2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Sparse factor analysis via likelihood and ℓ_1-regularization

Lipeng Ning and Tryphon T. Georgiou

***Abstract*—In this note we consider the basic problem to identify linear relations in noise. We follow the viewpoint of factor analysis (FA) where the data is to be explained by a small number of independent factors and independent noise. Thereby an approximation of the sample covariance is sought which can be factored accordingly. An algorithm is proposed which weighs in an ℓ_1-regularization term that induces sparsity of the linear model (factor) against a likelihood term that quantifies distance of the model to the sample covariance. The algorithm compares favorably against standard techniques of factor analysis. Their performance is compared first by simulation, where ground truth is available, and then on stock-market data where the proposed algorithm gives reasonable and sparser models.**

I. Introduction

We are interested in a most basic problem where linear models are sought from noisy data [10]. The salient feature of the problem, however, is that it involves typically a large number of variables. This of great importance in many areas in science, engineering, but most notably in econometrics and economic forecasting [4].

Consider the (column) vector

$$x = [x_1,\ \ldots, x_n]' \in \mathbb{R}^n$$

of real random variables $x_1, \ldots, x_n$, and a set of independent realizations of $x(1), \ldots, x(m)$ of the vector x making up

$$X = [x(1), \ldots, x(m)] \in \mathbb{R}^{n\times m}.$$

We assumed that these have zero mean and we set

$$\Sigma_{\text{sample}} = \frac{1}{m} XX'$$

to be the sample covariance.

From this point, various modeling postulates can be considered. Each makes different assumption about the nature of noise. Typically, one assumes that

$$x = \hat{x} + \tilde{x}$$

where $\hat{x}$ represents a noise-free variable and $\tilde{x}$ represents noise. Then, $\hat{x}$ and $\tilde{x}$ are assumed uncorrelated zero-mean (vectorial) random variables. The classical Frisch model postulates that

$$\tilde{\Sigma} := E(\tilde{x}\tilde{x}')$$

is diagonal whereas principle component analysis (PCA) assumes that $\tilde{\Sigma}$ has no structure whatsoever, but that the variance $\text{trace}(\tilde{\Sigma})$ is small.

Regarding the structure of the noise-free variables, it is assumed that these satisfy between them a number of linear relations. Thus, the rank of the noise-free covariance

$$\hat{\Sigma} = E(\hat{x}\hat{x}')$$

must be less than n. This $\text{rank}(\hat{\Sigma})$ is the number of "principle components" or "factors" which are needed to explain the data.

In the PCA literature, the rank is conveniently chosen according to a "break point" of the singular values of Σ_{sample}, while in the context of the Frisch problem, it is natural to seek $\hat{\Sigma}$ of minimal rank [10]–a problem which to large degree is still open. Factor analysis on the other hand focuses not not only on the number of independent components, which need to be small, but also on the structure of the postulated noise-free covariance $\hat{\Sigma}$. More specifically, it is assumed that

$$X = FV + \tilde{X}$$

and that the noise-free data $\hat{X} = FV$ is modeled by a constant coefficient matrix F, which is required to be sparse, and a matrix

$$V = [v(1), \ldots, v(m)],$$

where $v(t) = [v_1(t), \ldots, v_r(t)]'$, $t = 1, \ldots, m$ represent independent realizations of a set of random variables. The variables v_j are called the *factors* and F is the *factor loading* in factor analysis.

The underlying rationale for sparse factor loadings is so that each random variable $\hat{x}_i$ is accounted for by superimposing the effect of a small number of factors v_j. The assumption on $\tilde{x}$ is again that the entries are independent and hence that the covariance $\tilde{\Sigma}$ is diagonal. Further, without loss of generality, the covariance of v can be taken to be the identity. Thus the basic problem in factor analysis is as follows.

Problem: Given an $n \times n$ sample covariance Σ_{sample} and an integer $r < n$, determine a diagonal covariance $\tilde{\Sigma}$ and a sparse coefficient matrix $F \in \mathbb{R}^{n\times r}$ such that

$$\Sigma_{\text{sample}} \simeq FF' + \tilde{\Sigma}. \tag{1}$$

The purpose of this paper is to propose an algorithm for obtaining such a decomposition. The solution is not optimal, i.e., it is neither sparsest nor of least rank. Attaining either of these specs is a formidable objective. Instead, we consider a likelihood function together with a regularization term that promotes sparsity. The algorithm aims to approximate extrema of such a functional.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

II. Algorithm

Suppose that the data is drawn from a Gaussian distribution with covariance Σ. In the absence of any constraints the negative-log-likelihood

$$\frac{m}{2}\left(\text{logdet}(\Sigma) + \text{trace}(\Sigma_{\text{sample}}\Sigma^{-1}) + n\log(2\pi)\right)$$

is miminal at $\Sigma = \Sigma_{\text{sample}}$, while in general, it can be taken to quantify distance between Σ and Σ_{sample}.

Given the sample covariance Σ_{sample}, our interest is in obtaining a nearby value for Σ which admits a factorization as in (1) with a sparse factor F. The sparsity of F, often denoted by $\|F\|_0$, is the number of non-zero entries. It is known that the sum of the absolute values of its entries, namely $\|F\|_1$, is a good surrogate for sparsity. The use of the ℓ_1 norm for penalty to promote sparsity has a history in statistics, but was brought to prominence only recently, after a series of deep studies by Candès, Tao, Romberg, Donoho, Elad, and others. The area of these contributions is now commonly known as compressive sensing.

Thus, for our purposes, we propose as a compromise between fit and sparsity, a linear combination of the likelihood function and a weighted ℓ_1-norm of the corresponding factor. This linear combination of likelihood and ℓ_1-cost is thought of as a function of the entries F, $\tilde{\Sigma}$ in the decomposition

$$\Sigma = FF' + \tilde{\Sigma}$$

with $\tilde{\Sigma} \geq 0$ and diagonal. Thus, our problem can be expressed as follows:

$$\min_{F,\tilde{\Sigma}}\{\text{logdet}(FF'+\tilde{\Sigma}) + \text{trace}(\Sigma_{\text{sample}}(FF'+\tilde{\Sigma})^{-1}) + \lambda\|F\|_1 \mid \tilde{\Sigma} \geq 0,\ \tilde{\Sigma} \text{ is diagonal}\}. \quad (2)$$

The weight λ affects sparsity of the factor F.

Problem (2) is not convex. Since

$$(FF'+\tilde{\Sigma})^{-1} = E - GG'$$

with

$$\begin{aligned} E &= \tilde{\Sigma}^{-1}, && (3a)\\ G &= \tilde{\Sigma}^{-1}FM^{-1}, && (3b)\\ M &= (I + F'\tilde{\Sigma}^{-1}F)^{\frac{1}{2}} \\ &= (I - G'E^{-1}G)^{-\frac{1}{2}}, && (3c) \end{aligned}$$

then (2) becomes

$$\min_{G,E}\{-\text{logdet}(E-GG') + \text{trace}(\Sigma_{\text{sample}}(E-GG')) + \lambda\|E^{-1}GM\|_1 \mid E - GG' \geq 0,\ E \text{ is diagonal}\}. \quad (4)$$

For $G = G_k + \delta_G$, the first order approximation of GG' is

$$\delta_G G_k' + G_k\delta_G' + G_k G_k' =: [GG']_{k,\delta_G}. \quad (5)$$

Likewise, let $E = E_k + \delta_E$ where G_k, E_k for $k = 1, 2, \ldots$ represents iteratively obtained values for G and E (as in (6b-6c) below), respectively. Also M and F will be approximated iteratively. Since, from (3b) $F = E^{-1}GM$ we approximate F with the first order perturbation

$$E_k^{-1}(G_k + \delta_G)M_k,$$

where M_k represents the value of M from (3c) at E_k and G_k. In this expression, we neglect the linear perturbation terms corresponding to E and M. Our reasoning is as follows: first, E is diagonal and neglecting corrections will not affect the sparsity pattern of F while on the other hand the linear perturbation term involves E_k^{-2} which causes the iteration to be numerically sensitive. Then again, the linear perturbation in M is rather complicated (given in Appendix B) and the neglected term is not affecting the convergence claimed in the proposition below. Define

$$\begin{aligned} g_{G_k,E_k}(\delta_G,\delta_E) = &-\text{logdet}(E_k + \delta_E - [GG']_{k,\delta_G}) \\ &+ \text{trace}(\Sigma_{\text{sample}}(E + \delta_E - [GG']_{k,\delta_G})) \\ &+ \lambda\|E_k^{-1}(G_k+\delta_G)M_k\|_1 \end{aligned}$$

which is convex in δ_G and δ_E and approximates (4) at G_k and E_k. Moreover, the constraint $E - GG' \geq 0$ is equivalent to

$$\begin{bmatrix} E & G \\ G' & I \end{bmatrix} \geq 0.$$

Hence, we seek minima of (4) by solving

$$\begin{aligned} (\hat{\delta}_G, \hat{\delta}_E) = \ &\text{argmin}\,\{g_{G_k,E_k}(\delta_G,\delta_E) : \\ &\begin{bmatrix} E_k+\delta_E & G_k+\delta_G \\ G_k'+\delta_G' & I \end{bmatrix} \geq 0, \\ &\text{and } \delta_E \text{ diagonal}\}. \end{aligned} \quad (6a)$$

We choose step size $\alpha \in [0,\ 1]$ such that with

$$\begin{aligned} G_{k+1} &= G_k + \alpha\hat{\delta}_G && (6b)\\ E_{k+1} &= E_k + \alpha\hat{\delta}_E && (6c) \end{aligned}$$

the following inequality holds

$$g_{G_{k+1},E_{k+1}} < g_{G_k,E_k} - \sigma. \quad (6d)$$

Here, the constant $\sigma > 0$ determines a stopping criterion. Also, we have simplified the notation by setting

$$g_{G_k,E_k} := g_{G_k,E_k}(0,0).$$

If (6d) holds, the step size α is accepted, otherwise let $\alpha = \frac{1}{2}\alpha$. If the step size is smaller than a preselected $\epsilon > 0$, the iterations can be terminated.

Proposition: If $\lambda = 0$, then at point G_k, E_k, $(\hat{\delta}_G, \hat{\delta}_E)$ is descent direction for (4).

Proof: Since $\delta_G = \delta_E = 0$ satisfy the LMI in (6),

$$g_{G_k,E_k}(\hat{\delta}_G,\hat{\delta}_E) \leq g_{G_k,E_k}.$$

Moreover, $g_{G_k,E_k}(\delta_G,\delta_E)$ is convex in δ_G and δ_E, so

$$g_{G_k,E_k}(\epsilon\hat{\delta}_G, \epsilon\hat{\delta}_E) \leq g_{G_k,E_k}$$

for all $\epsilon \in [0,\ 1]$. If ϵ is small, the difference between the above two is of order ϵ or higher. When $\lambda = 0$, because of

(5), the difference between $g_{G_k,E_k}(\epsilon\hat{\delta}_G, \epsilon\hat{\delta}_E)$ and the objective function in (4) is of order ϵ^2 or higher. So $(\hat{\delta}_G, \hat{\delta}_E)$ is descent direction for (4) at E_k, G_k. ∎

If $\lambda > 0$ and is small enough, the decreased amount of the first two terms dominates the change of the last term. So the pair $(\hat{\delta}_G, \hat{\delta}_E)$ is still a descent direction. Since the objective function (4) is bounded below, then the algorithm will at least converge to a local minimum.

The algorithm can be initialized using suitable starting values for $\tilde{\Sigma}_0$ and F_0, and then setting

$$\begin{aligned} M_0 &= (I + F_0\tilde{\Sigma}_0^{-1}F_0)^{\frac{1}{2}} \\ E_0 &= \tilde{\Sigma}_0^{-1} \\ G_0 &= \tilde{\Sigma}_0^{-1}F_0M_0^{-1}. \end{aligned}$$

A starting value for F_0 can be chosen to contain the eigenvectors of Σ_{sample} corresponding to the largest r eigenvalues, scaled by the square root of corresponding eigenvalues, and $\tilde{\Sigma}_0$ may be set equal to the diagonal [1] of $\Sigma_0 - F_0F_0'$.

III. Background and the Varimax criterion

The problem to identify linear relations in data has its roots at least as far back as in the work of Gauss. "Least squares" has been a workhorse in engineering ever since. Early in the 20th century, following Ragnar Frisch, statisticians laid down alternative assumptions on the noise model and sought to understand the impact of such concepts on modeling.

There are several schools of thought. Most prominently, principal component analysis (PCA) which is based on the fact that singular value decomposition (SVD) allows for an exact and computationally simple analysis of data and covariance according to the hypothesis that noise has no structure while the signal-to-noise ratio is significant. In parallel, in disciplines such as psychometrics and econometrics where data is often dominated by noise [11], [16], assuming a more detailed noise-model is essential. Reasonable hypotheses often allow for more accurate models, albeit at a cost of an often computationally intractable problem. The main schools are that of Factor Analysis (FA) and Errors in variables (EIV). FA is based on the assumption of the independence of noise whereas EIV allow for more sophisticated models and is a broader research area [13], [14]. The assumption of independent noise components is natural in several applications in signal analysis and system identification, but most importantly in financial data. In [2] the sparse factor model was assumed in the study of gene expression genomics where a Bayesian work was introduced. In [3] the same objective function, penalized maximum likelihood function, was considered for the sparse factor analysis problem and a generalized expectation maximization algorithm was proposed. A recent study in [17] to achieve sparse PCA blares the distinction as it assumes a noise model with a diagonal structure. Further, in [17] a similar weighing of a likelihood function together with an ℓ_1-penalty is being proposed. See also [18] for an alternative view to sparse PCA. Dynamic modeling can also be formulated in a variety of ways [1], [4]. One way that this can be achieved is by seeking linear relations between time-shifted copies of time-series data and will not be discussed further. The problem to sparsify factor loadings has been a central issue in FA. A standard approach is to search for an $r \times r$ rotation matrix R, in conjunction with a search for the $n \times r$ factor loading F, so that the product FR is sparse. Cumbersome as it may seem, this viewpoint forms the basis of a technique presented in [9], [12]. The rotation matrix R is sought to maximize the following *Varimax* criterion:

$$\sum_{j=1}^{r}\left(\sum_{i=1}^{n}(FR)_{ij}^4 - \frac{1}{n}(\sum_{i=1}^{n}(FR)_{ij}^2)^2\right).$$

This expression is precisely the sum of variances of the squares of the elements of FR. The rationale for seeking extrema of this expression rests on the observation that the Frobenius norm of FR is not affected by R, which is a rotation matrix. Then, if this is chosen so as to render the variance of the *squares* of the entries large, FR will necessarily have most entries small. This criterion is now standard and is included as part of Matlab in the subroutine `factoran` which is widely used for factor analysis.

IV. Examples

Example 1: First we compare the performance of the proposed method against `factoran` (which uses the Varimax criterion) using simulation. To this end, we generate a random factor loading matrix $F \in \mathbb{R}^{100\times 10}$ with only two nonzero entries per row, and a realization $V \in \mathbb{R}^{10\times 400}$ of random factors from a multivariate normal distribution with covariance equal to the identity. Figure 1 displays a color-coded representation of the entries of F, side by side with the entries of the estimated factor loadings using our algorithm (second column) and one produced by `factoran` (third column). To highlight the difference we display in Figure 2 the same, as a binary plot, thresholding at a low amplitude. This shows an almost perfect agreement of the result of our algorithm with the "ground truth." Finally, Figure 3 shows the relative values of entries across one of the columns of the factor loading matrix F (5^{th} column) and compares, likewise, the "ground truth" with the result of the algorithm presented in this paper and the output of factoran. It is seen that both algorithms are consistent.

Example 2: We process time-series data corresponding to 30 different stocks taken from 4 different sectors. The stocks and sectors are listed in the Appendix A. The data are taken over a period of four weeks sampled at 2 minute intervals.

[1] This starting choice is the maximum likelihood solution with F_0 given [12]; $\tilde{\Sigma}_0$ may need to be slightly modified so as to be positive definite by adding small diagonal positive entries.

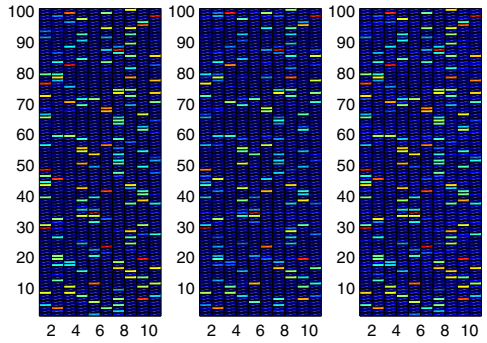

Fig. 1. Factor loadings: i) "ground truth" (left), ii) estimated using proposed method (middle), iii) estimated using `factoran` (right).

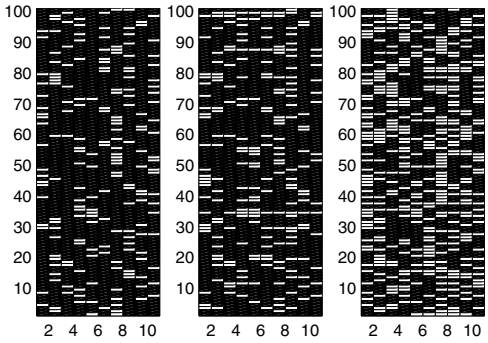

Fig. 2. Binary values for factor loadings: i) "ground truth" (left), ii) estimated using proposed method (middle), iii) estimated using `factoran`, by thresholding at a small value.

The same database has been used in [15]. We compute the correlation matrix which is shown in Figure 4 in color-coded format. From the figure it is possible to discern strong correlation between certain stocks. This in turn may suggest a possible common underling factor. Otherwise noise may be the dominant component. We compute factor loadings using the method that was proposed earlier and compare with the factor loadings obtained using standard factor analysis (`factoran` routine in Matlab). These are displayed in Figure 5, again in a color-coded format. It is evident that the method that we propose gives sparser factor loadings in this example as well.

In general, we observe that both our algorithm and the `factoran` are largely consistent. In this example there is no "ground truth." Inspection of the sample covariance and further analysis can suggest whether values in the factor loadings seem appropriate, although such a claim is difficult to substantiate. Yet, it is apparent from Figure 5 that our algorithm relies less on small values in the factor loading to explain the data. Hence, the factor loading matrix F is sparser.

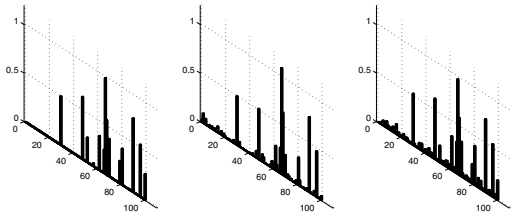

Fig. 3. Bar plots of the 5th column of the true sparse matrix (left), the estimated sparse matrix given by proposed method (middle), and the result by factor analysis (right).

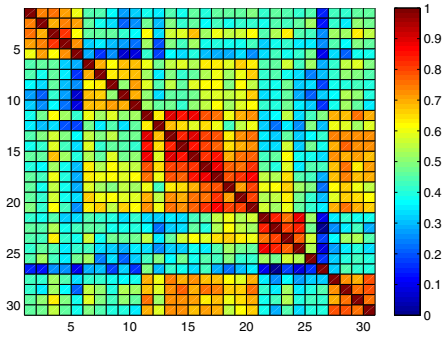

Fig. 4. Sample covariance of 30 stocks.

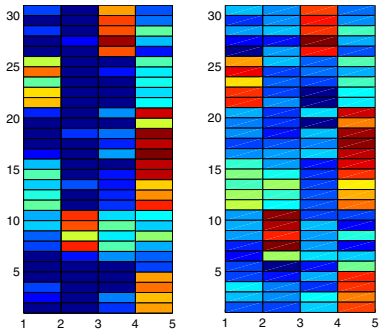

Fig. 5. Factor loadings: i) using the proposed method (left), using `factoran` (right); in both, small values below a threshold are set to zero.

ACKNOWLEDGMENT

The financial-stock data used in Example 2 was provided to us by Dr. Donatello Materassi.

REFERENCES

[1] B.D.O. Anderson and M. Deistler, "Generalized linear dynamic factor models," *Proceedings of the 47th IEEE Conference on Decision and Control*, pp. 1980-1985, 2008.

[2] C.M. Carvalho, J. Chang, J.E. Lucas, J.R. Nevins, Q. Wang and M. West, "High-dimensional sparse factor modeling: applications in gene expression genomics," *J Am Stat Assoc*, **103(484)**: 1438-1456, 2008.

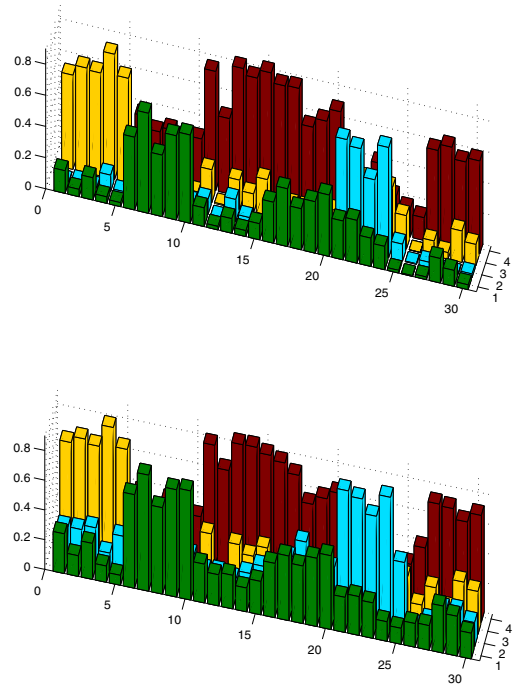

Fig. 6. Factor loadings (values in 3D bar): i) using the proposed method (left), using `factoran` (right); in both, small values below a threshold are set to zero.

[3] J. Choi, H, Zou and G. Oehlert, "A penalized maximum likelihood approach to sparse factor analysis," *Statistics and Its Iterface*, **3(4)**: 429-436, 2011.

[4] M. Deistler and B.D.O. Anderson, "Linear dynamic errors-in-variables models – Some structure theory," *Journal of Econometrics*, **41**: 39-63, 1989.

[5] R. Frisch, "Correlation and scatter," *Nordic Statistical Journal*, 1928.

[6] R. Frisch, "Statistical confluence analysis by means of complete regression systems," *Publication No. 5 of the University Institute of Economics*, Oslo, 1934.

[7] T.T. Georgiou, "Relative entropy and multivariable multidimensional moment problem," *IEEE Transactions on Information Theory*, **52(3)**): 1052-1066, 2006.

[8] K.G. Jöreskog,"Some contributions to maximum likelihood factor analysis," *Psychometrika*, **32(4)**: 443-482, 1967.

[9] H.F. Kaiser, "The Varimax criterion for analytic rotation in factor analysis," *Psychometrika*, **23(3)**: 187-200, 1958.

[10] R.E. Kalman, "System identification from noisy data," *Dynamical System II*, Eds A. R. Bednarek and L. Cesari, pp. 135-164. Academic Press, New York, 1982.

[11] T. Koopmans, "Linear regression analysis of economic time series," *Publication No. 20 of Netherlands Economic Institute*, Haarlem, 1937.

[12] D.N. Lawley and A.E. Maxweel, *Factor analysis as a statistical method*, 2nd Ed. New York: American Elsevier Publishing Co. 1971.

[13] C.A. Los, "Identification of a linear system from inexact data: a three-variable example," *Computers & Mathematics with Applications*, **17(8/9)**: 1285-1304, 1989.

[14] C.A. Los, "The prejudices of least squares, principal components and common factor schemes," *Computers & Mathematics with Applications*, **17**): 1269-1283, 1989.

[15] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Transactions on Automatic Control*, **55(8)**): 1860-1871, 2010.

[16] O. Reiersøl, "Confluence analysis by means of lag moments and other methods of confluence analysis," *Econometrica*, **9(1)**: 1-24, January, 1941.

[17] M.O. Ulfarsson and V. Solo, "Sparse variable PCA using geodesic steepest descent," *IEEE Transactions on Signal Processing*, **56(12)**: 5823-5832, 2008.

[18] D.M. Witten, R. Tibshirani, T. Hastie, "A penalized matrix decomposition, with applications to sparse principle component and cannonical correlation analysis," *Biostatistics*, **10(3)**: 515-534, 2009.

APPENDIX A

Name	Sector	Industry
D	Utilities	Electric Utilities
DTE	Utilities	Electric Utilities
DUK	Utilities	Electric Utilities
ED	Utilities	Electric Utilities
EIX	Utilities	Electric Utilities
ADI	Technology	Semiconductors
ALTR	Technology	Semiconductors
AMAT	Technology	Semiconductors
AMD	Technology	Semiconductors
BRCM	Technology	Semiconductors
CMA	Financial	Regional Banks
FHN	Financial	Regional Banks
FITB	Financial	Regional Banks
HBAN	Financial	Regional Banks
KEY	Financial	Regional Banks
MER	Financial	Investment Services
MS	Financial	Investment Services
NYX	Financial	Investment Services
SCHW	Financial	Investment Services
TROW	Financial	Investment Services
APA	Energy	Oil & Gas Operations
APC	Energy	Oil & Gas Operations
CHK	Energy	Oil & Gas Operations
DVN	Energy	Oil & Gas Operations
DYN	Energy	Oil & Gas Operations
ABK	Financial	Insurance (Prop.& Casualty)
ACE	Financial	Insurance (Prop.& Casualty)
AIG	Financial	Insurance (Prop.& Casualty)
ALL	Financial	Insurance (Prop.& Casualty)
CB	Financial	Insurance (Prop.& Casualty)

APPENDIX B

Given A and Δ,

$$e^{A+\Delta} = e^A + \int_0^1 e^{(1-\tau)A}\Delta e^{\tau A} d\tau + o(\|\Delta\|), \quad (7a)$$

$$\log(A+\Delta) = \log A + \int_0^\infty (A+\tau I)^{-1}\Delta(A+\tau I)^{-1} d\tau + o(\|\Delta\|), \quad (7b)$$

see e.g. [7]. Denote $\Delta := G_k' E_k G_k - G' E G$ and write

$$M = (I - G'EG)^{-\frac{1}{2}} = (I - G_k' E_k G_k + \Delta)^{-\frac{1}{2}} = e^{-\frac{1}{2}\log(I - G_k' E_k G_k + \Delta)}.$$

Then from (7)

$$M = M_k - \frac{1}{2}\int_0^1 \int_0^\infty M_k^{1-\lambda}(I - G_k' E_k G_k + \tau I)^{-1}\Delta \times (I - G_k' E_k G_k + \tau I)^{-1} M_k^{\lambda} d\tau d\lambda + o(\|\Delta\|).$$