specific assumption as to the distribution of \mathbf{Z} when the factors are unobserved.¹⁷

Rubin and Thayer derive the conditional expectation of the sufficient statistics for the case of a multivariate normal prior on \mathbf{Z} . In this case $\boldsymbol{\Omega} = (\mathbf{B}, \boldsymbol{\tau}^2, \mathbf{R})$ where \mathbf{R}

¹⁷The parameter space, $\boldsymbol{\Omega}$ will of course also depend on the specific distributional form for \mathbf{Z} .

is the $q \times q$ factor covariance matrix.¹⁸ We expand slightly on their notation defining

$$\begin{aligned}
\mathbf{Z} &\equiv E[\mathbf{Z}|\mathbf{Y}, \Omega] = \mathbf{Y}\boldsymbol{\delta}, \\
E[C_{ZZ}|\mathbf{Y}, \Omega] &= \boldsymbol{\delta}^T C_{YY} \boldsymbol{\delta} + \boldsymbol{\Delta} \\
&= n^{-1} \mathbf{Z}^T \mathbf{Z} + \boldsymbol{\Delta}, \\
E[C_{ZY}|\mathbf{Y}, \Omega] &= \boldsymbol{\delta}^T C_{YY} = n^{-1} \mathbf{Z}^T \mathbf{Y}, \text{ and} \\
E[C_{YY}|\mathbf{Y}, \Omega] &= C_{YY} = n^{-1} \mathbf{Y}^T \mathbf{Y}.
\end{aligned} \tag{6}$$

Under the assumption that \mathbf{Z} is multivariate normal, the values of $\boldsymbol{\delta}$ and $\boldsymbol{\Delta}$ are given by Gauss's Theorem:

$$\begin{aligned}
\boldsymbol{\delta} &= (\tau^2 + \mathbf{B}^T \mathbf{R} \mathbf{B})^{-1} (\mathbf{B}^T \mathbf{R}), \\
\boldsymbol{\Delta} &= \mathbf{R} - (\mathbf{R} \mathbf{B}) (\tau^2 + \mathbf{B}^T \mathbf{R} \mathbf{B})^{-1} (\mathbf{B}^T \mathbf{R}).
\end{aligned} \tag{7}$$

Woodbury's identity simplifies the inversion of the $p \times p$ matrix in (7) $p \times p$ diagonal and $q \times q$ matrices as

$$\begin{aligned}
(\tau^2 + \mathbf{B}^T \mathbf{R} \mathbf{B})^{-1} &= \tau^{-2} - \tau^{-2} \mathbf{B}^T (\mathbf{R}^{-1} + \mathbf{B} \tau^{-2} \mathbf{B}^T)^{-1} \mathbf{B} \tau^{-2}, \\
&= \tau^{-2} - \tau^{-2} \mathbf{B}^T (\mathbf{R}^{-1} + \mathbf{F})^{-1} \mathbf{B} \tau^{-2},
\end{aligned} \tag{8}$$

where $\mathbf{F} \equiv \mathbf{B} \tau^{-2} \mathbf{B}^T$. The *EM* algorithm for the normal factor model thus involves the E-step where

$$\boldsymbol{\delta} = \tau^{-2} \mathbf{B}^T \mathbf{R} - \tau^{-2} \mathbf{B}^T (\mathbf{R}^{-1} + \mathbf{F})^{-1} \mathbf{F} \mathbf{R}, \quad \text{which can be further simplified as}$$

¹⁸Since \mathbf{Y} is assumed to have no missing data and is de-meant, we can assume that the factor scores have zero mean and ignore the means of the factors in Ω . If \mathbf{Y} has partially missing observations, however, the convergence rate of the *EM* algorithm is improved by explicitly including the factor means in the parameter set. See Liu, Rubin, and Wu, (1998).

$$\begin{aligned}
&= \tau^{-2} \mathbf{B}^T (\mathbf{R}^{-1} + \mathbf{F})^{-1}, \\
\Delta &= \mathbf{R} - \mathbf{R} \mathbf{B} \delta, \quad \text{and} \\
\mathbf{Z} &= \mathbf{Y} \delta.
\end{aligned} \tag{9}$$

The M-step is given as;

$$\begin{aligned}
\mathbf{B} &= [\mathbf{Z}^T \mathbf{Z} + n \Delta]^{-1} \mathbf{Z}^T \mathbf{Y}, \\
\mathbf{E} &= \mathbf{Y} - \mathbf{Z} \mathbf{B}, \\
\tau^2 &= n^{-1} \text{diag}(\mathbf{Y}^T \mathbf{E}) = \text{diag}(n^{-1} \mathbf{E}^T \mathbf{E} + \mathbf{B}^T \Delta \mathbf{B}), \quad \text{and} \\
\mathbf{R} &= n^{-1} \hat{\mathbf{Z}}^T \mathbf{Z} + \Delta.
\end{aligned} \tag{10}$$

Note that for the normal-factor model, \mathbf{E} represents “pseudo residuals”, in the sense that they are based on the estimated factor scores, \mathbf{Z} . As such, the usual regression estimate of the residual variance, $n^{-1} \text{diag}(\mathbf{E}^T \mathbf{E})$ must be adjusted by adding the term $\mathbf{B}^T \Delta \mathbf{B}$.

V Non-Normal Priors

While the normal factor model is perhaps the most popular method of latent factor modeling, some authors have expressed concern with regards to the assumption of a normal prior on the factor scores. Roll and Ross (1980) express concern with regards to this assumption stating “unknown biases and inconsistencies may be introduced”. At the opposite end of the spectrum for stochastic factor models is the case of an uninformed or vague prior on \mathbf{Z} .¹⁹

Under the assumption of a vague prior on the factor scores, the log likelihood

¹⁹See Mayekawa (1985), Press and Shigemasa (1989,1997), and Rowe (2002).

function can be expressed as;

$$LL = -\frac{n}{2} \log |\boldsymbol{\tau}^2| - \frac{1}{2} \sum_{i=1}^n (\mathbf{y}_i - \mathbf{z}_i \mathbf{B}) \boldsymbol{\tau}^{-2} (\mathbf{y}_i - \mathbf{z}_i \mathbf{B})^T \quad (11)$$

Note that $E(\mathbf{z}_i)$ now explicitly appears in the log likelihood as a parameter and thus require nq degrees of freedom in estimating \mathbf{Z} . While Rubin and Thayer (1982) developed the *EM* algorithm for the case of normal factors, *EM* can be applied to any prior distribution assumption. As noted above, the general form of the complete-data sufficient statistics depends only the assumptions regarding the errors, \mathbf{E} . Thus given our earlier assumptions regarding \mathbf{E} , and the assumption of a vague prior on \mathbf{Z} , the conditional expectation of the E-step sufficient statistics are given as;

$$\begin{aligned} E[C_{YY} | \mathbf{Y}, \boldsymbol{\Omega}] &= C_{YY} = n^{-1} \mathbf{Y}^T \mathbf{Y} \\ E[C_{YZ} | \mathbf{Y}, \boldsymbol{\Omega}] &= C_{YY} \boldsymbol{\tau}^{-2} \mathbf{B}^T (\mathbf{B} \boldsymbol{\tau}^{-2} \mathbf{B}^T)^{-1} \\ &= C_{YY} \boldsymbol{\delta} = n^{-1} \mathbf{Y}^T \mathbf{Z} \\ E[C_{ZZ} | \mathbf{Y}, \boldsymbol{\Omega}] &= \boldsymbol{\delta}^T C_{YY} \boldsymbol{\delta} + (\mathbf{B} \boldsymbol{\tau}^{-2} \mathbf{B}^T)^{-1} \\ &= n^{-1} \mathbf{Z}^T \mathbf{Z} + \boldsymbol{\Delta} \end{aligned} \quad (12)$$

Thus the *EM* algorithm for the vague prior involves the E-step given by

$$\begin{aligned} \mathbf{Z} &= \mathbf{Y} \boldsymbol{\delta} \\ \boldsymbol{\delta} &= \boldsymbol{\tau}^{-2} \mathbf{B}^T (\mathbf{B} \boldsymbol{\tau}^{-2} \mathbf{B}^T)^{-1} \\ &= \boldsymbol{\tau}^{-2} \mathbf{B}^T \mathbf{F}^{-1} \\ \boldsymbol{\Delta} &= (\mathbf{B} \boldsymbol{\tau}^{-2} \mathbf{B}^T)^{-1} \\ &= \mathbf{F}^{-1} \end{aligned} \quad (13)$$

and M-step given by

$$\begin{aligned}
\mathbf{B} &= (\mathbf{Z}^T \mathbf{Z} + n\mathbf{\Delta})^{-1} \mathbf{Z}^T \mathbf{Y} \\
\tau^2 &= n^{-1} \text{diag}(\mathbf{Y}^T \mathbf{E}) = \text{diag}(n^{-1} \mathbf{E}^T \mathbf{E} + \mathbf{B}^T \mathbf{\Delta} \mathbf{B}) \\
\mathbf{E} &= \mathbf{Y} - \mathbf{Z} \mathbf{B}
\end{aligned} \tag{14}$$

We can also develop the *EM* algorithm for the LSMFA and equivalently PFA by assuming a degenerate or point prior for the factor scores. As is typically the case, the expected value of \mathbf{Z} for the degenerate prior is the same as under the vague prior. The only difference in the *EM* algorithm under the degenerate prior is that $\mathbf{\Delta} = \mathbf{0}$. Thus the *EM* algorithm given in (12) and (13) with $\mathbf{\Delta} = \mathbf{0}$ is equivalent to the LSMFA estimators in (4).

The standard argument for Bayesian estimation is that if one has prior information concerning the value of a model variable from other sources than the current data set, then the information should be included in the final determination of the estimated parameter value. The problem then, is simply how much weight to give to the value based on prior information, versus the value indicated by the data at hand. As the amount of observed data increases, the influence of the data will overwhelm the influence of the prior, and the Bayesian estimates of the parameter value will approach that which would be obtained from the data alone.

We present *EM* algorithms for factor analysis for the cases of both a degenerate and vague prior regarding the (missing) factor scores. The resulting *EM* algorithms are very similar to that of Rubin and Thayer for the normal prior, since the complete-data algorithm for all cases is OLS regression. The three algorithms differ only in the value of the hyper parameters, $\mathbf{\Delta}$, a $q \times q$ matrix of the uncertainties in the factor scores given the returns and parameters, and $\mathbf{\delta}$, a $p \times q$ matrix that is used

to create the expected factor scores given the data and parameters. The following table summarizes the differences in The *EM* algorithm for the three different priors:

Table I

Hyperparameter	Prior Distribution on Factor Scores		
	Normal	Vague	Degenerate
δ	$\tau^{-2} \mathbf{B}^T (\mathbf{R}^{-1} + \mathbf{F})^{-1}$	$\tau^{-2} \mathbf{B}^T \mathbf{F}^{-1}$	
Δ	$(\mathbf{R}^{-1} + \mathbf{F})^{-1}$	\mathbf{F}^{-1}	0
Table I lists the values of the hyperparameters, δ and Δ , described in Rubin and Thayer (1982), for each of the three priors.			

Note, that as the number of variables, p , increases, the eigenvalues of \mathbf{F} becomes large while \mathbf{R}^{-1} remains approximately constant. Each additional security increases information about the (missing) factor scores and, hence, the eigenvalues of \mathbf{F} , which leads to $\lim_{p \rightarrow \infty} (\mathbf{R}^{-1} + \mathbf{F}) = \mathbf{F}$. Thus as the amount of information regarding the factor scores increases with p , the construction of the expected value of the factor scores under the normal, vague, and degenerate priors converge. Likewise, the estimates of the factor uncertainties under the normal and vague priors also converge. Furthermore, as \mathbf{F} increases without bound, \mathbf{F}^{-1} will approach $\mathbf{0}$ as in the case of the degenerate prior. Thus for data sets with a very large number of variables, p , as we have with security returns, the impact of assuming a normal versus vague prior distribution on the parameter estimates should be negligible.²⁰

The above results hold for the case where $p \rightarrow \infty$. It is of interest to practitioners to determine approximately how large of value for p is *large enough* such

²⁰Williams (1978)

that the differences in the three models are insignificant. We estimate both 5 and 10-factor models using five years of CRSP historical daily returns for all stocks with no missing returns over the period 1999-2003.²¹ This results in a total of 3599 return series with 1265 return observations per security. Each of the three models are run until the largest absolute change in the unique variance estimates is less than .0000000005.²² Since the *EM* algorithm is an iterative process, it must be seeded with initial parameter estimates. In all models the initial values of the unique variances is set to unity. Due to the strong initial convergence properties of the *EM* algorithm, the initial estimates of the factor loadings, \mathbf{B} , are simply set to random values.²³

Table 2-a lists the correlation coefficients between the (biased) maximum likelihood estimates of the unique variances of each model for the case of five factors. Note that the coefficients between the degenerate and the two stochastic priors are equal to six decimal places, while the coefficient between the vague and normal priors is equal to nine decimal places. Table 2-b lists the coefficient of multiple determination between each of the factor scores from the degenerate prior model and the vague and normal prior models. Tables 3-a and 3-b show the same calculations for a ten factor model.

²¹We restrict the analysis to no missing returns for simplicity. The *EM* algorithm for each of the three cases can be extended to account for partially missing returns. See Little and Rubin (1987).

²²This is approximately machine precision for the cumulative precision error.

²³Slightly fewer iterations are needed to obtain convergence if the initial parameter values are set to the parameter estimated from an OLS regression on the factor score estimates obtained from PCA of the returns.

5-Factor Model

Table 2-a

Correlation between unique variance estimates		
	Degenerate	Vague
Vague	0.99999976107075	
Normal	0.99999974318853	0.9999999966803

Table 2-b

Coefficient of Multiple Determination - R^2		
Factor	Vague	Normal
1	0.99809006690431	0.99849457278232
2	0.99961970087516	0.99995728116171
3	0.99865157451845	0.99999371907491
4	0.99984274114329	0.99986981939403
5	0.99999897978171	0.99755218288660

10-Factor Model

Table 3-a

Correlation between unique variance estimates		
	Degenerate	Vague
Vague	0.99999980150497	
Normal	0.99999978155629	0.99999999947440

Table 3-b

Coefficient of Multiple Determination - R^2		
Factor	Vague	Normal
1	0.99926270577202	0.99990167675355
2	0.99994183239856	0.99997544665440
3	0.99994341422415	0.99961339662613
4	0.99996918134376	0.99993522822175
5	0.99991684845893	0.99998395035008
6	0.99987331687017	0.99999313581024
7	0.99991676872900	0.99993933620604
8	0.99991795377910	0.99982162001874
9	0.99995708277285	0.99994792662562
10	0.99999948799603	0.99946202043190

Conclusions:

The *EM* algorithm for factor analysis clearly defines the differences between several popular methods for estimating statistical factor models. When cast in a Bayesian framework, the differences are due to differences in the distributional prior for the factor scores. As in all cases of Bayesian estimation, the impact of the specific

prior diminishes as the amount of data increases. For the case of security returns, the large number of different securities result in virtually no significant economic differences between the various models.

References

- [1] Anderson, T.W. and Herman Rubin. "Statistical inference in factor analysis.", *Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5, 1956.
- [2] Bartholomew, D. J. (1987), *Latent Variable Models and Factor Analysis*, Charles Griffin & Company LTD, London.
- [3] Bentler, P. M. and Jeffery S. Tanaka. (1983), "Problems with *EM* Algorithms for ML Factor Analysis." *Psychometrika*, Vol. 48, 247-251.
- [4] Brown, Stephen J. and Mark I. Weinstein, "A new approach to testing asset pricing models: The bilinear paradigm." *Journal of Finance*, vol. 38, June 1983.
- [5] Chen, Chan-Fu. (1981), "The *EM* Approach to the Multiple Indicators and Multiple Causes Model via the Estimation of the Latent Variable." *Journal of the American Statistical Association*, Vol. 76 No. 375, 704-708.
- [6] Connor, Gregory, "Three types of factor models: A comparison of their explanatory power.", *Financial Analysts Journal*, vol 51: pp42-46, May/June 1995.
- [7] Connor, Gregory and Robert A. Korajczyk, "Performance measurement with the arbitrage pricing theory: A new framework for analysis.", *Journal of Financial Economics*, 15:373-394, March 1986.
- [8] Dempster, Arthur P., N.M. Laird, and Donald B. Rubin. (1977), "Maximum Likelihood from Incomplete Data via the E-M Algorithm." *Journal of the Royal Statistical Society Series B*, Vol. 39, 1-38.
- [9] Johnson, Richard A. and Dean W. Wichern. (1982). (1988), *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- [10] Jöreskog, K. G. (1969), "A General Approach to Confirmatory Maximum Likelihood Factor Analysis." *Psychometrika*, Vol. 34, 183-202.
- [11] Jöreskog, K. G. (1977), "Factor Analysis by Least Squares and Maximum-Likelihood Methods." in *Statistical Methods for Digital Computers*, (edited by Kurt Enslein et al.), John Wiley & Sons, New York, pp125-153.
- [12] Lawley, D. N. and A. E. Maxwell. (1971), *Factor Analysis as a Statistical Method*, Second edition. London: Butterworth.
- [13] Lawley, D. N. (1941). Further investigations in factor estimation. Proc. Roy. Soc. Edinburgh Sec. A 61 176-185.

- [14] Lehmann, Bruce N. and David M. Modest. (1988), "The Empirical Foundations of the Arbitrage Pricing Theory." *Journal of Financial Economics*, Vol. 21, 213-254.
- [15] Little, Roderick J. A. and Donald B. Rubin. (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
- [16] Pastor, Lubos and Stambaugh, Robert F. "The Equity Premium and Structural Breaks." *Journal of Finance*, 2001, 56(4), pp. 1207-39.
- [17] Pastor, Lubos and Stambaugh, Robert F. "Costs of Equity Capital and Model Mispricing." *Journal of Finance*, 1999, 54(1), pp. 67-121.
- [18] Press, S.J. and K. Shigemasu, *Bayesian inference in factor analysis.*, chapter 15. Springer-Verlag 1989.
- [19] Press, S.J. and K. Shigemasu, "Bayesian inference in factor analysis-revised", Technical report 243, Department of Statistics, University of California Riverside, May 1997.
- [20] Roll, R. and S. A. Ross (1980), "An empirical investigation of the arbitrage pricing theory." *Journal of Finance*, 35(5):1073-1103, December 1980.
- [21] Ross, Stephen A. "The arbitrage theory of capital asset pricing", *Journal of Economic Theory*, vol. 13 pp. 341-360, December.
- [22] Rowe, D.B. and S. J. Press (1998), "Gibbs sampling and hill climbing in bayesian factor analysis." Technical report, Department of Statistics, University of California, Riverside, CA, May 1998.
- [23] Rowe, Daniel B. (2000), "A bayesian factor analysis model with generalized prior information", Technical Report 1099, Division of the Humanities and Social Sciences, California Institute of Technology, August
- [24] Rowe, Daniel B. (2002), *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*, CRC Press, Boca Raton, FL, USA
- [25] Rubin, Donald B. and Dorothy T. Thayer. (1982), "EM Algorithms for ML Factor Analysis." *Psychometrika*, Vol. 47, 69-76.
- [26] Rubin, Donald B. and Dorothy T. Thayer. (1983), "More on EM for ML Factor Analysis." *Psychometrika*, Vol. 48, 253-257.
- [27] Stroyny, Alvin L. (1991), "Heteroskedasticity and the estimation of systematic risk", PhD thesis, University of Wisconsin - Madison

- [28] Zhenyu Wang (2005) “A Shrinkage Approach to Model Uncertainty and Asset Allocation.” *Review of Financial Studies*, 18 (2), forthcoming.
- [29] Whittle, P. (1952) “On principal components and least squares methods of factor analysis” *Skandinavisk Aktuarietidskrift*, vol. 36, pp. 223-239.
- [30] Wu, C. F. J. (1983), ”On the convergence properties of the *EM* algorithm”, *Ann. Statist.* vol. 11, pp. 95-103.
- [31] Young, G. (1940) “Maximum likelihood estimation and factor analysis”, *Psychometrika*, vol. 6, pp. 49:53