

Estimating a Combined Linear Factor Model

\C:\PCTeX32\texinput\CombinedLinearModel4s.tex - Thursday-October-24-03

Alvin L. Stroyny
Chairman
EM Applications Limited

St. Martins House,
16 St. Martin's LeGrand,
London EC1A4EN,
+44 (020) 73978395
astroyny@emapplications.com

Copyright ©2003

All rights reserved, no part of this material may be reproduced in whole or part without the expressed written permission of the author.

Estimating a Combined Linear Model

Summary:

Most linear factor models used in portfolio risk management employ one of three basic estimation procedures, least squares regression on time-series data, (weighted) least squares regression on fundamental accounting variables, or factor analysis. These are often referred to as *economic*, *fundamental*, and *statistical* factor models. A variety of arguments have been offered as to why one approach or another is purportedly “better” than the others. We feel that each approach has merit and that there is no reason to use any one method to the exclusion of the others. We present an algorithm for estimating a *combined* linear model that incorporates the basic features of all three approaches in a single simultaneous estimation procedure. Under a set of appropriate assumptions, the resulting parameter values are maximum likelihood estimates. The simultaneous estimation procedure allows for some extensions of the linear model as well.

Introduction:

Linear factor models are widely used for modeling portfolio risk. Three popular approaches most frequently used in modeling security returns are:

- a) time-series regression with known factors, and estimated betas (assumed constant across observations),
- b) cross-sectional regressions using known fundamental/technical variables as proxies for betas which may vary from period to period, and estimated factor values, and
- c) factor analysis where factor values and betas are both missing and must be estimated (again beta is typically assumed to be constant across observations).

One of the first multi-factor models of security returns in the financial literature is that of King (1966), using multiple industry based return indices. Security “betas” were estimated using ordinary least squares (OLS) regression of each security return series on specified industry portfolios. BIRR is a consulting firm that constructs a variety of proprietary economic data series used to estimate a linear factor model for security returns.¹ While the methodology used to create each of the factor series can be quite technically advanced, the estimation of the security *betas* utilize the same basic linear (times series) regression methodology as in King (1966). This approach is often referred to as economic factor models. Professional investment managers may feel that choosing data series that they are comfortable with is an advantage of this model, since they may

¹BIRR is a commercial portfolio risk consulting firm operated by E. Burmeister, J. Ingersoll, R. Roll, and S. Ross. The company’s web site is www.birr.com.

feel that they have the ability to predict changes in the direction and/or volatility of the series.

Rosenberg (1974) proposed a multi-factor model that included market and industry components. This approach has been extended by BARRA to develop a “fundamental” factor model that is widely used by investment managers.² Again, the proprietary statistical methodology used to create the “fundamental” firm characteristic variables can be quite complex. Once these variables are created, the statistical estimation of the factors is done using weighted least squares regression. One of the main advantages often cited for the fundamental factor model approach is that by utilizing frequently updated accounting data, this approach can predict changes in a company’s risk profile before methods that rely solely on historical return data.

Roll and Ross (1980) used factor analysis to estimate a common factor structure for equity returns. Their use of Joreskog’s algorithm along with computational limitations of the time prevented them analyzing large numbers of securities in a single estimation. Factor analysis models of security returns gained wide-spread use in the mid 80’s as computing power became cheaper and new estimation techniques were developed that made it feasible to apply factor analysis to a data set containing a large number of securities. Connor and Korajczyk (1986) utilized a principal factor analysis algorithm capable of working with very large data sets. EM Applications is one of several consulting firms that offer a variety of factor models that combine both economic data series as well as statistical factors in various global markets.³ One of the main advantages that is often cited for factor analysis is that it is completely data-driven and thus not subject to human-specified available data biases.

Connor (1995) examined the explanatory power of the three forms of factor models. He finds that the fundamental approach performs slightly better than the other two according to his criteria. While the various arguments both for and against each of the three basic forms of factor models makes it difficult to choose a single “best” approach to building a factor model, most everyone would agree that each of the three methods have useful features not found in the other two. This would seem to raise the question, “Why not use the best parts of all three methods?”. The objective of this chapter is to present an algorithm for obtaining maximum likelihood estimates of a linear factor model that incorporates *all three* of the basic forms.

A Combined Linear Factor Model:

²BARRA is a commercial portfolio risk consulting firm founded by Barr Rosenberg. The company’s web site is www.barra.com.

³EM Applications is a UK-based consulting firm. The company’s web site is www.emapplications.com.

We use the following convention for notation; row and column vectors are denoted as lower case boldface letters, matrices are denoted as upper case boldface letters, while scalars and matrix/vector elements are denoted as non-bold lower case letters. The i subscript is used to refer to the observation (time) domain and the j subscript to refer to the variable (stock) domain. There will be some instances of three-dimensional matrices, in which case we will use an upper-case bold letter with a subscript to designate a two-dimensional “slice” of the full three-dimensional matrix.

The usual form of the economic and statistical factor models can be represented in matrix form as

$$Y = ZB + E \quad (1)$$

where Y is the $(n \times p)$ matrix of n return observations on p securities, Z is the $(n \times q)$ matrix of n observations on q factor variables, B is the $(q \times p)$ matrix of q factor sensitivities (betas) on p securities, and E is the $(n \times p)$ matrix of n residuals on p securities. One of the features of the usual form above models is that the factor sensitivities matrix, B , is typically assumed constant across the i observations.⁴ This assumption allows for compact notation as above.

In the case of fundamental factor models, the factor sensitivity matrix takes on a third dimension associated with each i -observation. Thus the full three-dimensional data set is expressed as a series of two-dimensional slices, B_i , $i = 1, \dots, n$. It is this large volume of fundamental data that makes the information potential of this type of model appealing. Note that almost any observable feature of a security prior to period i can be used as an input “fundamental” data characteristic matrix, B_i . In particular, past characteristics of security returns, $g(y_{i-1}, y_{i-2}, \dots)$, can be input at time i as deterministic variables. Thus an infinite class of “technical” variables can be constructed from past price series alone. This form of model has been popularized by BARRA using both (industry-normalized) fundamental accounting data as well as technical price/volume type variables. Due to the three-dimensional form of the factor sensitivities, the model is specified in the following notation:

$$y_i = z_i B_i + e_i, \quad (i = 1, 2, \dots, n), \quad (2)$$

where y_i refers to the $(1 \times p)$ row-vector of i -th return observation on p securities, z_i is the $(1 \times q)$ row-vector of the i -th observation on q factor variables, B_i is now the $(q \times p)$ matrix of

⁴While B is typically assumed to be constant, this is not a strict requirement of these models and a variety of time-dependent forms can be utilized.

the now the i -th realization of the now observation dependent factor sensitivity matrix, and e_i is the $(1 \times p)$ row-vector of the i -th observation residuals.

Denoting the three forms described above by the subscripts, a, b, c , we can express a combined linear factor model that incorporates all three basic forms:

$$y_i = z_{ai}B_a + z_{bi}B_{bi} + z_{ci}B_c + e_i \quad (i = 1, 2, \dots, n), \quad (3)$$

where y_i is the $(1 \times p)$ row-vector of i -th period returns on p -securities, z_{ai} is the $(1 \times q_a)$ row-vector of observed (regression-type) factors, B_a is the $(q_a \times p)$ matrix of the estimated regression slope coefficients (assumed constant across all observations), z_{bi} is the $(1 \times q_b)$ row-vector of estimated factors for period i , B_{bi} is the $(q_b \times p)$ matrix of i -th period firm characteristics (e.g. BARRA-type fundamental/technical variables), z_{ci} is the $(1 \times q_c)$ row-vector of estimated statistical factors for period i , B_c is the $(q_c \times p)$ matrix of estimated factor loadings, (again assumed constant across i) and e_i is the $(1 \times p)$ row-vector of i -th period residuals, where $i = 1, 2, \dots, n$. We do not address any of the asset pricing issues associated with estimates of the mean returns on securities.⁵ As such, the standard estimates of the intercepts in both the n and p dimensions can be incorporated in the model by including a unit column vector in Z_a and a unit row vector in B_{bi} or by constraining one row of B_c to equal a unit row vector.

An Extended Model:

An extension to the above model is to allow for a firm-specific parameter, A_j , $(q_b \times q_b)$, assumed constant across observations (time), controlling the intensity of the response to Z_{bi} . The model now becomes:

$$y_{ij} = z_{ai}b_{aj} + z_{bi}A_jb_{bij} + z_{ci}b_{cj} + e_{ij} \quad (i = 1, 2, \dots, n), \quad (j = 1, 2, \dots, p). \quad (4)$$

Equation (3) can be viewed as a special case of (4) where $A_j \equiv I$, $(j = 1, 2, \dots, p)$. This second form of the model allows for a *separable* response function combining a time-specific, firm-constant market variable, z_{bi} , and a time-constant, firm-specific intensity variable, A_j . Also note that A_j is the j th slice of another three-dimensional matrix. As a result, the notation must now be specified at the individual y_{ij} element level.

There are several advantages to this extended form of the linear model. Since the b_{bij} are constructed variables, there is the potential for scaling errors, (e.g. D/E ratios for banks not

⁵Additional Arbitrage Pricing Theory (Ross 1976)-type pricing restrictions could of course be imposed on the estimation procedure.

being comparable to those for technology stocks. By estimating A_j , the model would statistically adjust the response where industry or other a priori scaling schemes do not fully account for these differences.⁶ A variety of restrictions can be imposed on the form of A_j such as requiring a single value for all securities within a specified group (e.g. industry). There is now an indeterminacy in the model due to the multiplicative form of the Z_b and A . To avoid this problem, we need to add an additional uniqueness constraint, such as requiring that the average parameter value across all firms of the diagonal elements of A_j be equal to one:

$$p^{-1} \sum_{j=1}^p a_{jkk} = 1, \quad (k = 1, \dots, qb). \quad (5)$$

Model Estimation:

There are several methods that can be used to obtain parameter estimates for the models described above. Even if we restrict ourselves to obtaining maximum likelihood estimates, there still are several forms of algorithms that can be employed. While methods that calculate the Hessian matrix typically converge in the fewest iterations, the computational burden per iteration grows with the square of the number of parameters in the model making the method impractical in many applications where more than 10,000 financial securities may be involved. This is one of the problems Roll and Ross (1980) encountered when they tried to estimate a statistical factor model using Joreskog's algorithm.⁷

Dempster, Laird, and Rubin (DLR 1977) introduced the concept of the Expectation-Maximization (*EM*) algorithm for finding maximum likelihood parameter estimates in problems involving missing data. They point out one of the more intriguing aspects of the theory is the application of *EM* to problems involving *conceptually* missing data, mentioning factor analysis as one example. In these cases, the corresponding *EM* algorithms lead to simple iterative procedures since the corresponding *complete*-data mle problems have one-step solutions via OLS regression. Rubin and Thayer (1982) develop the *EM* algorithm for the case of traditional factor analysis, under the assumption that the missing factors are jointly-normal with the disturbances.

While their algorithm produced the same mle parameter estimates (to a rotation), it proved to be considerably slower than Joreskog's method for the particular problem. Rubin and Thayer also

⁶Fundamental data is typically "normalized" to the industry average in attempt to adjust for these differences. The advantage of this form of model is that it allows for within-industry scaling as well.

⁷While current CPU power and greater memory size allow for larger problems, Joreskog's algorithm also contains an additional term, not part of the likelihood function, that prevent's estimation when the number of variables, p , exceeds the number of observations, n . This result has lead many to erroneously conclude that maximum likelihood factor analysis is not possible in this case.

inadvertently impose an artificial constraint in their algorithm that prevented it from obtaining the steepest ascent path.⁸ Stroyny (1991) showed that when this constraint is removed, the convergence rate for the EM factor analysis algorithm actually improves as the number of variables, p , increase. Simulation evidence showed that at between 200-300 variables, the EM algorithm converged quicker than Joreskog’s method in terms of total CPU time even though it required more iterations.

In applying *EM* to missing data problems, the *E*-step involves taking the expectation of the *sufficient statistics* of the corresponding complete-data problem, (*not* simply the expected value of the missing data, though for some problems these are one in the same), conditioned on the observed data and current estimates of the parameters. Evaluating the expectation operator in the *E*-step requires a specific prior for the distributional form of the factors. The general theory of *EM*, however, can be applied to any distributional prior.

Before we can proceed with estimating the model, several assumptions must be made regarding the forms of statistical processes involved. We will assume that the disturbances, e_{ij} are identically distributed normal random variables with variance $var(e_{ij}) = \sigma_j^2, i = 1, \dots, n$, and are independent both between different securities, $cov(e_{ij}, e_{ik}) = 0, j \neq k$, and through time, $cov(e_{ij}, e_{hj}) = 0, h \neq i$. As is typically done, we will treat the unknown factor sensitivities, B_a and B_c as well as $A_j, j = 1, \dots, p$, as parameters to be estimated. The observed factors, Z_a and observed fundamental variables, $B_i, (i = 1, \dots, n)$, are treated as given non-stochastic variables.

The main difference among the various statistical methods for estimating factor models with latent factors has to do with the assumptions regarding the distributional form of the factor variables, Z_b and Z_c in the above model. The latent factors can be treated as either stochastic or non-stochastic variables. Both approaches have advantages and disadvantages. If the unknown factors are treated as stochastic variables, a distributional form must be specified. Again, each specific form of distribution will have advantages and drawbacks. One popular form of stochastic factor model is to treat the factors as normally distributed random variables as in the traditional factor analysis method of Joreskog and as assumed in the EM algorithm for factor analysis of Rubin and Thayer (1982). Stroyny and Rowe, 2002 (SR2002), utilized the EM algorithm to examine differences in factor analysis algorithms under a variety of prior distributional forms for the “missing” factor scores. They extended the normal prior EM algorithm of Rubin and Thayer (1982) to the case of a vague prior on the (stochastic) factor scores. SR2002 also showed that the

⁸Liu, Rubin, and Wu (1998) develop the PX-EM algorithm which addresses the type of problem encountered by Rubin and Thayer.

EM algorithm can be applied to the case of non-stochastic factors by assuming a degenerate or point prior, in which case, the EM algorithm is equivalent to the least squares method of factor analysis of Lawley (1942).

The main advantage of a non-stochastic factor model, (or equivalently a degenerate prior), is that the estimated residuals, E , are directly observable. This is not the case with stochastic factors as the total error term now includes the unknown variability in the factor score, (multiplied by the factor loading), as well as the disturbance term. One of the main limitations of non-stochastic factor models is that they are often unstable, with tendencies to estimate zero residual variances.⁹ Another limitation is a degrees of freedom is required for each observation on each factor score, or a total of $(n \times (q_b + q_c))$.

The main advantage of assuming a normal prior for the factor scores is a reduction in the number of parameters to be estimated. The reduced number of parameters contributes to the stability of the estimation procedure, reducing the tendency of zero residual variance estimates (e.g. Heywood cases). SR2002 show that as the number of variables, p , increase, the parameter estimates under the normal, vague, and degenerate priors converge. They also find that stability issues associated with initial parameter estimates and the tendency for estimating zero residual variances for the non-stochastic factor models tend to diminish as well as p increase.¹⁰

We develop an estimation algorithm under the assumption of non-stochastic factors that allows us to use a relatively simple conditional maximization (CM) algorithm.¹¹ The CM algorithm alternates between calculating values of the first set of parameters that maximize the likelihood given the second set of parameters, and maximizing the likelihood over the values of the second set of parameters given the first. The process is repeated until convergence is obtained.

The key to the CM algorithm for the linear factor models outlined above is that the maximization problem for each group of parameters in the model, is in the form of a *linear regression*. As such, we can directly write the (conditional) maximizing parameter estimates using the usual form of the sufficient statistics for either OLS or WLS regression. Under the assumption of non-stochastic factors, the resulting algorithm is also consistent with the general theory of EM under the assumption of a degenerate prior for the factors. In effect, the E-step of the EM algorithm

⁹See Anderson and Rubin (1956).

¹⁰The least squares method of factor analysis (LSMFA) was introduced by Lawley (1943). While the approach was intuitively simple, the method proved to be unstable resulting in estimates of zero residual variances. While less frequent, this also occurs in stochastic factor models as well.

¹¹The conditional maximization algorithm is also referred to as the alternating variable method.

becomes redundant since the expected values of the sufficient statistics given the observed data and current values of the parameters are equivalent to the sufficient statistics evaluated at the expected value of the parameters.

Conditional Maximization:

For the first form of the combined linear model, the parameters can be separated into two groups, z_{bi}, z_{ci} and B_a, B_c . The CM algorithm alternately maximizes the conditional log likelihood function for one set of parameters given the current values of the other group. The CM equation for the estimates of the first group of parameters, z_{bi}, z_{ci} , is given by the WLS regression as:

$$z_{bci} = (y_i - z_{ai} B_a) \Sigma^{-1} B_{bci}^T (B_{bci} \Sigma^{-1} B_{bci}^T)^{-1} \quad (i = 1, 2, \dots, n), \quad (6)$$

where;

$$\begin{aligned} B_{bci} &\equiv \begin{bmatrix} B_{bi} \\ B_c \end{bmatrix} && ((q_b + q_c) \times p), \\ z_{bci} &\equiv [z_{bi} \ z_{ci}] && (1 \times (q_b + q_c)), \\ e_i &= y_i - z_{ai} B_a - z_{bi} B_i - z_{ci} B_c && (1 \times p), \text{ and} \\ \Sigma &= n^{-1} \text{diag}(E^T E) && (p \times p). \end{aligned}$$

The additional parameters, B_a and B_c , are simply treated as constants based on their latest estimated values. In similar fashion, the second group parameter estimates, B_a, B_c , are given by the OLS (matrix) regression equations as:

$$B_{ac} = (Z_{ac}^T Z_{ac})^{-1} Z_{ac}^T \left(Y - \begin{bmatrix} z_{b1} B_{b1} \\ z_{b2} B_{b2} \\ \dots \\ z_{bn} B_{bn} \end{bmatrix} \right), \quad (7)$$

where

$$\begin{aligned} B_{ac} &\equiv \begin{bmatrix} B_a \\ B_c \end{bmatrix} && ((q_a + q_c) \times p), \text{ and} \\ Z_{ac} &\equiv [Z_a \ Z_c] && (n \times (q_a + q_c)). \end{aligned}$$

The parameter estimates from equation (6), Z_b and Z_c , are now treated as constants and evaluated at their latest estimated values. The CM algorithm thus iterates between equations (6) and (7), until convergence is obtained.

The CM algorithm can be started with either equation (6) or (7) and appropriate starting values for the group of parameters in the second equation. A variety of approaches can be taken regarding the initial parameter values. If we begin the algorithm say, with equation (6), estimating z_{bi}, z_{ci} , then we need to have initial estimates of B_a, B_c , and Σ . Initial estimates for B_a can be easily obtained by OLS regression of Y on Z_a . Initial estimates for the residual variance parameter, $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, are used on a relative basis to weight the individual observations and thus the scale of the initial estimates is irrelevant. Several alternative initial values for Σ are I , (e.g. equal weighting), $n^{-1} \text{diag}(Y^T Y)$, (e.g. proportional to total variance), or $n^{-1} \text{diag}(E^T E)$ where $E = Y - Z_a B_a$ from the OLS regression for the initial estimates of B_a .

We have found that in many instances, a pseudo random number generator can be used to generate the initial estimates of B_c . The initial convergence rate of the CM algorithm is extremely rapid in this case such that only a few additional iterations are saved by using improved estimates of B_c . If convergence problems do occur, then the factor loading estimates corresponding to the first q_c eigenvalues from either principal components or factor analysis of either $\text{cov}(Y)$, or $\text{cov}(E)$, the residuals from the OLS regression of B_a , could be utilized as improved estimates. Under the assumption of independent, identically distributed normal errors, e_{ij} , the CM algorithm as defined in equations (6) and (7), results in least-squares estimates of the model parameters which under the assumption of non-stochastic factors, will also be the maximum likelihood estimates.

Heterogeneous Errors:

One of the major problems associated with security return data is the presence of outliers. These are often associated with economic events specific to an individual company, such as earnings announcements, corporate actions, verdicts in lawsuits, etc. As such, the assumption of identically distributed errors is typically violated. We can relax the assumption in the above model to allow for heterogeneous error variances across observations, (time periods), analogous to the standard weighted least squares (WLS) regression model by assuming that the $(n \times 1)$ variance-scaling vector, $\theta_j = (\theta_{1j}, \dots, \theta_{nj})$, is known *a priori*, and θ_{ij} is defined such that:

$$\sum_{i=1}^n \theta_{ij} = n, \quad (j = 1, \dots, p). \quad (8)$$

The residual variance for the i -th observation of the j -th security is now given as:

$$\sigma_{ij}^2 = v_j \theta_{ij}, \quad (9)$$

where v_j is the variance scale constant for security j defined as:

$$v_j = n^{-1}(e_j^T \Theta_j e_j) / (\ell_n^T \Theta \ell_n), \quad (10)$$

ℓ_n is an $(n \times 1)$ unit column vector, and the $(n \times n)$ weighting matrix for security j , Θ_j , is given as:

$$\Theta_j = \text{diag}(\theta_{1j}^{-1}, \theta_{2j}^{-1}, \dots, \theta_{nj}^{-1}) \quad (n \times n). \quad (11)$$

As a result, the residual covariance matrix, Σ_i , will now also be time dependent and is defined as:

$$\Sigma_i = \text{diag}(\sigma_{i1}^2, \sigma_{i2}^2, \dots, \sigma_{ip}^2) \quad (p \times p) \quad (i = 1, \dots, n). \quad (12)$$

The CM equation for z_{bi} and z_{ci} now accounts for the heterogeneity in the disturbances (through time) and is given as:

$$z_{bci} = (y_i - z_{ai} B_a) \Sigma_i^{-1} B_{bci}^T (B_{bci} \Sigma_i^{-1} B_{bci}^T)^{-1} \quad (i = 1, 2, \dots, n). \quad (13)$$

Similarly, the estimates of B_a and B_c must account for the specific heterogeneous pattern of each different security, j . The OLS regression estimates of B_a and B_c in equation (7) are accordingly modified to a corresponding WLS form as:

$$b_{acj} = (Z_{ac}^T \Theta_j Z_{ac})^{-1} Z_{ac}^T \Theta_j \left(y_j - \begin{bmatrix} z_{b1} b_{b1j} \\ z_{b2} b_{b2j} \\ \dots \\ z_{bn} b_{bnj} \end{bmatrix} \right) \quad (j = 1, \dots, p), \quad (14)$$

where b_{acj} is the j -th column of B_{ac} as defined in equation 7. The CM algorithm now iterates through equations (9) to (14) until convergence is obtained. The model can be extended to allow for simultaneous estimates of the weighting matrix under a variety of assumptions regarding the specific form of the heterogeneity. We limit our analysis to the case where the weights are given exogenously.

Estimating the Extended Model:

If we now relax the restriction that $A_j \equiv I$, $\forall j$, the model estimation becomes slightly more complex. The model parameters will now include a third group for the CM of A_j . The estimation equations for z_{bci} and b_{acj} are a straight-forward extension of the previous section equations since A_j is treated as a constant, although the resulting matrix notation becomes slightly more

involved. The CM estimates for the factor scores, z_{bci} are again of the form of a WLS regression as:

$$z_{bci} = (y_i - z_{ai}B_a) \Sigma_i^{-1} G_{bci}^T (G_{bci} \Sigma_i^{-1} G_{bci}^T)^{-1} \quad (i = 1, 2, \dots, n), \quad (15)$$

where

$$G_{bci} = \begin{bmatrix} G_{bi} \\ B_c \end{bmatrix} \quad ((q_b + q_c) \times p), \quad \text{and}$$

$$G_{bi} \equiv [A_1 b_{bi1} \quad A_2 b_{bi2} \quad \dots \quad A_p b_{bip}] \quad (q_b \times p).$$

Likewise, the estimate for b_{acj} is essentially the same as it was for the previous model in equation (11) except that A_j now appears as an additional constant:

$$b_{acj} = (Z_{ac}^T \Theta_j Z_{ac})^{-1} Z_{ac}^T \Theta_j \left(y_j - \begin{bmatrix} z_{b1} A_j b_{b1j} \\ z_{b2} A_j b_{b2j} \\ \dots \\ z_{bn} A_j b_{bnj} \end{bmatrix} \right) \quad (j = 1, \dots, p). \quad (16)$$

As mentioned earlier, an intercept term for the fundamental factor model can be incorporated by having a unit row vector in either B_{bi} or B_c . When using the extended form of the model with A_j , it will be notationally more convenient to constrain a row of B_c to equal a unit row vector due to the interaction of A_j with B_{bi} .¹² In this case the estimates of b_{acj} are given by:

$$b_{acj}^* = (Z_{ac}^{*T} \Theta_j Z_{ac}^*)^{-1} Z_{ac}^{*T} \Theta_j \left(y_j - \begin{bmatrix} z_{b1} A_j b_{b1j} \\ z_{b2} A_j b_{b2j} \\ \dots \\ z_{bn} A_j b_{bnj} \end{bmatrix} - z_{c1} \right) \quad (j = 1, \dots, p),; \quad (17)$$

where z_{c1} is the $(n \times 1)$ column-vector corresponding to the first row of b_{cj} which is constrained to equal a unit row-vector for all securities, b_{acj}^* is the $((q_a + q_c - 1) \times 1)$ vector of estimated factor sensitivities corresponding to Z_a and the 2nd through q_b columns of Z_c , (e.g. the columns corresponding to the unconstrained rows of b_{cj}), and Z_{ac}^* is the $(n \times (q_a + q_c - 1))$ matrix formed from Z_a and Z_c^* , where Z_c^* represents the ‘‘unconstrained’’ factors and is constructed from the 2nd through q_c columns of Z_c , corresponding to the unconstrained factor loadings, B_c^* .

¹²If the first row of B_{bi} in the extended model is set equal to a unit row vector for estimating an intercept, then an additional constraint should be placed on the first row and column of A_j such that $a_{1,1} = 1$ and $a_{1,k} = a_{k,1} = 0$ for $k = 2, \dots, q_b$. Without this restriction, the form of term $z_{bi} a_{j,1,1}$ is equivalent to a regular factor analysis term in $z_{ci} b_{cj}$. While the model can estimate A_j in this case, the algorithm will tend to converge more slowly as the two forms of the ‘‘statistical’’ factors are estimated in opposing steps.

The estimation of A_j requires some additional algebraic manipulation to simplify the form of the model. We assume that B_{bi} does *not* contain a unit row vector for an intercept term.¹³ Let $vec()$ denote the operation creating a $(mk \times 1)$ column vector by stacking the columns of the $(m \times k)$ argument matrix, and \otimes denote the Kronecker product operator. The elements of the $(q_b \times q_b)$ matrix A_j can now be expressed as a $(q_b^2 \times 1)$ column vector:

$$vec(A_j) \equiv [a_{11j} \ a_{21j} \ \dots \ a_{q_b q_b j}]^T \quad (q_b^2 \times 1). \quad (18)$$

Define h_{ij} as the $(1 \times q_b^2)$ row vector formed by the Kronecker product of b_{bij}^T and z_{bi} and H_j as the $(n \times q_b^2)$ matrix formed by stacking the n -rows of h_{ij} , $i = 1, \dots, n$ as:

$$h_{ij} \equiv b_{bij}^T \otimes z_{bi} \quad (1 \times q_b^2), \quad \text{and}$$

$$H_j = \begin{bmatrix} h_{1,j} \\ h_{2,j} \\ \dots \\ h_{n,j} \end{bmatrix} \quad (n \times q_b^2). \quad (19)$$

With the above definitions, we can now express the term:

$$z_{bi} A_j b_{bij} = h_{ij} vec(A_j) \quad (1 \times 1), \quad (20)$$

and model (2) in the form:

$$y_j = Z_a b_{bj} + H_j vec(A_j) + Z_c b_{cj} + e_j. \quad (21)$$

With this definition, the second part of the M-step conditional maximization process can be expressed as a (WLS) linear regression to estimate A_j :

$$vec(A_j) = (y_j - Z_a b_{aj} - Z_c b_{cj})^T \Theta_j H_j (H_j^T \Theta_j H_j)^{-1}. \quad (22)$$

Discussion:

The construction of the economic factors, Z_a , may be done such that they are either directly comparable to certain economic data series, and thus typically somewhat collinear, or constructed to be standardized orthogonal variates, arranged in some meaningful ordering of importance. For example, a simple three factor model comprised of a market index and two industry/sector indexes would begin with a market proxy such as the S&P 500, and then use an “energy” index that

¹³If an intercept is estimated, we assume it is in the form of a constraint on B_c containing a unit row vector as discussed above.

has been designed to be orthogonal to the market index as factor 2, and finally use a “financial” index that is orthogonal to both the market and the “energy” series. The predicted values and estimated residual variances would of course be unchanged if the order of index construction was reversed. We assume that any issues regarding the make up of the economic factors are addressed prior to model estimation and thus treat $cov(Z_a)$ as given. Similarly, the form of $cov(Z_b)$ will be determined by the methods used to construct the “fundamental” data, $B_{bi}, i = 1, \dots, n$, and thus we also treat $cov(Z_b)$ as given.¹⁴ As a result, the covariance structure between the economic and fundamental factors, $cov(Z_{ab})$, is treated as given.

The “statistical” factor analysis portion model, $Z_c B_c$, ($Z_c^* B_c^*$ if the first row of B_c is constrained to a unit vector), results in the usual rotational indeterminacy problems. These are compounded due to the presence of the Z_a and Z_b factors as well. A “pure” rotation of the factors does not impact the value of the likelihood function but simply moves the parameter values to a point of equal “altitude” on the ridge of the likelihood surface.¹⁵ The choice of any rotational transformations is of course subject to individual interpretation. The statistical factors, Z_c , can be rotated to an orthogonal form (i.e. linearly independent of one another). In addition, the unconstrained factors, Z_c^* , can be scaled to have unit variance as well such that $cov(Z_c^*) = I$.¹⁶ Additional constraints are required to produce a unique rotation of the factor scores.¹⁷

Since the economic and statistical factors are multiplied by a static (i.e. through time) factor loading matrices, B_a, B_c , the statistical factors can also be rotated to be orthogonal to Z_a without impacting the value of the likelihood function. Note that due to the time-variation in the fundamental variable matrix, b_{bi} , it is not possible to simply orthogonalize the statistical factors, Z_c to the fundamental factors, Z_b without impacting the value of the likelihood function. This is due to the fact that the factor loading matrix for each variable, $B_j, (n \times q_b)$, has a different time-variation “pattern” for each variable.¹⁸ While it is thus possible to define a unique rotation for the entire factor space, there are nonetheless an infinite number of these and thus the choice of a specific rotation and its interpretation is left to the user.

¹⁴It is our understanding that BARRA also performs a post-estimation process to address collinearity issues and reduce the dimensionality of the estimated fundamental factors, Z_b .

¹⁵The estimates of the residual variance, s are also invariant to “pure” rotational transformations.

¹⁶The unit variance of Z_c^* is of course achieved by scaling the corresponding B_c^* . Since the factor loadings for the constrained factors are defined as a unit vector, the variance of the constrained factors cannot be rescaled.

¹⁷A variety of methods can be used to obtain a unique rotation such as ordering by decreasing eigenvalues of the (weighted) predicted factor covariance matrix.

¹⁸One way to view this problem is to note that any rotation of the fundamental and statistical factors effectively transfers part of the time variation of the fundamental factors to the statistical factors and vice-versa. The variation from the statistical factors are now compounded by the variation in B_j and the previous interaction effect between Z_b and B_j is now lost as the transferred variation is now multiplied by a constant B_{cj} .

Some Simulation Evidence:

As mentioned earlier, non-stochastic factor models have a tendency to produce estimates of zero residual variance when the number of variables, p , is small. The above models have a non-stochastic factor component and thus will also have this characteristic. A simulation data set is constructed with $n = 100, p = 50, q_a = 2, q_b = 2, q_c = 3$. The first column of Z_a is set to a unit column vector and the first row of B_c to a unit row vector. We assume a homoskedastic residual variance process (e.g. $\theta_{ij} = 1 \forall i, j$).

In the first set of simulation trials, the residual variance is scaled such that the average multiple correlation coefficient across all variables is .85, a high value relative to what might be expected using historical security returns. The initial parameter estimates of B_a are based on an OLS regression of Y on Z_a . To test the robustness of the algorithm to the initial parameter estimates, the same simulated data set is used to estimate the model for a variety of random initial parameter estimates for the residual standard deviation, σ_j , and the unconstrained statistical factor loadings, B_c^* . The initial values of σ_j are drawn from a uniform (1,6) distribution and B_c^* from a $N(0, 1)$ distribution. For small data sets, $p = 50$, the algorithm consistently converged to a zero estimated value of σ_j for at least one security. All else equal, increasing the number of variables to $p = 200$, resulted in consistent convergence in 20 iterations on average to the same value of the log likelihood for all 200 trials. A unique rotation is applied as described above to the factors such that the estimated covariance structure of all the factors is identical for all trials.¹⁹

A second simulation data set is then constructed with the residual variance scaled such that the average multiple correlation coefficient across all variables is .27, which is more in line with what might be expected with say historical weekly return data. The average number of iterations required for convergence increase to approximately 60 iterations. With $p = 200$, all trials converged to non-zero estimates of σ_j and identical values of the log likelihood and factor covariance structures. While the simulation results are of course subject to the particular simulation construction, it would seem reasonable to expect similar results with actual data for a model with a similar average multiple correlation coefficient. One caveat to note, however, is that the simulated data had homogeneous errors by design. The presence of heteroskedasticity in actual return data, may present more of a problem.

¹⁹The rotation of the unconstrained statistical factors is based on the eigenvalues of the predicted factor covariance matrix, $B_c^{*T} Z_c^{*T} Z_c^* B_c^*$.

We also examine the performance of the algorithm using historical daily return data on 500 stocks for 499 days. An equally weighted return index is created as the single "economic" data series in addition to a unit vector in Z_a . The previous day's return for each stock is used as the single "fundamental/technical" data item, B_{bi} . Three statistical factors, one with B_c constrained and two unconstrained are also estimated. The initial estimates of the parameters are constructed as in the above simulations.

In four of the ten trials, the algorithm converged to a value of the likelihood function of 237263.1765710403 and to a value of 237219.4498132423 in the remaining six trials, indicating the presence of at least two local maxima of the likelihood surface. We estimated the model again using the variance estimate from the residuals of the B_a regressions for the initial value of the residual variance. The initial values of the unconstrained factor loadings, B_c^* , are estimated by regressing Y on the first two principal components obtained from the residual covariance matrix from same B_a regressions. With these initial parameter values, the algorithm converges to the larger of the two values of the log likelihood function given above.

Model Extensions:

The assumption of non-stochastic factors results in the estimates of Z_b and Z_c being treated as parameters and thus using $n(q_b + q_c)$ degrees of freedom. The model could be extended to allow for a normal prior on the factors which would involve only the $q_b + q_c$ estimated means of the factor series and the estimated factor covariance parameter, R . This reduction in the size of the parameter space should reduce the tendency to estimate zero variances in small data sets as well less of a tendency to create factors that over fit the data. The CM algorithm above would now become an ECM algorithm as described in Meng and Rubin (1993).

One of the prominent characteristics of security return data is that often some of the returns are missing, particularly when using daily data. Many approaches to estimating factor models use ad hoc procedures for dealing with missing data points. The EM algorithm can be easily adapted to account for (partially) missing returns in Y as well under certain assumptions regarding the missing data process.²⁰

As mentioned earlier, a characteristic of security return data is the presence of occasional "outliers". The above algorithms account for heteroskedastic errors when the variance scaling

²⁰For example, the factor estimation procedures used by EM Applications Ltd. are consistent with maximum likelihood estimation under the assumption that the data are missing at random as defined by Rubin (1976).

parameter, v_{ij} , is known a priori. The above algorithms can be extended to simultaneously estimate the variance scaling process under a variety of assumptions. Obviously it is not possible to estimate each v_{ij} as a parameter since the model would be under-identified. In the case of the standard time-series regression model, the assumption that the disturbances arise from say a two-component mixture-normal model results in an iteratively re-weighted least-squares (IRLS) *EM* algorithm of Dempster, Laird, and Rubin 1980. A similar approach can be taken to extend the above models as well.

Conclusion:

We present a combined linear factor model that incorporates all three basic factor model types, economic, fundamental, and statistical, in a single simultaneous estimation procedure. The algorithm is consistent with the general theory of *EM* estimation of DLR1977 under the assumption of non-stochastic factors, but can be extended to several forms of stochastic factor models as well. Under the assumption of non-stochastic factors, the algorithm simplifies to conditional maximization algorithm. Several extensions of the model are presented, allowing for heteroskedastic errors and a separable response function for the fundamental data types. Under appropriate assumptions, the parameter estimates are maximum likelihood estimates. Simulation results show that the iterations of the algorithm result in monotonic increasing values of the log likelihood function. We find an example of at least two local maxima in the case of historical return data. While improving the initial parameter estimates result in convergence to the larger of the two value of the log likelihood function, we can not be certain that we have in fact obtained the global maximum of the log likelihood function. We note several areas for extentions of the model.

While the above algorithm provides a theoretical framework for estimating a combined linear model, the major elements of estimating such a model will be in the construction of sound economic and fundamental data series. We would also expect that the inclusion of both data types will result in a significant degree of multicollinearity in the resulting factors. Additional work would need to be done to reduce the size of the data sets to avoid issues of over-fitting the data set. The estimation of the statistical factors can be controlled to avoid collinearity issues with the economic and fundamental leaving only the issue of how many additional factors to estimate. Additional research will be required to determine if the performance of the combined model exceeds each of the individual model approaches.

References

- [1] Anderson, T. W., and Herman Rubin, (1956), “Statistical inference in factor analysis”, Berkeley Symposium (5) pp. 111- 150.
- [2] Brown, Stephan J., and Mark I. Weinstein (1983), “A New Approach to Testing Asset Pricing Models: The Bilinear Paradigm”, *Journal of Finance*, Vol 38, pp. 711-43.
- [3] Conner, Gregory, (1995), “Three Types of Factor Models: A Comparison of Their Explanatory Power.” *Financial Analysts Journal* 51 (May/June 1995): 42-46.
- [4] Connor, Gregory and Robert A. Korajczyk (1986), “Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis.” *Journal of Financial Economics* 15 (March 1986): 373-394.
- [5] 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.
- [6] Dempster, Arthur P., N.M. Laird, and Donald B. Rubin. (1980), “Iteratively Re weighted Least Squares for Linear Regression when Errors are Normal/Independent Distributed.” *Multivariate Analysis – V* North-Holland Publishing.
- [7] G. Feeney and D. Hester, “Stock market indices: A principal component analysis,” Cowles Foundation, vol. Monograph 19(39), pp. 110–138, 1967.
- [8] Johnson, Richard A., and Dean W. Wichern. (1988), *Applied Multivariate Statistical Analysis*, Prentice-Hall, New Jersey.
- [9] Jöreskog , K. G. (1969), “A General Approach to Confirmatory Maximum Likelihood Factor Analysis.” *Psychometrika*, Vol. 34, 183-202.
- [10] King, Benjamin F., “Market and Industry Factors in Stock Price Behavior” *Journal of Business* 39, No. 1, January 1966, pp. 139-190.
- [11] Lawley, D. N. and A. E. Maxwell. (1971), *Factor Analysis as a Statistical Method*, Second edition. London: Butterworth.
- [12] Lawley, D.N. (1942), “Further estimation in factor analysis.”, *Proceedings of the Royal Society of Edinburgh, Series A*. v.61, pp.176-185, 1942.

- [13] Little, Roderick J. A. and Donald B. Rubin. (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.
- [14] Liu, C., Rubin, B. D. and Wu, Y. (1998). "Parameter expansion to accelerate EM the PX-EM algorithm" *Biometrika* 85 pp. 755-770.
- [15] Meng, Xiao-Li; Rubin, Donald B., "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika* Vol. 80, 1993.
- [16] Roll, Richard W. and Stephen A. Ross. (1980), "An Empirical Investigation of the Arbitrage Pricing Theory." *Journal of Finance*, Vol. 35, pp. 1073-1103.
- [17] Rosenberg, Barr, "Extra-Market Components of Covariance in Security Returns." *Journal of Financial and Quantitative Analysis* 9 (March 1974): 263-274.
- [18] Ross, Stephen A., "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13 (December 1976): pp. 341-360.
- [19] D.B. Rubin (1976), "Inference and missing data." *Biometrika* 63, pp. 581-592.
- [20] Rubin, Donald B. and Dorothy T. Thayer. (1982), "*EM* Algorithms for ML Factor Analysis." *Psychometrika*, Vol. 47, pp. 69-76.
- [21] Rubin, Donald B. and Dorothy T. Thayer. (1983), "More on *EM* for ML Factor Analysis." *Psychometrika*, Vol. 48, pp. 253-257.
- [22] Stroyny, Alvin L. (1991), "Heteroskedasticity and the Estimation of Systematic Risk", Ph.D. Dissertation, The University of Wisconsin - Madison.
- [23] Stroyny, Alvin L. and Daniel Rowe (2002), "A re-examination of some popular latent factor estimation methods." Presented at the 15th Annual Investment Seminar, Cambridge, UK.
- [24] Whittle, P. (1952), "On Principal Components and Least Square Methods of Factor Analysis.", *Skandinavisk Aktuarietidskrift* 1952.