Stony Brook University

Log Band Fraction Approximation For Covariance Estimation and Low Volatility Strategy

A Dissertation Presented

by

Riyu Yu

to

The Graduate School

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Quantitative Finance

Stony Brook University

August 2015

Stony Brook University

The Graduate School

Riyu Yu

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Andrew Mullhaupt – Dissertation Advisor
Research Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Svetlozar Rachev – Chairperson of Defense
Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Haipeng Xing – Committee Member
Associate Professor
Department of Applied Mathematics and Statistics, Stony Brook University

Keli Xiao – External Committee Member
Assistant Professor
College of Business, Stony Brook University

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Log Band Fraction Approximation For Covariance Estimation and Low Volatility Strategy

by

Riyu Yu

Doctor of Philosophy

in

Applied Mathematics and Statistics

Quantitative Finance

Stony Brook University

2015

Structured matrix plays an important role in statistics, especially in covariance estimation. Band fraction representation is one of the efficient structures for matrices. In this dissertation, we study the metric tensor for the band fraction representation for the covariance matrix. We propose a new structure, the log band fraction representation, which gives smaller information distance and Hellinger distance than factor model and band fraction representation. We apply the log band fraction estimation in the portfolio optimization problem. We propose our long only strategy and 130-30 strategy, which significantly outperform the benchmarks, i.e., SPY, SPLV, and CSM. Transaction cost is considered in the portfolio construction process. The strategies proposed in this dissertation are fully investable.

Key Words: Log Band Fraction; Fisher Information; Long Only Portfolio; 130-30 Fund; Low Volatility Strategy

Contents

List of Tables

List of Figures

Acknowledgements

I need to first thank my academic advisor Prof. Andrew P. Mullhaupt. He was the one opened the door of quantitative finance for me. Under his guidance, I started diving into the mathematics world, from matrix analysis, signal processing to convex optimization, information geometry, etc. He is always available and patient for any discussion with students. He was always able to express his deep insight of mathematics and finance in a simple and interesting way. He was also the one keeps students aware of the importance of working with financial data. He directed the developing and the maintenance of the Trade and Quote database in the program, which benefits not only the students within the program, but also researchers across the whole university. With his over twenty years' experience as an extremely successful portfolio manager in Wall Street, he is one of the few professors who knows how to correctly apply the mathematics tools to the real fiance world. He teaches me not only his knowledge about math, but also his attitude doing research work. Starting from there I realize that I need to always keep the passion and curiosity of learning mathematics. Without any doubt, Prof. Mullaupt is the most popular Professor in the department and beloved by all students. I am thankful to all his help and kindness during my Ph.D. study and feel so fortunate that I had the opportunity to closely learn from him.

I want to thank Prof. Svetlozar Rachev for his help during my study. His guidance helps me study not only the course work but also the enormous academic world. His knowledge of risk management helps me broaden my understanding of the whole financial structure. Prof. Rachev spent huge effort directing the program. I started with zero background of quantitative finance when I was admitted to the program. Prof. Rachev's lectures, from stochastic calculus to portfolio theory, Bayesian methods, capital markets, helps me build a solid foundation for my further study.

I also want to thank Prof. Robert Frey for starting such a great program. I want to thank Prof. John Pinezich for his kindness and help during my research. I would like to express my appreciation to colleagues and friends including (but not limited to) Angela Tsao, Xu Dong, Tengjie Jia, Xiaoping Zhou, Barret Shao, Xiang Shi, Ke Zhang, Jaehyung, Choi, Michael Tiano, Rong Lin, for all the discussions and help during my Ph.D. life.

Last but most, I want to thank my parents, and all the family members, for all the unconditional support, without whom I would not succeed under any circumstance.

Vita, Publications and/or Fields of Study

I was born in Wenzhou, Zhejiang, P.R. China in 1988. I received my B.S. in Electrical Engineering at Beijing Institute of Technology in 2010.

I was admitted to the master program in Applied Mathematics and Statistics at Stony Brook University in 2011. One year later, I jointed to the Ph.D. program. My concentration is quantitative finance and my academic advisor is Professor Andrew P. Mullhaupt. My research interest include structured matrix analysis, fast algorithms for quadratic programming, adaptive digital filters, covariance estimation and shrinkage. The related research work include two working papers:

1. Riyu Yu, Andrew P. Mullhaupt. Log band fraction representation of covariance matrix. Working paper. (2015)

2. Riyu Yu, Andrew P. Mullhaupt. Low volatility strategy with log band fraction approximation. Working paper. (2015)

1 Introduction

Structured matrix plays an important role in mathematics and computations. Exploiting the matrix structure, such as sparsity, low rank property, Toeplitz, is the key to the design of faster and more efficient algorithms. By a structured matrix, typically we mean an $n \times n$ matrix with entries having a formulaic relationship, allowing the matrix to be specified by significantly fewer than n^2 parameters [1].

In numerical analysis, a sparse matrix is a matrix in which most elements are zeros. Conceptually, sparsity corresponds to the systems which are coupled loosely. It is beneficial to take advantage of the sparse structure of the matrix when storing and manipulating a sparse matrix. The general algorithms used for dense matrix may be inefficient when applied to large sparse matrices. Strassen [2] was the first to show that the naive matrix multiplication algorithm is not optimal and gave an $O(n^{2.81})$ algorithm. Since then many improvements followed. For example, several fast matrix multiplication algorithms can be found in Pan [5], Burgisser [4], Coppersmith and Winograd [3], Cohn and Umans [7]. Sparse approximation is a problem of estimating a sparse multidimensional vector given high dimensional observed data. It reduces the number of variables from high dimension space to the low dimension space by using the sparse structure. Lots of sparse approximation algorithm have been developed, such as matching pursuit [8] [9], orthogonal matching pursuit [10] and LASSO method [11].

Band matrix is a special type of sparse matrix whose non-zero entries are confined to a diagonal band. A band matrix can be linked in complexity to a rectangular matrix whose row dimension is equal to the bandwidth of band matrix. The amount of operations such as multiplications on band matrix falls so significantly that huge time and complexity would be saved. As sparse matrices lead themselves to more efficient algorithms than dense matrices, there has been much research that are focused on utilizing the band structure to improve efficiency and also finding approximations to the original high dimension system. For instance, Gilbert [12] shows the factorizations of banded matrices with banded inverses. Pan [6] devises parallel algorithms for solving banded linear system of equations and computing the determinant of a banded matrix. Remon [15] presents the Cholesky factorization of band matrices. Martin [16] proposes the symmetric decomposition of positive definite band matrices. Another example is Hessenberg matrix which is a special case of band matrix and finds itself wide applications in signal processing. [13] [14].

Low grade matrix is first introduced by Andrew P. Mullhaupt and Kurt Riedel [27]. They show that low grade matrix has fraction representation with two band matrices. Another decomposition called consecutive sub-block product representation is also given. The low grade structure is applied to the triangular input balance system [28], which is shown to have lots of advantages such as better numerical stability, when compared to other digital filters such as Burg and lattice filter. Semiseparable matrix has similar structure as low grade matrix. Chandrasekaran and Shiv [31] [32] [33] [34] [35] propose fast solvers of the linear equations by exploiting the hierarchically semiseparable representations (HSS). These representations are useful for the matrices whose off-diagonal blocks have small rank. Ming Gu [37] [38] [39] approximates the positive definite matrices with the semiseparable matrices with efficient and robust algorithms, by modifying the classical Cholesky decomposition procedure with low rank approximation.

Covariance estimation plays a fundamental role in multivariate statistical analysis [26] [40] [52] [57] [58] [59] [65] [67] [40] [41]. Estimation of covariance matrix is a problem of how to approximate the actual covariance matrix with the sample data from the true distributions. Factor model is widely used in the covariance estimation [54] [56] [76] [84]. It is a statistical method describing variability between observed variables and lower number of unobserved variables which are called factors. The observed variables are represented as the linear combination of the unobserved factors, plus the noise term. One advantage of factor model is that it is a reduction method which reduces the high dimension data to a lower dimension space with potential smaller amount of factors, which means computationally it is equivalent to low rank approximation of the actual covariance matrix.

Information geometry is a branch of mathematics that applies differential geometry to the field of probability theory and provides analysis for a wide range of domains such as information theory, statistical inference and neurocomputing. Rao [43] [44] [45] [46] [47] did extensive study on the information geometry of statistical models. He proposed a method for measuring distances between distributions in a parametric family. Rao's measure is based on the metric of Riemannian geometry, which is in terms of the elements of the Fisher information matrix. The Fisher-Rao metric and Kullback-Leibler divergence may be used to model experimental data. There is subsequential research after Rao's pioneering work [43] [17] [18]. Current applications of information geometry in statistics include the dimensionality reduction problem on statistical manifolds, as well as the preparation of

samplers for sequential Monte Carlo technique [19].

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model, which was recommended, analyzed and popularized by R. A. Fisher between 1912 and 1922. Much of the theory of maximum likelihood estimation was first developed for Bayesian statistics. The maximum likelihood estimation problem often involves convex optimization which includes Newton's method, conjugate gradient method and other fast optimization algorithms. Fisher scoring algorithm is a form of Newton's method used to solve maximum likelihood equations numerically. In the Fisher scoring algorithm, the Hessian of the objective function, which is required by Newton's method, is replaced by the Fisher information matrix.

Hankel matrix and Toeplitz matrix are another types of structured matrices of great importance arising naturally in the context of linear systems and signal processing [99] [119] [120] [121]. Hankel matrix and Toeplitz matrix are closely related to the impulse response of the linear time invariant systems, which lead to various applications such as system identification, model reduction and fast multiplication. AAK theorem, which is proposed by Adamjan, Arov and Krein, gives the best Hankel norm approximation to a given Hankel matrix [20]. Other reduction methods involves approximating Hankel matrix with respect to optimal L_2 norm and Frobenius norm [23], or solving Hankel matrix approximation using semedefinite programming [21], etc.

Recently, there is a fast growing interest in low volatility strategies. Lots of investors are looking for strategies which can provide protection in a volatile market and smaller drawdowns. Though the return could be possibly lower, the Sharpe ratio of the low volatility strategy could be potentially higher since the diversification and risk control are improved. SPLV, one of the low volatility ETFs, is launched on May 5, 2011. It is the first low volatility ETF in the market. The fund invest in the 100 symbols from S&P 500 with lowest realized volatility. It is shown that SPLV achieves higher Sharpe ratio than SPY and provide a good alternative to SPY since its low volatility attribute. Another popular class of strategies is called 130-30 strategy which extend the long only portfolio to long-short portfolio. A 130-30 portfolio has 130% of its capital in long position and 30% of its capital in short position, which typically uses a leverage ratio of 1.6. Thus, the 130-30 portfolio can be viewed as a market-neutral portfolio plus a long only portfolio, which provides hedge when market is falling. The Credit Suisse 130/30 index, which based on Andrew Lo and Pankaj Patel's [22]

construction, is a widely used 130/30 index, with corresponding ETF called CSM.

The organization of this dissertation is as following. In section 2, we review the low grade matrix and semiseparable approximation algorithms. The block band fraction representation is presented. In section 3, we show the calculation of the first and second derivatives of the log likelihood function with band fraction representation. Fisher information matrix of the band fraction representation is given, as well as the metric tensor. In section 4, we calculate the metric tensor of the factor model with respect to the complete data. We give the information distance and upper bound, based on different choices of diagonal matrix and factor loadings of the factor model. In section 5, we review several optimization methods such as maximum log likelihood estimator, Newton's method, Fisher scoring algorithm. Conjugate gradient method is shown can be used to improve the band fraction representation. We propose the log band fraction representation for the covariance matrix. The empirical simulations are given to show that we have better estimator with the log band fraction representation. In section 6, we apply the log band fraction technic to the portfolio optimization. We construct long only strategy and 130-30 strategy and compare to some widely used benchmarks, i.e., SPY, SPLV, CSM. We show that our long only strategy and 130-30 strategy are significantly outperforming the benchmarks. The transaction cost are included in the portfolio construction. The strategies we proposed are fully investable. In section 7, we give a conclusion of this thesis.

2 Band Fraction Representation

Factor models are widely used for structured covariance estimation. However, there are other structured estimation methods as well. As suggested by Tengjie Jia [29], the band fraction representation of covariance matrix would be a more accurate and faster method than factor model. In this section, we will first review the definition of low grade matrix and band matrix. We will show Ming Gu's semiseparable approximation algorithm and then propose the block decomposition from the semiseparable representation.

2.1 Low Grade Matrix

The lower grade (lgrade) of an $n \times n$ matrix A is the largest rank of any lower subdiagonal block of a symmetric partition of A [27]. The low grade matrix A can be approximately decomposed as the sum of an upper triangular matrix U and a rank d matrix d

$$A = U + V$$

A d-grade matrix A has two forms of matrix fraction representations:

$$M = G^{-1}N_1$$

and

$$M = Q^{-1}N_2$$

where N is unitary matrix and $G,\ N_1,\ N_2$ are banded matrices.

Definition 1. *(Mullhaupt, Riedel 2002) An $n \times n$ matrix is called lower banded with lower bandwidth (lwidth) d if $M_{ij} = 0$ for $i > j + d$. M is said to have strict lower bandwidth d if $M_{j+d,j} \neq 0$ for $1 \leq j \leq n - d$.*

Take $n = 5$ as example, a band 2 matrix has the following structure:

$$M = \begin{bmatrix} * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ 0 & * & * & * & 0 \\ 0 & 0 & * & * & * \end{bmatrix}$$

If lwidth$T \leq 0$, the matrix is called upper triangular The lower Hessenberg matrix has upper bandwidth 1, and upper Hessenberg matrix has lower bandwidth 1. Diagonal matrix has 0 upper and lower bandwidth.

Theorem 1. *(Mullhaupt, Riedel 2002). Suppose L is a lower triangular matrix with low-grade $\leq d-1$. For any $\varepsilon > 0$, there exists M and N which are also lower triangular matrix with bandwidth $\leq d$. s.t. $\| L - M^{-1}N \| < \varepsilon$.*

For a given covariance matrix Σ which is a positive definite matrix, let L be the Cholesky decomposition for Σ :

$$\Sigma = L * L^*$$

where L is a lower triangular matrix.

We can also write this as

$$\Sigma = R^* * R$$

where $R = L^T$. Therefore, R is a upper triangular matrix.

We can decompose R such that $R = M^{-1}N$. Therefore, the covariance matrix Σ can be represented by the band matrices M and N.

$$\Sigma = R'R = (M^{-1}N)^*(M^{-1}N)$$

There are various reasons that why we prefer band fraction representation to factor model. First, by using band fraction representation, the inverse of the covariance matrix Σ^{-1} stays band fraction.

$$\Sigma^{-1} = ((M^{-1}N)^*(M^{-1}N))^{-1} = (N^{-1}M)(N^{-1}M)^*$$

while the inverse of a factor model does not maintain factor structure anymore. By using Woodbury formula, we have the inverse of the factor model as following

$$\begin{aligned} (D + VV^*)^{-1} &= D^{-1} - D^{-1}V(I + V^*D^{-1}V)^{-1}V^*D^{-1} \\ &= D_1 - V_1V_1^* \end{aligned}$$

Second, covariance matrices are a convex subset of matrices. For any $0 < \alpha < 1$, the convex combination of two covariance matrices Σ_1 and Σ_2 is $\alpha\Sigma_1 + (1-\alpha)\Sigma_2$ which is still a covariance matrix. That is, for any non-zero vector x

$$x^T(\alpha\Sigma_1 + (1-\alpha)\Sigma_2)x = ax^T\Sigma_1 x + (1-\alpha)x^T\Sigma_2 x > 0$$

Moreover, covariance matrices are a global isometry of information distance on positive definite matrices:

$$I(\Sigma_1, \Sigma_2) = I(I, \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) = I(I, U\Lambda U^*) = I(I, \Lambda)$$

where U is an unitary matrix and Λ is a diagonal matrix.

The diagonal geodesic from I to Λ is $e^{t\Lambda}$ where $t \in [0, 1]$.

$$I(I, \Lambda(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})) = \frac{1}{2}\sum_{i=1}^{n}(\ln \lambda_i)^2$$

where λ_i is the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$.

Matrices of bandwidth $\leq d$ are a convex subset of matrices. For two pairs of band fraction representations (M_1, N_1) and (M_2, N_2) with bandwidth d, the convex combination $\alpha(M_1, N_1) + (1-\alpha)(M_2, N_2) = (aM_1 + (1-\alpha)M_1, \alpha N_1 + (1-\alpha)N_2)$ is still band fraction with bandwidth d and stays the same convex set. Band fraction representation preserves global isometry as well:

$$\begin{aligned} & I((M_1^{-1}N_1)^*(M_1^{-1}N_1), (M_2^{-1}N_2)^*(M_2^{-1}N_2)) \\ = \; & I((M_2^{-1}N_2)^{-*}(M_1^{-1}N_1)^*(M_1^{-1}N_1)(M_2^{-1}N_2)^{-1}, I) \\ = \; & I((M_0^{-1}N_0)^*(M_0^{-1}N_0), I) \end{aligned}$$

On the other hand, factor model does not have the global isometry property. Therefore, factor model does not stay in a nice manifold.

2.2 Semiseparable Approximations

Another widely used definition is semiseparable matrix, which is quite similar to the definition of low grade matrix.

Definition 2. *A matrix is a lower (upper) semiseparable matrix with semiseparability rank d if all of the submatrices that could be taken out the lower (upper) triangular block of the matrix have rank $\leq d$.*

It can be easily seen that a lower semiseparable matrix with semiseparablity rand d is a lower grade $d-1$ matrix.

To begin the semiseparable approximations for the symmetric positive definite matrices (SPD), let us first consider the Cholesky factorization procedure for any SPD matrix A:

For $k = 1, 2, ..., n$:

Cholesky factorize $R_{k,k}^T R_{k,k} := A_{k,k}$;

Compute $\quad R_{k,k+1:n} := R_{k,k}^{-T} A_{k,k+1:n}$;

Schur complement $A_{k+1:n,k+1:n} := A_{k+1:n,k+1:n} - R_{k,k+1:n}^T \times R_{k,k+1:n}$;

The output of the procedure above is the upper triangular matrix:

$$R = \begin{bmatrix} R_{1,1} & R_{1,2} & \cdots & R_{1,n} \\ & R_{2,2} & \cdots & R_{2,n} \\ & & \ddots & \vdots \\ & & & R_{n,n} \end{bmatrix} \text{ such that } A = R^T R$$

Ming Gu [37] modified the Cholesky factorization procedure and embedded the semiseparable matrix construction scheme into the procedure. His fast algorithm constructed an SPD semiseparable matrix which approximates S and preserves the product AZ for any given $Z \in R^{N \times d}$, $d \ll N$ and any given tolerance $\tau > 0$. It preserves the actions of A on certain vectors (directions):

$$S^T S = A + O(\sqrt{\|A\|_2 \, \tau}) \text{ and } S^T S Z = AZ$$

where

$$Z = \begin{bmatrix} Z_1 \\ \ddots \\ Z_n \end{bmatrix}$$

and S is an upper triangular semiseparable matrix of the following form

$$S = \begin{bmatrix} D_1 & S_{1,2} & \cdots & S_{1,n} \\ & D_2 & \cdots & S_{2,n} \\ & & \ddots & \vdots \\ & & & D_n \end{bmatrix}$$

with $D_k's$ upper triangular and $S_{k,t} = U_k W_{k+1} ... W_{t-1} V_t^T$.

Tengjie Jia [29] proposes an iterative method to construct the band fraction representation. In this paper, we will show that we can construct the band fraction (M, N) directly from Ming Gu's semiseparable matrix given the embedded special structure.

2.3 Block Banded Matrix M and N

The upper semiseparable matrix S constructed from Ming Gu's algorithm is a highly structured matrix. Each block element $S_{k,t}$ is a product of $U_k, W_{k+1}, ..., W_{t-1}$ and V_t. By decomposing the upper triangular semiseparable matrix S, we can get the block banded matrix M and M, which again can be represented by $U_k, W_{k+1}, ..., W_{t-1}$ and V_t.

2.3.1 4×4 block case

First, let us start from a simple case. Suppose S is a 4×4 block semiseparable matrix with the following form

$$S = \begin{bmatrix} D_1 & S_{1,2} & S_{1,3} & S_{1,4} \\ & D_2 & S_{2,1} & S_{2,2} \\ & & D_3 & S_{3,3} \\ & & & D_n \end{bmatrix}$$

substitute $S_{k,t} = U_k W_{k+1} ... W_{t-1} V_t^T$ into element of S, we have

$$S = \begin{bmatrix} D_1 & U_1 V_2^T & U_1 W_2 V_3^T & U_1 W_2 W_3 V_4^T \\ & D_2 & U_2 V_3^T & U_2 W_3 V_4^T \\ & & D_3 & U_3 V_4^T \\ & & & D_n \end{bmatrix}$$

Thus, by constructing M and N as following

$$M = \begin{bmatrix} I & -U_1 W_2 U_2^{-1} & & \\ & I & -U_2 W_3 V_3^{-1} & \\ & & I & 0 \\ & & & I \end{bmatrix}$$

and

$$N = \begin{bmatrix} D_1 & U_1 V_2^T - U_1 W_2 U_2^{-1} D_2 & & \\ & D_2 & U_2 V_3^T - U_2 W_3 U_3^{-1} D_3 & \\ & & D_3 & U_3 V_4^T \\ & & & D_4 \end{bmatrix}$$

we have

$$\begin{bmatrix} I & -U_1W_2U_2^{-1} & & \\ & I & -U_2W_3U_3^{-1} & \\ & & I & 0 \\ & & & I \end{bmatrix} \begin{bmatrix} D_1 & U_1V_2^T & U_1W_2V_3^T & U_1W_2W_3V_4^T \\ & D_2 & U_2V_3^T & U_2W_3V_4^T \\ & & D_3 & U_3V_4^T \\ & & & D_n \end{bmatrix}$$

$$= \begin{bmatrix} D_1 & U_1V_2^T - U_1W_2U_2^{-1}D_2 & & \\ & D_2 & U_2V_3^T - U_2W_3U_3^{-1}D_3 & \\ & & D_3 & U_3V_4^T \\ & & & D_4 \end{bmatrix}$$

That is

$$MS = N$$

where M and N are block banded matrices with block bandwidth 1. We can have the band fraction representation of S by taking the inversion of M

$$S = M^{-1}N$$

2.3.2 General $n \times n$ block case

Now let's consider the general case that S is $n \times n$ block upper semiseparable matrix which can be written as

$$S = \begin{bmatrix} D_1 & U_1V_2^T & \cdots & U_1W_2...W_{j-1}V_j^T & \cdots & U_1W_2W_3...W_{n-1}V_n^T \\ & D_2 & \ddots & \vdots & & \vdots \\ & & \ddots & U_iW_{i+1}...W_{i-1}V_j^T & & \vdots \\ & & & D_{i+1} & & \vdots \\ & & & & \ddots & U_{n-1}V_n^T \\ & & & & & D_n \end{bmatrix}$$

By the same decomposition procedure, we can construct M and N with the following form

$$M = \begin{bmatrix} I & -U_1 W_2 U_2^{-1} & & & & \\ & \ddots & \ddots & & & \\ & & & I & -U_{n-2} W_{n-1} U_{n-1}^{-1} & \\ & & & & I & 0 \\ & & & & & I \end{bmatrix}$$

$$N = \begin{bmatrix} D_1 & U_1 V_2^T - U_1 W_2 U_2^{-1} D_2 & & & & \\ & \ddots & \ddots & & & \\ & & & D_{n-2} & U_{n-2} V_{n-1}^T - U_{n-2} W_{n-1} U_{n-1}^{-1} D_{n-1} & \\ & & & & D_{n-1} & U_{n-1} V_n^T \\ & & & & & D_n \end{bmatrix}$$

That is, the block diagonal elements of M are identity matrices

$$M_{i,i} = I,\ i = 1, ..., n$$

and the superdiagonal element of M has the form

$$M_{i,i+1} = \begin{cases} -U_i W_{i+1} U_{i+1}^{-1}, & i = 1, ..., n-2 \\ 0, & i = n-1 \end{cases}$$

while the block diagonal elements of are diagonal matrices D_i

$$N_{i,i} = D_i,\ i = 1, ..., n$$

and the superdiagonal elements of N has the form

$$N_{i,i+1} = \begin{cases} U_i V_{i+1}^T - U_i W_{i+1} U_{i+1}^{-1} D_{i+1}, & i = 1, ..., n-2 \\ U_{n-1} V_n^T, & i = n-1 \end{cases}$$

Therefore, we have

$$\begin{bmatrix} I & -U_1W_2U_2^{-1} & & & & \\ & \ddots & \ddots & & & \\ & & & I & -U_{n-2}W_{n-1}U_{n-1}^{-1} & \\ & & & & I & 0 \\ & & & & & I \end{bmatrix}$$

$$\times \begin{bmatrix} D_1 & U_1V_2^T & \cdots & U_1W_2...W_{j-1}V_j^T & \cdots & U_1W_2W_3...W_{n-1}V_n^T \\ & D_2 & \ddots & \vdots & & \vdots \\ & & \ddots & U_iW_{i+1}...W_{i-1}V_j^T & & \vdots \\ & & & D_{i+1} & & \vdots \\ & & & & \ddots & U_{n-1}V_n^T \\ & & & & & D_n \end{bmatrix}$$

$$= \begin{bmatrix} D_1 & U_1V_2^T - U_1W_2U_2^{-1}D_2 & & & & \\ & \ddots & \ddots & & & \\ & & & D_{n-2} & U_{n-2}V_{n-1}^T - U_{n-2}W_{n-1}U_{n-1}^{-1}D_{n-1} & \\ & & & & D_{n-1} & U_{n-1}V_n^T \\ & & & & & D_n \end{bmatrix}$$

That is, the $n \times n$ block upper semiseparable matrix S can by represented as the band fraction of M and N :

$$S = M^{-1}N$$

3 Fisher Information Matrix of Band Fraction

For an observable random variable X and an unknown parameter θ, Fisher information measures the amount information X carries about θ. It is the variance of the score, which is the expected value of the observed information. In this section, we will calculate the fisher information matrix for the band fraction representation of the covariance matrix.

Let us consider the multivariate normal distribution $Y \sim N(\mu, \Sigma)$, the log likelihood is

$$\log(L) = \log P(Y \mid \mu, \Sigma) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}tr(\Sigma^{-1}C)$$

where $C = (x-\mu)(x-\mu)^T$ is the sample covariance matrix and x is the observed sample data.

We have the band fraction representation for covariance matrix as following

$$\Sigma = (M^{-1}N)^*(M^{-1}N) = N^*M^{-*}M^{-1}N$$

The inverse of the covariance can be represented by band fraction as well

$$\Sigma^{-1} = N^{-1}MM^*N^{-*}$$

where M and N are upper triangular matrices with band width d.

Given the band fraction representation, the log likelihood function can be written as

$$\log L = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(N^*M^{-*}M^{-1}N) - \frac{1}{2}tr(N^{-1}MM^*N^{-*}C)$$

By definition, the Fisher Information matrix is

$$I = -E[(\frac{\partial}{\partial\theta}\log(L))^2]$$

In this section, we will derive the Fisher Information matrix by calculating the second derivatives for $-\frac{1}{2}\log\det(N^*M^{-*}M^{-1}N)$ and $-\frac{1}{2}tr(N^{-1}MM^*N^{-*}C)$ respectively. In the derivation, following two properties of matrix derivative are used

$$\frac{\partial X^{-1}}{\partial X_{rs}} = -X^{-1}\frac{\partial X}{\partial X_{rs}}X^{-1}$$

$$\frac{\partial \log\det A}{\partial x} = Tr(A^{-1}\frac{\partial A}{\partial x})$$

3.1 First Derivative of Log Likelihood for Band Fraction

We will calculate the first derivative of the log likelihood function with respect to M_{rs} and N_{rs}. To simplify the calculation, let us derive the first derivatives for $-\frac{1}{2}\log\det(N^*M^{-*}M^{-1}N)$ first.

$$\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial M_{rs}} = Tr(N^{-1}MM^*N^{-*} \times \frac{\partial(N^*M^{-*}M^{-1}N)}{\partial M_{rs}})$$

In order to calculate $\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial M_{rs}}$, we need to solve $\frac{\partial(N^*M^{-*}M^{-1}N)}{\partial M_{rs}}$

$$\begin{aligned}
\frac{\partial(N^*M^{-*}M^{-1}N)}{\partial M_{rs}} &= N^*\frac{\partial M^{-*}}{\partial M_{rs}}M^{-1}N + N^*M^{-*}\frac{\partial M^{-1}}{\partial M_{rs}}N \\
&= -N^*M^{-*}\frac{\partial M^*}{\partial M_{rs}}M^{-*}M^{-1}N - N^*M^{-*}M^{-1}\frac{\partial M}{\partial M_{rs}}M^{-1}N \\
&= -N^*M^{-*}E_{sr}M^{-*}M^{-1}N - N^*M^{-*}M^{-1}E_{rs}M^{-1}N
\end{aligned}$$

where we use the notation

$$E_{rs} = \begin{cases} e_re_s^T, & if\ (r,s)\ in\ the\ band \\ 0, & otherwise \end{cases}$$

therefore

$$\begin{aligned}
&\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial M_{rs}} \\
&= Tr(N^{-1}MM^*N^{-*} \times (-N^*M^{-*}E_{sr}M^{-*}M^{-1}N - N^*M^{-*}M^{-1}E_{rs}M^{-1}N)) \\
&= Tr(-E_{sr}M^{-*} - M^{-1}E_{rs}) \\
&= Tr(-e_se_r^TM^{-*} - M^{-1}e_re_s^T) \\
&= -2(M^{-1})_{sr}
\end{aligned}$$

Similarly, we can get the first order derivative of $\log\det(N^*M^{-*}M^{-1}N)$ with respect to ∂N_{rs}

$$\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial N_{rs}} = Tr(N^{-1}MM^*N^{-*} \times \frac{\partial(N^*M^{-*}M^{-1}N)}{\partial N_{rs}})$$

since

$$\begin{aligned}
\frac{\partial(N^*M^{-*}M^{-1}N)}{\partial N_{rs}} &= \frac{\partial N^*}{\partial N_{rs}}M^{-*}M^{-1}N + N^*M^{-*}M^{-1}\frac{\partial N}{\partial N_{rs}} \\
&= E_{sr}M^{-*}M^{-1}N + N^*M^{-*}M^{-1}E_{rs}
\end{aligned}$$

plugging $\frac{\partial(N^*M^{-*}M^{-1}N)}{\partial N_{rs}}$ back to $\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial N_{rs}}$, we have

$$\begin{aligned}
&\frac{\partial \log\det(N^*M^{-*}M^{-1}N)}{\partial N_{rs}} \\
= \;& Tr(N^{-1}MM^*N^{-*} \times (E_{sr}M^{-*}M^{-1}N + N^*M^{-*}M^{-1}E_{rs})) \\
= \;& Tr(N^{-*}E_{sr} + N^{-1}E_{rs}) \\
= \;& Tr(N^{-*}e_se_r^T + N^{-1}e_re_s^T) \\
= \;& 2(N^{-1})_{sr}
\end{aligned}$$

Next, let us calculate the first derivative for $\frac{1}{2}tr(\Sigma^{-1}C) = \frac{1}{2}tr(N^{-1}MM^*N^{-*}C)$ with respect to M_{rs}

$$\begin{aligned}
\frac{\partial tr(N^{-1}MM^*N^{-*}C)}{\partial M_{rs}} &= tr(\frac{\partial(N^{-1}MM^*N^{-*}C)}{\partial M_{rs}}) \\
&= tr(N^{-1}E_{rs}M^*N^{-*}C + N^{-1}ME_{sr}N^{-*}C)
\end{aligned}$$

and the first derivative for $\frac{1}{2}tr(\Sigma^{-1}C) = \frac{1}{2}tr(N^{-1}MM^*N^{-*}C)$ with respect to N_{rs}

$$\begin{aligned}
\frac{\partial tr(N^{-1}MM^*N^{-*}C)}{\partial N_{rs}} &= tr(\frac{\partial(N^{-1}MM^*N^{-*}C)}{\partial N_{rs}}) \\
&= tr(-N^{-1}\frac{\partial N}{\partial N_{rs}}N^{-1}MM^*N^{-*}C - N^{-1}MM^*N^{-*}\frac{\partial N^*}{\partial N_{rs}}N^{-*}C) \\
&= tr(-N^{-1}E_{rs}N^{-1}MM^*N^{-*}C - N^{-1}MM^*N^{-*}E_{sr}N^{-*}C)
\end{aligned}$$

Thus, we have the first derivative for the log likelihood summarized below:

$$\frac{\partial \log(L)}{\partial M_{rs}} = (M^{-1})_{sr} - \frac{1}{2}tr(N^{-1}E_{rs}M^*N^{-*}C + N^{-1}ME_{sr}N^{-*}C)$$

$$\frac{\partial \log(L)}{\partial N_{rs}} = -(N^{-1})_{sr} - \frac{1}{2}tr(-N^{-1}E_{rs}N^{-1}MM^*N^{-*}C - N^{-1}MM^*N^{-*}E_{sr}N^{-*}C)$$

3.2 Element of Fisher information Matrix for Band Fraction

Next, we will derive the second derivatives of the log likelihood by calculating for each term in the log likelihood function. First, let us calculate the second derivative of $\log\det(N^*M^{-*}M^{-1}N)$

with respect to ∂M_{rs} and ∂M_{tk}:

$$\begin{aligned}
\frac{\partial^2 \log\det(N^* M^{-*} M^{-1} N)}{\partial M_{rs} \partial M_{tk}} &= \frac{\partial(2Tr(-M^{-1}E_{rs}))}{\partial M_{tk}} \\
&= -2Tr(M^{-1}\frac{\partial M}{\partial M_{tk}} M^{-1}E_{rs}) \\
&= -2Tr(M^{-1}E_{tk}M^{-1}E_{rs}) \\
&= -2Tr(M^{-1}e_t e_k^T M^{-1} e_r e_s^T) \\
&= -2(M^{-1})_{st}(M^{-1})_{kr}
\end{aligned}$$

We can see that the $\frac{\partial^2 \log\det(N^* M^{-*} M^{-1} N)}{\partial M_{rs} \partial M_{tk}}$ only depends on M, and is not a function of the other band matrix N.

Similarly, we can have

$$\frac{\partial^2 \log\det(N^* M^{-*} M^{-1} N)}{\partial M_{rs} \partial N_{tk}} = \frac{\partial(2Tr(-M^{-1}E_{rs}))}{\partial N_{tk}} = 0$$

and

$$\begin{aligned}
\frac{\partial^2 \log\det(N^* M^{-*} M^{-1} N)}{\partial N_{rs} \partial N_{tk}} &= \frac{\partial(2Tr(N^{-1}E_{rs}))}{\partial N_{tk}} \\
&= -2Tr(N^{-1}\frac{\partial N}{\partial N_{tk}} N^{-1}E_{rs}) \\
&= -2Tr(N^{-1}E_{tk}N^{-1}E_{rs}) \\
&= -2Tr(N^{-1}e_t e_k^T N^{-1} e_r e_s^T) \\
&= -2(N^{-1})_{st}(N^{-1})_{kr}
\end{aligned}$$

The second derivatives of $tr(N^{-1}MM^*N^{-*}C)$ is calculated as following:

$$\begin{aligned}
\frac{\partial^2 tr(N^{-1}MM^*N^{-*}C)}{\partial M_{rs} \partial M_{tk}} &= \frac{\partial(tr(N^{-1}E_{rs}M^*N^{-*}C + N^{-1}ME_{sr}N^{-*}C))}{\partial M_{tk}} \\
&= tr(N^{-1}E_{rs}E_{kt}N^{-*}C + N^{-1}E_{tk}E_{sr}N^{-*}C)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 tr(N^{-1}MM^*N^{-*}C)}{\partial M_{rs}\partial N_{tk}} &= \frac{\partial(tr(N^{-1}E_{rs}M^*N^{-*}C + N^{-1}ME_{sr}N^{-*}C))}{\partial N_{tk}} \\
&= tr(-N^{-1}\frac{\partial N}{\partial N_{tk}}N^{-1}E_{rs}M^*N^{-*}C - N^{-1}E_{rs}M^*N^{-*}\frac{\partial N^*}{\partial N_{tk}}N^{-*}C + ... \\
&\quad -N^{-1}\frac{\partial N}{\partial N_{tk}}N^{-1}ME_{sr}N^{-*}C - N^{-1}ME_{sr}N^{-*}\frac{\partial N^*}{\partial N_{tk}}N^{-*}C) \\
&= tr(-N^{-1}E_{tk}N^{-1}E_{rs}M^*N^{-*}C - N^{-1}E_{rs}M^*N^{-*}E_{kt}N^{-*}C + ... \\
&\quad -N^{-1}E_{tk}N^{-1}ME_{sr}N^{-*}C - N^{-1}ME_{sr}N^{-*}E_{kt}N^{-*}C)
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial^2 tr(N^{-1}MM^*N^{-*}C)}{\partial N_{rs}\partial N_{tk}} \\
= &\frac{\partial(tr(-N^{-1}E_{rs}N^{-1}MM^*N^{-*}C - N^{-1}MM^*N^{-*}E_{sr}N^{-*}C))}{\partial N_{tk}} \\
= &\ tr(N^{-1}\frac{\partial N}{\partial N_{tk}}N^{-1}E_{rs}N^{-1}MM^*N^{-*}C + N^{-1}E_{rs}N^{-1}\frac{\partial N}{\partial N_{tk}}N^{-1}MM^*N^{-*}C + ... \\
&\ +N^{-1}E_{rs}N^{-1}MM^*N^{-*}\frac{\partial N^*}{\partial N_{tk}}N^{-*}C + N^{-1}\frac{\partial N}{\partial N_{tk}}N^{-1}MM^*N^{-*}E_{sr}N^{-*}C + ... \\
&\ +N^{-1}MM^*N^{-*}\frac{\partial N^*}{\partial N_{tk}}N^{-*}E_{sr}N^{-*}C + N^{-1}MM^*N^{-*}E_{sr}N^{-*}\frac{\partial N^*}{\partial N_{tk}}N^{-*}C) \\
= &\ tr(N^{-1}E_{tk}N^{-1}E_{rs}N^{-1}MM^*N^{-*}C + N^{-1}E_{rs}N^{-1}E_{tk}N^{-1}MM^*N^{-*}C + ... \\
&\ +N^{-1}E_{rs}N^{-1}MM^*N^{-*}E_{kt}N^{-*}C + N^{-1}E_{tk}N^{-1}MM^*N^{-*}E_{sr}N^{-*}C + ... \\
&\ +N^{-1}MM^*N^{-*}E_{kt}N^{-*}E_{sr}N^{-*}C + N^{-1}MM^*N^{-*}E_{sr}N^{-*}E_{kt}N^{-*}C)
\end{aligned}$$

In order to get the fisher information matrix, the next step is to take the expectation of the second derivatives of the log likelihood function. By using the fact

$$E(C) = (M^{-1}N)^*(M^{-1}N) = N^*M^{-*}M^{-1}N$$

we have

$$E(\frac{\partial^2 \log\det(\Sigma)}{\partial M_{rs}\partial M_{tk}}) = E(\frac{\partial^2 \log\det(N^*M^{-*}M^{-1}N)}{\partial M_{rs}\partial M_{tk}}) = -2(M^{-1})_{st}(M^{-1})_{kr}$$

$$E(\frac{\partial^2 \log\det(\Sigma)}{\partial M_{rs}\partial N_{tk}}) = E(\frac{\partial^2 \log\det(N^*M^{-*}M^{-1}N)}{\partial M_{rs}\partial N_{tk}}) = 0$$

and

$$E(\frac{\partial^2 \log\det(\Sigma)}{\partial N_{rs}\partial N_{tk}}) = E(\frac{\partial^2 \log\det(N^*M^{-*}M^{-1}N)}{\partial N_{rs}\partial N_{tk}}) = -2(N^{-1})_{st}(N^{-1})_{kr}$$

Next, let us apply the same procedure to the second derivatives of $tr(\Sigma^{-1}C)$

$$
\begin{aligned}
& E(\frac{\partial^2 tr(\Sigma^{-1}C)}{\partial M_{rs}\partial M_{tk}}) \\
= \;& E(tr(N^{-1}E_{rs}E_{kt}N^{-*}C + N^{-1}E_{tk}E_{sr}N^{-*}C)) \\
= \;& tr(N^{-1}E_{rs}E_{kt}N^{-*}N^{*}M^{-*}M^{-1}N + N^{-1}E_{tk}E_{sr}N^{-*}N^{*}M^{-*}M^{-1}N) \\
= \;& tr(E_{rs}E_{kt}M^{-*}M^{-1} + E_{tk}E_{sr}M^{-*}M^{-1}) \\
= \;& 2tr(E_{kt}M^{-*}M^{-1}E_{rs}) \\
= \;& 2(M^{-*}M^{-1})_{tr}e_s^T e_k
\end{aligned}
$$

$$
\begin{aligned}
& E(\frac{\partial^2 tr(\Sigma^{-1}C)}{\partial M_{rs}\partial N_{tk}}) \\
= \;& E(tr(-N^{-1}E_{tk}N^{-1}E_{rs}M^{*}N^{-*}C - N^{-1}E_{rs}M^{*}N^{-*}E_{kt}N^{-*}C + ... \\
& -N^{-1}E_{tk}N^{-1}ME_{sr}N^{-*}C - N^{-1}ME_{sr}N^{-*}E_{kt}N^{-*}C)) \\
= \;& tr(-N^{-1}E_{tk}N^{-1}E_{rs}M^{*}N^{-*}N^{*}M^{-*}M^{-1}N - N^{-1}E_{rs}M^{*}N^{-*}E_{kt}N^{-*}N^{*}M^{-*}M^{-1}N + ... \\
& -N^{-1}E_{tk}N^{-1}ME_{sr}N^{-*}N^{*}M^{-*}M^{-1}N - N^{-1}ME_{sr}N^{-*}E_{kt}N^{-*}N^{*}M^{-*}M^{-1}N) \\
= \;& tr(-E_{tk}N^{-1}E_{rs}M^{-1} - E_{rs}M^{*}N^{-*}E_{kt}M^{-*}M^{-1} + ... \\
& -E_{tk}N^{-1}ME_{sr}M^{-*}M^{-1} - E_{sr}N^{-*}E_{kt}M^{-*}) \\
= \;& -(N^{-1})_{kr}(M^{-1})_{st} - (M^{*}N^{-*})_{sk}(M^{-*}M^{-1})_{tr} - (N^{-1}M)_{ks}(M^{-*}M^{-1})_{rt} - (N^{-*})_{rk}(M^{-*})_{ts} \\
= \;& -2(N^{-1})_{kr}(M^{-1})_{st} - 2(M^{*}N^{-*})_{sk}(M^{-*}M^{-1})_{tr}
\end{aligned}
$$

and

$$
\begin{aligned}
& E(\frac{\partial^2 tr(\Sigma^{-1}C)}{\partial N_{rs}\partial N_{tk}}) \\
= \; & E(tr(N^{-1}E_{tk}N^{-1}E_{rs}N^{-1}MM^*N^{-*}C + N^{-1}E_{rs}N^{-1}E_{tk}N^{-1}MM^*N^{-*}C + ... \\
& +N^{-1}E_{rs}N^{-1}MM^*N^{-*}E_{kt}N^{-*}C + N^{-1}E_{tk}N^{-1}MM^*N^{-*}E_{sr}N^{-*}C + ... \\
& +N^{-1}MM^*N^{-*}E_{kt}N^{-*}E_{sr}N^{-*}C + N^{-1}MM^*N^{-*}E_{sr}N^{-*}E_{kt}N^{-*}C)) \\
= \; & tr(N^{-1}E_{tk}N^{-1}E_{rs}N^{-1}MM^*N^{-*}N^*M^{-*}M^{-1}N + N^{-1}E_{rs}N^{-1}E_{tk}N^{-1}MM^*N^{-*}N^*M^{-*}M^{-1}N + ... \\
& +N^{-1}E_{rs}N^{-1}MM^*N^{-*}E_{kt}N^{-*}N^*M^{-*}M^{-1}N + N^{-1}E_{tk}N^{-1}MM^*N^{-*}E_{sr}N^{-*}N^*M^{-*}M^{-1}N + ... \\
& +N^{-1}MM^*N^{-*}E_{kt}N^{-*}E_{sr}N^{-*}N^*M^{-*}M^{-1}N + N^{-1}MM^*N^{-*}E_{sr}N^{-*}E_{kt}N^{-*}N^*M^{-*}M^{-1}N) \\
= \; & tr(E_{tk}N^{-1}E_{rs}N^{-1} + E_{rs}N^{-1}E_{tk}N^{-1} + ... \\
& +E_{rs}N^{-1}MM^*N^{-*}E_{kt}M^{-*}M^{-1} + E_{tk}N^{-1}MM^*N^{-*}E_{sr}M^{-*}M^{-1} + ... \\
& +N^{-*}E_{kt}N^{-*}E_{sr} + N^{-*}E_{sr}N^{-*}E_{kt}) \\
= \; & (N^{-1})_{kr}(N^{-1})_{st} + (N^{-1})_{st}(N^{-1})_{kr} + (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr} + ... \\
& +(N^{-1}MM^*N^{-*})_{ks}(M^{-*}M^{-1})_{rt} + (N^{-*})_{rk}(N^{-*})_{ts} + (N^{-*})_{ts}(N^{-*})_{rk} \\
= \; & 4(N^{-1})_{kr}(N^{-1})_{st} + 2(N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}
\end{aligned}
$$

In conclusion, the element of Fisher Information matrix for band fraction is given as following:

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial M_{tk}}) = (M^{-1})_{st}(M^{-1})_{kr} - (M^{-*}M^{-1})_{tr}e_s^T e_k$$

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (N^{-1})_{kr}(M^{-1})_{st} + (M^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}$$

$$
\begin{aligned}
E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) &= -(N^{-1})_{st}(N^{-1})_{kr} - 2(N^{-1})_{kr}(N^{-1})_{st} - (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr} \\
&= -(N^{-1})_{kr}(N^{-1})_{st} - (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}
\end{aligned}
$$

3.3 Bandwidth d = 1 case

In this section, we will give a simplified example of the band fraction matrices so that we can have a better understanding of the structure embedded. Let us consider the a special case when bandwidth $d = 1$. Since the diagonal elements of M are all ones, the parameter space is $\theta = (M_{12}, ..., M_{n-1,n}, N_{1,1}, ..., N_{n,n}, N_{12}, ..., N_{n-1,n})$, where n is the dimension of

the problem. Let us write down each element of Fisher Information matrix for band $d = 1$ first.

Since

$$E(\frac{\partial^2 \log L}{\partial M_{rs} \partial M_{tk}}) = (M^{-1})_{st}(M^{-1})_{kr} - (M^{-*}M^{-1})_{tr} e_s^T e_k$$

in the case of $d = 1$, we have $M_{rs} = M_{r,r+1}, M_{tk} = M_{t,t+1}$

$$\begin{aligned} E(\frac{\partial^2 \log L}{\partial M_{rs} \partial M_{tk}}) &= E(\frac{\partial^2 \log L}{\partial M_{rs} \partial M_{tk}}) \\ &= (M^{-1})_{r+1,t}(M^{-1})_{t+1,r} - (M^{-*}M^{-1})_{tr} e_{r+1}^T e_{t+1} \end{aligned}$$

Since M is upper triangular matrix, M^{-1} is upper triangular as well. In order to have $(M^{-1})_{r+1,t}(M^{-1})_{t+1,r} \neq 0$, we need

$$\begin{aligned} r+1 &\leq t \\ t+1 &\leq r \end{aligned}$$

that is

$$r + 1 \leq t \leq r - 1$$

which is impossible. Thus

$$(M^{-1})_{r+1,t}(M^{-1})_{t+1,r} = 0$$

$$E(\frac{\partial^2 \log L}{\partial M_{rs} \partial M_{tk}}) = -(M^{-*}M^{-1})_{tr} e_{r+1}^T e_{t+1} = \begin{cases} -(M^{-*}M^{-1})_{tr}, & for\ r = t \\ 0, & otherwise \end{cases}$$

For

$$E(\frac{\partial^2 \log L}{\partial N_{rs} \partial N_{tk}}) = -(N^{-1})_{kr}(N^{-1})_{st} - (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}$$

we discuss the following three cases: 1) $r = s,\ t = k$. 2) $r = s,\ k = t + 1$ and 3) $s = r + 1,\ k = t + 1$

Case 1, for $r = s,\ t = k$

$$E(\frac{\partial^2 \log L}{\partial N_{rs} \partial N_{tk}}) = -(N^{-1})_{tr}(N^{-1})_{rt} - (N^{-1}MM^*N^{-*})_{rt}(M^{-*}M^{-1})_{tr}$$

since

$$(N^{-1})_{tr}(N^{-1})_{rt} = \begin{cases} (N^{-1})_{rr}(N^{-1})_{rr}, & r = t \\ 0, & r \neq t \end{cases}$$

we have

$$E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) = \begin{cases} -(N^{-1})_{rr}(N^{-1})_{rr} - (N^{-1}MM^*N^{-*})_{rt}(M^{-*}M^{-1})_{tr}, & r = t \\ -(N^{-1}MM^*N^{-*})_{rt}(M^{-*}M^{-1})_{tr} \quad , & r \neq t \end{cases}$$

Case 2, for $r = s,\ k = t+1$

$$E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) = -(N^{-1})_{t+1,r}(N^{-1})_{rt} - (N^{-1}MM^*N^{-*})_{r,t+1}(M^{-*}M^{-1})_{tr}$$

and

$$(N^{-1})_{t+1,r}(N^{-1})_{rt} = 0$$

since $t+1 \leq r \leq t$ is not satisfied, we have

$$E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) = -(N^{-1}MM^*N^{-*})_{r,t+1}(M^{-*}M^{-1})_{tr}$$

Case 3, for $s = r+1,\ k = t+1$

$$\begin{aligned} E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) &= -(N^{-1})_{t+1,r}(N^{-1})_{r+1,t} - (N^{-1}MM^*N^{-*})_{s,t+1}(M^{-*}M^{-1})_{tr} \\ &= -(N^{-1}MM^*N^{-*})_{s,t+1}(M^{-*}M^{-1})_{tr} \end{aligned}$$

To conclude the three cases above, we have

$$E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) = \begin{cases} -(N^{-1})_{rr}(N^{-1})_{rr} - (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}, & r = t = s = k \\ -(N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}, & otherwise \end{cases}$$

Next, let us consider the cross term

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (N^{-1})_{kr}(M^{-1})_{st} + (M^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}$$

we have two cases 1) $s = r+1,\ t = k$ and 2) $s = r+1,\ k = t+1$

Case 1, for $s = r+1,\ t = k$

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (N^{-1})_{tr}(M^{-1})_{r+1,t} + (M^*N^{-*})_{r+1,t}(M^{-*}M^{-1})_{tr}$$

In order to have

$$(N^{-1})_{tr}(M^{-1})_{r+1,t} \neq 0$$

we need

$$\begin{aligned} t &\leq r \\ r+1 &\leq t \end{aligned}$$

which is impossible. Thus

$$(N^{-1})_{tr}(M^{-1})_{r+1,t} = 0$$

and we have

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (M^* N^{-*})_{r+1,t}(M^{-*}M^{-1})_{tr}$$

Case 2, for $s = r+1,\ k = t+1$

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (N^{-1})_{t+1,r}(M^{-1})_{r+1,t} + (M^* N^{-*})_{r+1,t+1}(M^{-*}M^{-1})_{tr}$$

similarly, we have

$$(N^{-1})_{t+1,r}(M^{-1})_{r+1,t} = 0$$

therefore,

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (M^* N^{-*})_{r+1,t+1}(M^{-*}M^{-1})_{tr}$$

To sum up, for the case when bandwidth $d = 1$, we have the elementwise Fisher information matrix

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial M_{tk}}) = -(M^{-*}M^{-1})_{tr}e_{r+1}^T e_{t+1} = \begin{cases} -(M^{-*}M^{-1})_{tr},\ for\ r = t \\ 0, \qquad otherwise \end{cases}$$

$$E(\frac{\partial^2 \log L}{\partial N_{rs}\partial N_{tk}}) = \begin{cases} -(N^{-1})_{rr}(N^{-1})_{rr} - (N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr},\ r = t = s = k \\ -(N^{-1}MM^*N^{-*})_{sk}(M^{-*}M^{-1})_{tr}, \qquad otherwise \end{cases}$$

$$E(\frac{\partial^2 \log L}{\partial M_{rs}\partial N_{tk}}) = (M^* N^{-*})_{sk}(M^{-*}M^{-1})_{tr}$$

3.4 Metric Tensor of Band Fraction

Given two normal distributions $N(\mu, C_1^2)$ and $N(\mu, C_2^2)$, we have the information distance [43]

$$I^2(C_1^2, C_2^2) = I^2(C_2^{-1}C_1^2C_2^{-1}, I) = \frac{1}{2}\sum_{k=1}^{n}(\log \lambda_k)^2$$

where λ_k are the eigenvalues of $C_1^{-1}C_2^2C_1^{-1}$. The first equivalence comes from the global isometry of Information distance.

For two band fraction representations $C_1 = (M_1^{-1}N_1)^*(M_1^{-1}N_1)$ and $C_2 = (M_2^{-1}N_2)^*(M_2^{-1}N_2)$, we can use the global isometry property as well

$$I((M_1^{-1}N_1)^*(M_1^{-1}N_1),(M_2^{-1}N_2)^*(M_2^{-1}N_2)) = I((M_2^{-1}N_2)^{-*}(M_1^{-1}N_1)^*(M_1^{-1}N_1)(M_2^{-1}N_2)^{-1},I)$$

Let $C_0^2 = (M_2^{-1}N_2)^{-*}(M_1^{-1}N_1)^*(M_1^{-1}N_1)(M_2^{-1}N_2)^{-1}$ and represent C_0^2 with band fraction matrices (M_0, N_0)

$$C_0^2 = (M_0^{-1}N_0)^*(M_0^{-1}N_0)$$

the above distance becomes

$$\begin{aligned} I((M_1^{-1}N_1)^*(M_1^{-1}N_1),(M_2^{-1}N_2)^*(M_2^{-1}N_2)) &= I((M_0^{-1}N_0)^*(M_0^{-1}N_0),I) \\ &= I(M_0^{-*}M_0^{-1},N_0^{-*}N_0) \end{aligned}$$

We have already calculated the information matrix for general band fraction representation. For the special case $C^2 = M^{-*}M^{-1}$ with $N = I$, we have the entries of fisher information matrix

$$g_{ij} = -E(\frac{\partial^2 \log L}{\partial M_{rs}\partial M_{tk}}) = -(M^{-1})_{st}(M^{-1})_{kr} + (M^{-*}M^{-1})_{tr}e_s^T e_k$$

with the parameterization $\theta = (M_{11}...,M_{n,n},M_{12},...,M_{n-1,n},...,M_{1,1+d},...,M_{n-d,n})'$, the metric tensor is

$$\begin{aligned} ds^2 &= \sum_{i,j=1}^{r} g_{i,j}(\theta)d\theta_i d\theta_j \\ &= \sum_{i=1}^{n}(-(M^{-1})_{ii}^2 + (M^{-*}M^{-1})_{ii})dM_{ii}^2 + \sum_{i=1}^{n-1}(M^{-*}M^{-1})_{ii}dM_{i,i+1}^2 + ... + \sum_{i=1}^{n-d}(M^{-*}M^{-1})_{ii}dM_{i,i+d}^2 \end{aligned}$$

3.5 Another Form with the Factorization $\Sigma = M^{-1}NN^*M^{-*}$

The covariance matrix can be also factored as

$$\Sigma = (M^{-1}N)(M^{-1}N)^* = N^*M^{-*}M^{-1}N$$

where M and N are lower triangular band matrix.

Similarly, we can have the first derivatives of the log likelihood function

$$\frac{\partial \log(L)}{\partial M_{rs}} = (M^{-1})_{sr} - \frac{1}{2}tr(E_{sr}N^{-*}N^{-1}MC + M^*N^{-*}N^{-1}E_{rs}C)$$

$$\frac{\partial \log(L)}{\partial N_{rs}} = -(N^{-1})_{sr} + \frac{1}{2}tr(M^*N^{-*}E_{sr}N^{-*}N^{-1}MC + M^*N^{-*}N^{-1}E_{rs}N^{-1}MC)$$

and the second derivatives of the log likelihood function

$$E(\frac{\partial^2 \log L}{\partial M_{rs} \partial M_{tk}}) = (M^{-1})_{st}(M^{-1})_{kr} - (N^{-*}N^{-1})_{rt}(M^{-1}NN^{*}M^{-*})_{ks}$$

$$E(\frac{\partial^2 \log L}{\partial M_{rs} \partial N_{tk}}) = (N^{-1})_{kr}(M^{-1})_{st} + (N^{-*}N^{-1})_{rt}(N^{-*}M^{-*})_{ks}$$

$$E(\frac{\partial^2 \log L}{\partial N_{rs} \partial N_{tk}}) = -(N^{-*}N^{-1})_{tr}e_s^T e_k - (N^{-1})_{st}(N^{-1})_{kr}$$

4 Metric Tensor and Upper Bound of Information Distance for Factor Model with respect to complete data

Factor model is widely used in covariance estimation. However, the information geometry of factor model is still an ongoing research topic. In this section, we will calculate the metric tensor of factor model with respect to the complete data. We will then give the information distance and upper bound of it with difference choices of D and V.

4.1 Metric Tensor for Factor Model with respect to complete data

Let us consider the factor model

$$y = Vx + \varepsilon$$

where y is the observed data, V is the factor loading, $x \sim N(0_{d\times 1}, I)$ is the factor, $\varepsilon \tilde{} N(0, I_{n\times n})$ is the error term.

Let us define the complete data

$$z \triangleq \begin{bmatrix} y \\ x \end{bmatrix} \tilde{} N(0, \begin{bmatrix} D+VV^* & V \\ V & I \end{bmatrix})$$

with the covariance matrix

$$\Sigma = \begin{bmatrix} D+VV^* & V \\ V & I \end{bmatrix}$$

the parameters are

$$\theta = \begin{bmatrix} diag(D) \\ vec(V) \end{bmatrix}$$

the log likelihood with respect to z is

$$L_z(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}z^T\Sigma^{-1}z$$

By using the property of schur complement, we have

$$\begin{aligned} \log\det(\Sigma) &= \log\det(\begin{bmatrix} D+VV^* & V \\ V & I \end{bmatrix}) \\ &= \log(\det(I)\det(D+VV^*-VI^{-1}V^*)) \\ &= \log\det(D) \end{aligned}$$

The inverse of Σ is given by

$$\Sigma^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}V \\ -V^*D^{-1} & I + V^*D^{-1}V \end{bmatrix}$$

and

$$\begin{aligned} z^T\Sigma^{-1}z &= \begin{bmatrix} y \\ x \end{bmatrix}^T \begin{bmatrix} D^{-1} & -D^{-1}V \\ -V^*D^{-1} & I + V^*D^{-1}V \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} \\ &= y^TD^{-1}y - 2x^TV^*D^{-1}y + x^T(I + V^*D^{-1}V)x \\ &= (y^T - x^TV^T)D^{-1}(y - Vx) + x^Tx \end{aligned}$$

we can rewrite the log likelihood with respect to z as

$$L_z(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log\det(D) - \frac{1}{2}((y^T - x^TV^T)D^{-1}(y - Vx) + x^Tx)$$

Let $D = diag(D_{11}, D_{22}, ..., D_{nn})$. The first derivative with respect to D_{ii} is

$$\frac{\partial L_z(\theta)}{\partial D_{ii}} = -\frac{1}{2}\frac{\partial \log\det(D)}{\partial D_{ii}} - \frac{1}{2}((y^T - x^TV^T)\frac{\partial D^{-1}}{\partial D_{ii}}(y - Vx) + x^Tx)$$

since

$$\frac{\partial \log\det(D)}{\partial D_{ii}} = D_{ii}^{-1}$$

and

$$\begin{aligned} \frac{\partial D^{-1}}{\partial D_{ii}} &= -D^{-1}\frac{\partial D}{\partial D_{ii}}D^{-1} \\ &= -D^{-1}e_ie_i^TD^{-1} \end{aligned}$$

thus the log likelihood function becomes

$$\begin{aligned} \frac{\partial L_z(\theta)}{\partial D_{ii}} &= -\frac{1}{2}D_{ii}^{-1} + \frac{1}{2}(y^T - x^TV^T)D^{-1}e_ie_i^TD^{-1}(y - Vx) \\ &= -\frac{1}{2D_{ii}} + \frac{(y^T - x^TV^T)_i^2}{2D_{ii}^2} \end{aligned}$$

The first derivative with respect to V_{r_1,r_2} is

$$\begin{aligned} \frac{\partial L_z(\theta)}{\partial V_{r_1,r_2}} &= -\frac{1}{2}\frac{\partial((y^T - x^TV^T)D^{-1}(y - Vx))}{\partial V_{r_1,r_2}} \\ &= x^Te_{r_2}e_{r_1}^TD^{-1}(y - Vx) \\ &= \frac{(y - Vx)_{r_1}x_{r_2}}{D_{r_1}} \end{aligned}$$

The second derivative of log likelihood is

$$\frac{\partial L_z(\theta)}{\partial D_{ii}\partial D_{jj}} = \begin{cases} \frac{1}{2D_{ii}^2} - \frac{(y^T - x^T V^T)_i^2}{D_{ii}^3} & i = j \\ 0 & i \neq j \end{cases}$$

$$\frac{\partial L_z(\theta)}{\partial D_{ii}\partial V_{r_1,r_2}} = \frac{(y^T - x^T V^T)_i}{D_{ii}^2}(-x^T e_{r_2} e_{r_1}^T e_i)$$

$$\begin{aligned} \frac{\partial L_z(\theta)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}} &= -x^T e_{r_2} e_{r_1}^T D^{-1} e_{r_3} e_{r_4} x \\ &= \begin{cases} -\frac{x_{r_2} x_{r_4}}{D_{r1}} & r_1 = r_3 \\ 0 & otherwise \end{cases} \end{aligned}$$

since

$$E(\frac{(y^T - x^T V^T)_i^2}{D_{ii}^3}) = -\frac{1}{2D_{ii}^2}$$

$$E(\frac{\partial L_z(\theta)}{\partial D_{ii}\partial V_{r_1,r_2}}) = 0$$

$$E(\frac{\partial L_z(\theta)}{\partial V_{r_1,r_2}\partial V_{r_3,r_4}}) = \begin{cases} -\frac{1}{D_{r1}} & r_1 = r_2 = r_3 = r_4 \\ 0 & otherwise \end{cases}$$

the Fisher information matrix with respect to z is in the form of

$$I_z = \begin{bmatrix} diag(\frac{1}{2D_{ii}^2}) & 0 \\ 0 & diag(\frac{1}{D_{r_1}}) \end{bmatrix}$$

the metric tensor is

$$g_{ij}(\theta) = \begin{bmatrix} vec(D) \\ vec(V) \end{bmatrix}^T I_z(\theta) \begin{bmatrix} vec(D) \\ vec(V) \end{bmatrix}$$

where

$$\begin{bmatrix} vec(D) \\ vec(V) \end{bmatrix} = \begin{bmatrix} D_{11} \\ D_{22} \\ \vdots \\ D_{nn} \\ V_{11} \\ V_{12} \\ \vdots \\ V_{1d} \\ V_{21} \\ V_{22} \\ \vdots \\ V_{2d} \\ \vdots \\ V_{n1} \\ V_{n2} \\ \vdots \\ V_{nd} \end{bmatrix}$$

Rao's distance metric is

$$\begin{aligned} ds^2 &= \partial D_{11}^2 \frac{1}{2D_{11}^2} + \partial D_{22}^2 \frac{1}{2D_{22}^2} + ... + \partial D_{11}^2 \frac{1}{2D_1^2} + ... \\ &\quad + \partial V_{11}^2 \frac{1}{D_1} + \partial V_{12}^2 \frac{1}{D_1} + ... + \partial V_{1d}^2 \frac{1}{D_1} + ... \\ &\quad + \partial V_{21}^2 \frac{1}{D_2} + \partial V_{22}^2 \frac{1}{D_2} + ... + \partial V_{2d}^2 \frac{1}{D_2} + ... \\ &\quad + ... \\ &\quad + \partial V_{n1}^2 \frac{1}{D_n} + \partial V_{n2}^2 \frac{1}{D_n} + ... + \partial V_{nd}^2 \frac{1}{D_n} \\ &= \frac{1}{2} tr(dDD^{-2}dD) + tr(D^{-1}dVdV^*) \end{aligned}$$

4.2 Upper Bound of the information distance with respect to z

In metric geometry, a geodesic is a curve that is everywhere locally a distance minimizer.

When considering the information distance between two factor models (D_1, V_1) and

(D_2, V_2), we distinguish three case: 1) $D_1 \neq D_2,\ V_1 = V_2$. 2) $D_1 = D_2,\ V_1 \neq V_2$. 3) $D_1 \neq D_2,\ V_1 \neq V_2$.

4.2.1 Factor models differ only in diagonal part D_1 and D_2

In this case, the metric tensor reduces to

$$\begin{aligned} ds^2 &= \frac{1}{2} tr(dDD^{-2}dD) \\ &= \frac{1}{2}\sum_{i=1}^{n} \frac{\partial D_{ii}^2}{D_{ii}^2} \end{aligned}$$

thus, the information distance between (D_1, V) and (D_2, V) is

$$I_z\left(D_1, V\right), \left(D_2, V\right) = \frac{1}{2} tr(\ln(D_1^{-1}D_2)^2)$$

4.2.2 Factor models differ only in the factor loading V_1 and V_2

the metric tensor reduces to

$$\begin{aligned} ds^2 &= tr(D^{-1}dVdV^*) \\ &= D^{-1}tr(dVdV^*) \\ &= \partial V_{11}^2 \frac{1}{D_{11}} + \partial V_{12}^2 \frac{1}{D_{11}} + ... + \partial V_{1d}^2 \frac{1}{D_{11}} + ... \\ &\quad + \partial V_{21}^2 \frac{1}{D_{22}} + \partial V_{22}^2 \frac{1}{D_{22}} + ... + \partial V_{2d}^2 \frac{1}{D_{22}} + ... \\ &\quad + ... \\ &\quad + \partial V_{n1}^2 \frac{1}{D_{nn}} + \partial V_{n2}^2 \frac{1}{D_{nn}} + ... + \partial V_{nd}^2 \frac{1}{D_{nn}} \end{aligned}$$

the information distance between (D, V_1) and (D, V_2) is

$$I_z\left(\left(D, V_1\right), \left(D, V_2\right)\right) = tr(D(V_1 - V_2)(V_1 - V_2)^*)$$

4.2.3 Factor models differ in both factor loadings V and diagonal part D

In this case, we will give a upper bound for the information distance. Let us construct two curves: one is the geodesic from (D_1, V_1) to (D_2, V_1) and the other one is from (D_2, V_1) to (D_2, V_2). The geodesic from (D_1, V_1) to (D_2, V_2) is bounded by

$$\begin{aligned} I_z((D_1, V_1), (D_2, V_2)) &\leq I_z((D_1, V_1), (D_2, V_1)) + I_z((D_2, V_1), (D_2, V_2)) \\ &= \frac{1}{2} tr(\ln(D_1^{-1}D_2)^2) + tr(D_2(V_1 - V_2)(V_1 - V_2)^*) \end{aligned}$$

by symmetry„ we can have another bound

$$\begin{aligned} I_z((D_1,V_1),(D_2,V_2)) &\leq I_z((D_2,V_2),(D_1,V_2)) + I_z((D_1,V_1),(D_1,V_2)) \\ &= \frac{1}{2}tr(\ln(D_1^{-1}D_2)^2) + tr(D_1(V_1-V_2)(V_1-V_2)^*) \end{aligned}$$

Thus, the upper bound for the information distance with respect to z is given by

$$\begin{aligned} I_z((D_1,V_1),(D_2,V_2)) &\leq \frac{1}{2}tr(\ln(D_1^{-1}D_2)^2) + ... \\ &\quad \min(tr(D_2(V_1-V_2)(V_1-V_2)^*), tr(D_1(V_1-V_2)(V_1-V_2)^*)) \end{aligned}$$

5 Optimization Method

In this section, we will first review the maximum likelihood estimator and Newton step. We then introduce the fisher scoring algorithm which is used in the improved band fraction estimation algorithm. The numerical results are provided, comparing to the factor model and band fraction without fisher scoring.

5.1 Maximum Likelihood Estimators

Maximum-likelihood estimation (MLE) is a way of estimating the parameters of a statistical model [105]. Given a statistical model and applied to a data set, maximum-likelihood estimation gives estimates for models's parameters. In general, it selects the set of parameter values that maximizes the likelihood function.

Suppose an i.i.d. sample $X = (X_1, X_2, ..., X_n)$ from a distribution with pdf $f(x; \theta)$, the likelihood function is

$$L(\theta|X) = \prod_{i=1}^{n} f(x_i; \theta)$$

The log likelihood function is

$$\log L(\theta|X) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

The maximum likelihood estimator is given by

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta|X)$$

As the sample size n goes to infinity, the maximum-likelihood estimator has the following properties

1. Consistency: the sequence of maximum-likelihood estimators converges to the true value in probability

$$\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

2. Asymptotic normality: as n increases, the distribution of the maximum-likelihood estimator tends to the Gaussian distribution, with mean θ and covariance matrix as the inverse of the Fisher information matrix

$$\sqrt{n}(\theta'_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta)^{-1})$$

3. Functional invariance: if $\widehat{\theta}$ is the maximum-likelihood estimator of θ, then maximum-likelihood estimator of $g(\theta)$ is $g(\widehat{\theta})$, for any function $g(\theta)$.

4. Efficiency: maximum-likelihood estimator achieves the Cramer-Rao lower bound as n goes to infinity.

5.2 Gradient descent method

Gradient descent is a first-order optimization algorithm. To find a local minimum of a function $f(x)$ using gradient descent, one takes steps proportional to the negative of the gradient $\nabla f(x)$ (or of the approximate gradient) of the function at the current point.

The procedure of Gradient descent method is as follows:

given a starting point $x \in dom\ f$

repeat

1) $\Delta x := -\nabla f(x)$

2) Line search. Choose step size t via exact or backtracking line search.

3) Update. $x := x + t\Delta x$.

Until stopping criterion is satisfied.

5.3 Newton's Method

Given a function $f(x)$, for $x \in dom\ f$, the vector

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

is called the Newton step [49]. The positive definiteness of $\nabla^2 f(x)$ shows that

$$\nabla^2 f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$

unless $\nabla f(x) = 0$ (then x is the optimal point). The Newton step is also the steepest descent direction at x, with respect to the quadratic norm defined by the Hessian function $\nabla^2 f(x)$

$$\|u\|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}$$

Thus, Newton's method can be considered as one of the steepest descent methods.

The procedure of Newton's method is:

Given a starting point $x \in dom\ f$, tolerance $\epsilon > 0$, repeat:

1. Compute the Newton step and decrement

$$\Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x); \quad \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

2. Stopping criterion. quit if $\lambda^2/2 \leq \epsilon$
3. Line search. Choose step size t by backtracking line search.
4. Update. $x := x + t\Delta x_{nt}$.

Newton step is independent of linear affine changes of coordinates and insensitive to the condition number of the sublevel sets of the objective. In general, the convergence of Newton's method is rapid.

5.4 Fisher Scoring Algorithm

In statistics, Fisher's scoring algorithm is a form of Newton's method and is used to solve the maximum likelihood problem numerically.

Let $Y_1, ..., Y_n$ be i.i.d. random variables with twice differentiable p.d.f. $f(y; \theta)$. To the calculate the maximum likelihood estimator θ_*, suppose we have a starting point θ_0, and the score function $V(\theta)$. The Talor expansion of $V(\theta)$ about θ_0 is

$$V(\theta) \approx V(\theta_0) - J(\theta_0)(\theta - \theta_0)$$

where

$$J(\theta_0) = -\sum_{i=1}^{n} \nabla \nabla^T |_{\theta=\theta_0} \log f(Y_i; \theta)$$

is the observed information matrix about θ_0. Using $V(\theta_*) = 0$

$$\theta_* = \theta_0 + J^{-1}(\theta_0) V(\theta_0)$$

therefore, we have the update formula of the algorithm

$$\theta_{m+1} = \theta_m + J^{-1}(\theta_m) V(\theta_m)$$

and it is shown that under certain regularity conditions, the algorithm converges $\theta_m \to \theta_*$.

In practice, we can use the Fisher information matrix instead.

$$I(\theta) = E(J(\theta))$$

which gives us the Fisher scoring algorithm

$$\theta_{m+1} = \theta_m + I^{-1}(\theta_m) V(\theta_m)$$

5.5 Conjugate Gradient Method

Conjugate gradient methods are widely used for unconstrained optimization [49]

$$\min\{f(x) : x \in R^n\}$$

where $f : R^n \to R$ is a continuously differentiable function, bounded from below. A nonlinear conjugate gradient method generates a sequence x_k, $k \geq 1$, starting from an initial guess $x_0 \in R^n$, using the recurrence

$$x_{k+1} = x_k + \alpha_k d_k$$

where the positive step size α_k is obtained by a line search, and the directions d_k are generated by the rule

$$\begin{aligned} d_{k+1} &= -g_{k+1} + \beta_k d_k \\ d_0 &= -g_0 \end{aligned}$$

here β_k is the conjugate gradient update parameter and $g_k = \nabla f(x_k)^T$. Let $y_k = g_{k+1} - g_k$. Following are several different choices of β_k

$$\begin{aligned} \beta_k^{HS} &= \frac{g_{k+1}^T y_k}{d_k^T y_k} \quad \text{(Hestenes and Stiefel)} \\ \beta_k^{FR} &= \frac{\|g_{k+1}\|^2}{\|g_k\|^2} \quad \text{(Fletcher and Reeves)} \\ \beta_k^{DY} &= \frac{\|g_{k+1}\|^2}{d_k^T y_k} \quad \text{(Dai and Yuan)} \end{aligned}$$

5.6 Line Search Method

Exact line search is one of the line search methods. The step size t is chosen to minimize $f(x)$ along the line $\{x + t\Delta x \mid t \geq 0\}$:

$$t = \arg\min_{s \geq 0} f(x + s\Delta x)$$

It is used when the cost of the minimization is low compared to the cost of computing the line search direction.

In practice, we usually use inexact line search. That is, the step size is chosen such that $f(x)$ is approximately minimized along the line $\{x + t\Delta x \mid t \geq 0\}$. Backtesting line search is one inexact line search method appears to be simple and effective:

Given a descent direction Δx for f at $x \in dom\ f,\ \alpha \in (0, 0.5),\ \beta \in (0, 1)$.

$t := 1$.

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$,

$t := \beta t$.

Backtracking line search stops when the step size t satisfies

$$t = 1,\ or\ t \in (\beta t_0, t_0]$$

The termination conditions for the CG line search are often based on some version of the Wolfe conditions. The standard Wolfe conditions are

$$\begin{aligned} f(x_k + \alpha_k d_k) - f(x_k) &\leq \delta \alpha_k g_k^T d_k \\ g_{k+1}^T d_k &\geq \sigma g_k^T d_k \end{aligned}$$

where d_k is a descent direction and $0 < \delta \leq \sigma < 1$. The strong Wolfe conditions are

$$\begin{aligned} f(x_k + \alpha_k d_k) - f(x_k) &\leq \delta \alpha_k g_k^T d_k \\ |g_{k+1}^T d_k| &< -\sigma g_k^T d_k \end{aligned}$$

Ideally, we would like to terminate the line search in a CG algorithm when the standard Wolfe conditions are satisfied. For some CG algorithms, however, stronger versions of the Wolfe conditions are needed to ensure convergence and to enhance stability.

5.7 First Derivatives of Information Distance of Band Fraction

The procedure of band fraction estimation with the MLE is:

Given the sample covariance from the observed data x and tolerance τ:

1. Calculate the semiseparable approximation S.

2. Decompose S into the band matrix (M_0, N_0), which could be used as the initial point in Fisher scoring algorithm.

3. Compute the first derivative and Fisher information matrix for any given (M, N) and update the parameters based on the updating formula.

4. Find the optimal band fraction representation (M, N) which gives the maximum log likelihood estimator under the given tolerance τ.

However, MLE does not necessarily guarantee the minimum information distance estimator. In most applications, we would prefer the minimum information distance estimator.

Let us consider the information distance between two covariance matrices Σ_1 and Σ_2

$$I^2 = \frac{1}{2}\sum_{k=1}^{n}(\log\lambda_k)^2$$

where λ_k are the eigenvalues of $\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}$. In the covariance estimation Σ_1 would be the sample covariance matrix and $\Sigma_2 = (M^{-1}N)(M^{-1}N)^*$ is the band fraction estimation to Σ_1. We can rewrite this formula as

$$\begin{aligned} I^2 &= \frac{1}{2}trace(\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^*) \\ &= \frac{1}{2}trace(\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})) \\ &= \frac{1}{2}trace(\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \end{aligned}$$

where we use the fact that $\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}) = \log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})^*$ since $\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})$ is symmetric matrix.

Let us calculate the first derivative of I^2

$$\begin{aligned} \frac{\partial I^2}{\partial M_{ij}} &= \frac{1}{2}tr(2(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})^{-1}\frac{\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}}{\partial M_{ij}} \times ... \\ &\quad \times\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\ &= tr((\Sigma_1^{1/2}(M^{-1}N)^{-*}(M^{-1}N)^{-1}\Sigma_1^{1/2})(-\Sigma_1^{-1/2}M^{-1}E_{ij}M^{-1}NN^*M^{-*}\Sigma_1^{-1/2} - ... \\ &\quad -\Sigma_1^{-1/2}M^{-1}NN^*M^{-*}E_{ji}M^{-*}\Sigma_1^{-1/2})\times\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\ &= tr((\Sigma_1^{1/2}(M^*N^{-*}N^{-1})(-E_{ij}M^{-1}NN^*M^{-*}\Sigma_1^{-1/2} - NN^*M^{-*}E_{ji}M^{-*}\Sigma_1^{-1/2})\times ... \\ &\quad \times\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\ &= tr((-\Sigma_1^{1/2}M^*N^{-*}N^{-1}E_{ij}M^{-1}NN^*M^{-*}\Sigma_1^{-1/2} - \Sigma_1^{1/2}E_{ji}M^{-*}\Sigma_1^{-1/2})\times ... \\ &\quad \times\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \end{aligned}$$

Let

$$M1 = M^{-1}NN^*M^{-*}\Sigma_1^{-1/2}\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}))\Sigma_1^{1/2}M^*N^{-*}N^{-1}$$

$$M2 = M^{-*}\Sigma_1^{-1/2}\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}))\Sigma_1^{1/2}$$

then

$$\frac{\partial I^2}{\partial M_{ij}} = -(M1)_{ji} - (M2)_{ij}$$

Similarly, let us calculate $\frac{\partial I^2}{\partial N_{ij}}$

$$\begin{aligned}
\frac{\partial I^2}{\partial N_{ij}} &= \frac{1}{2} tr(2(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})^{-1} \frac{\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}}{\partial N_{ij}} \times ... \\
&\quad \times \log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\
&= tr(\Sigma_1^{1/2}(M^{-1}N)^{-*}(M^{-1}N)^{-1}\Sigma_1^{1/2}(\Sigma_1^{-1/2}M^{-1}E_{ij}N^*M^{-*}\Sigma_1^{-1/2} + ... \\
&\quad + \Sigma_1^{-1/2}M^{-1}NE_{ji}M^{-*}\Sigma_1^{-1/2}) \times \log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\
&= tr(\Sigma_1^{1/2}M^*N^{-*}N^{-1}M(M^{-1}E_{ij}N^*M^{-*}\Sigma_1^{-1/2} + M^{-1}NE_{ji}M^{-*}\Sigma_1^{-1/2}) \times ... \\
&\quad \times \log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})) \\
&= tr((\Sigma_1^{1/2}M^*N^{-*}N^{-1}E_{ij}N^*M^{-*}\Sigma_1^{-1/2} + \Sigma_1^{1/2}M^*N^{-*}E_{ji}M^{-*}\Sigma_1^{-1/2}) \times ... \\
&\quad \times \log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}))
\end{aligned}$$

let

$$N1 = N^*M^{-*}\Sigma_1^{-1/2}\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2}))\Sigma_1^{1/2}M^*N^{-*}N^{-1}$$

$$N2 = M^{-*}\Sigma_1^{-1/2}\log(\Sigma_1^{-1/2}(M^{-1}N)(M^{-1}N)^*\Sigma_1^{-1/2})\Sigma_1^{1/2}M^*N^{-*}$$

Then

$$\frac{\partial I^2}{\partial N_{ij}} = (N1)_{ji} + (N2)_{ij}$$

We can use the conjugate gradient method to get the local minimum information distance estimator for the covariance matrix with good initial starting point.

5.7.1 Log Band Fraction Structure for Covariance Matrix

Instead of assuming the band fraction structure of the covariance matrix, we propose to use the band fraction structure to the logarithm of the covariance matrix Σ. That is,

$$\log \Sigma = (M^{-1}N)(M^{-1}N)^*$$

Since the sample covariance matrix is semidefinite positive, we can not take logarithm to the sample covariance directly. One way to solve with this is to use shrinkage. We can also do a higher order band fraction approximation to the sample covariance matrix to make it positive definite. We also need to scale the covariance matrix to a make the minimum eigenvalue of the sample covariance greater than 1.

5.8 Empirical Results

In this section, we compare the factor model, band fraction approximation and log band fraction approximation in terms of information distance and Hellinger distance. We compare the results based on both randomly generated data and empirical price data from S&P 500.

5.8.1 Randomly Generated Covariance Matrix

First, let us assume that we have a randomly generated covariance matrix Σ where Σ is a $n \times n$ matrix. We generate samples from the distribution $N(0, \Sigma)$. We apply factor model, band fraction, and log band fraction approximation to the sample covariance $\widehat{\Sigma}$, respectively. There are many measures we can use to assess the quality of the approximation methods. Here we calculate the log likelihood, Hellinger distance and information distance between the approximated covariance matrix and the true covariance matrix. We use $n = 100$ as the dimension size and choose $d = 30$ as the number of factors and $d + 1$ as the bandwidth of M and N. We randomly test for $N = 1000$ cases. Following figure shows the information distance for factor model, band fraction and log band fraction for the N randomly generated cases.

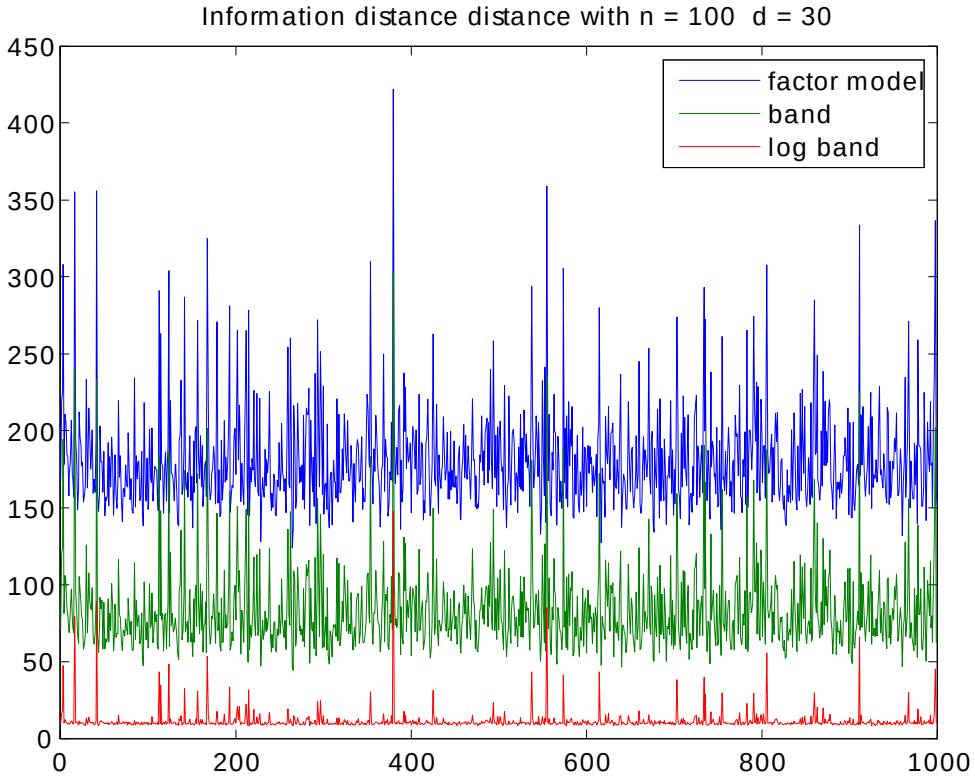

Figure 1: information distance for factor model, band fraction approximation, and log band fraction approximation for randomly generated covariance matrix with N=1000

We can see that log band fraction approximation gives the minimum information distance among the three methods, while factor model gives the worst approximation in terms of information distance. In order to see how much it improves, we can compute the relative difference between information distances

$$relative\ diff = \frac{dist1 - dist2}{dist1}$$

Following figure shows the relative difference between the information distances from band fraction and log band fraction approximation. We can see that the distance is improved by adopting the log band fraction structure for the covariance matrix.

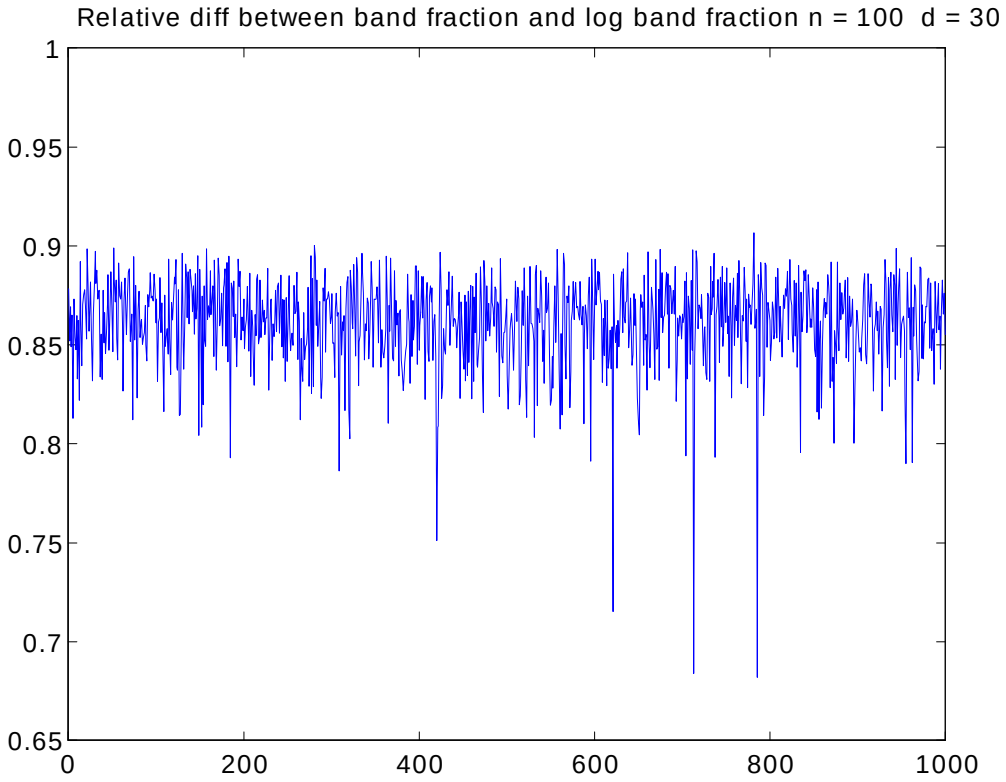

Figure 2: Relative difference of information distances of band fraction and log band fraction approximation ((distband-distlog band)/(distband)) for randomly generated covariance matrix

We can also compare the Hellinger distance for those methods. In the following figure, we can see that while factor model and band fraction give approximation with Hellinger distance close to 1, log band fraction is giving smaller Hellinger distance between the approximated matrix and the true covariance matrix.

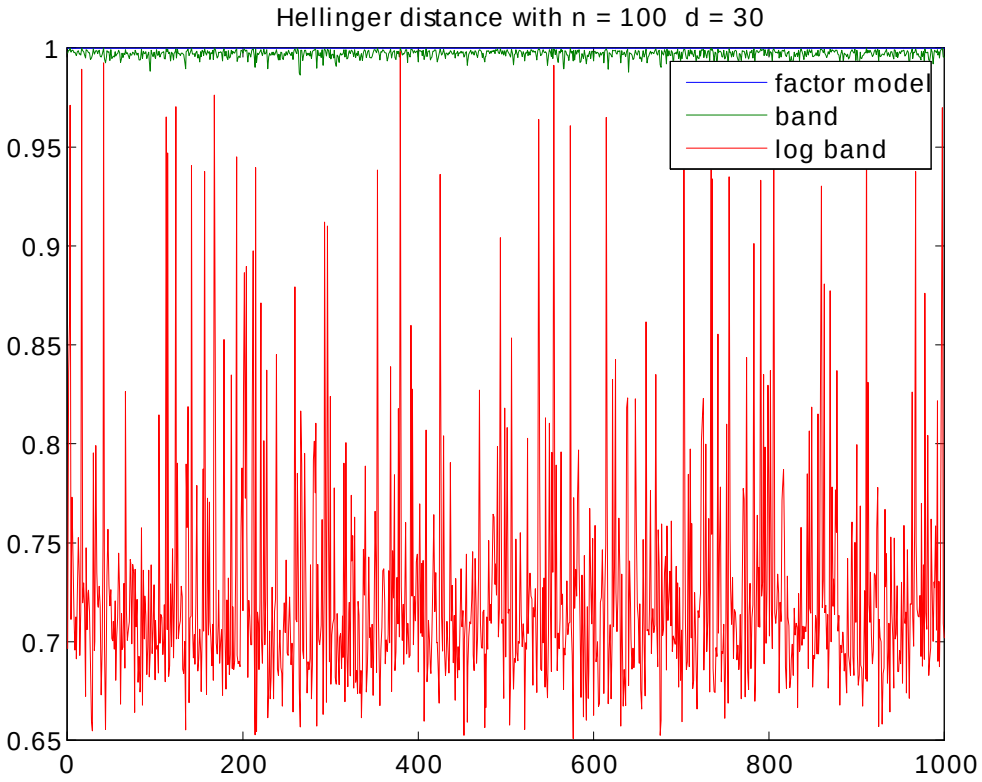

Figure 3: Hellinger distance for factor model, band fraction approximation, and log band fraction approximation for randomly generated covariance matrix with N=1000

5.8.2 Random Covariance Matrix with Factor Structure

What if the true covariance matrix actually has a factor structure ? Let us assume that the covariance matrix $\Sigma = D + VV'$ where D is $n \times n$ and V is $n \times p$. Similar as the previous section, we use factor model, band fraction and log band fraction to get the approximated covariance matrix with $d < p$. Following figure shows that log band fraction still gives better approximation than the factor model and band fraction even when the true covariance matrix is assumed to be factor structure.

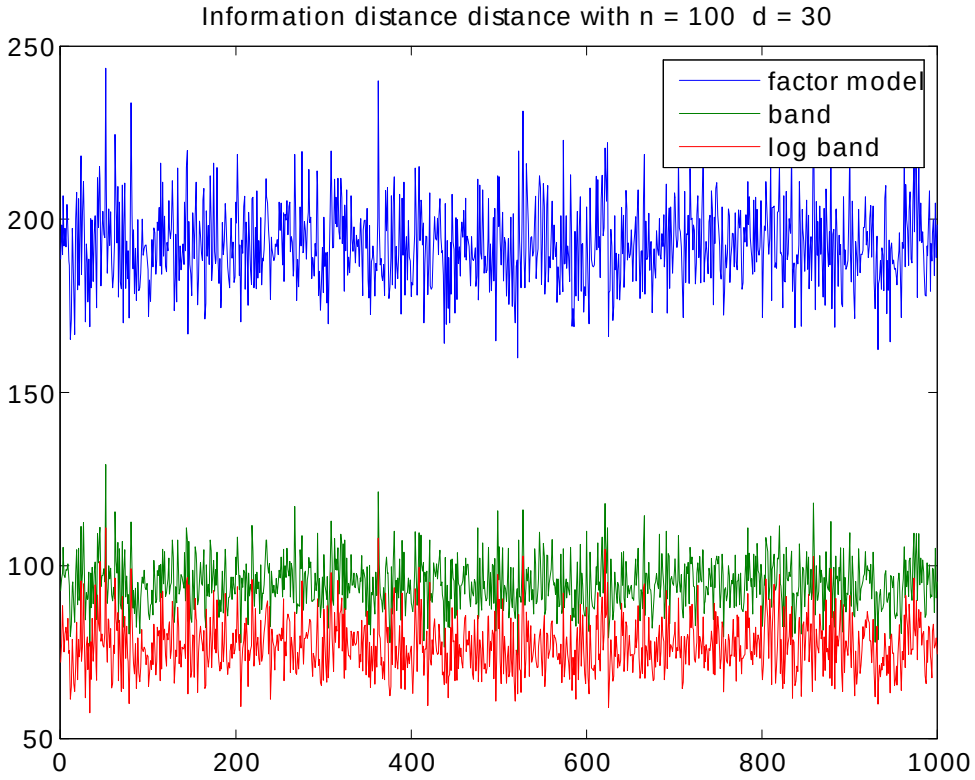

Figure 4: Information distance for factor model, band fraction approximation, and log band fraction approximation for covariance matrix with factor structure with N=1000

We can see that the log band fraction is still outperforming the band fraction approximation in the this case.

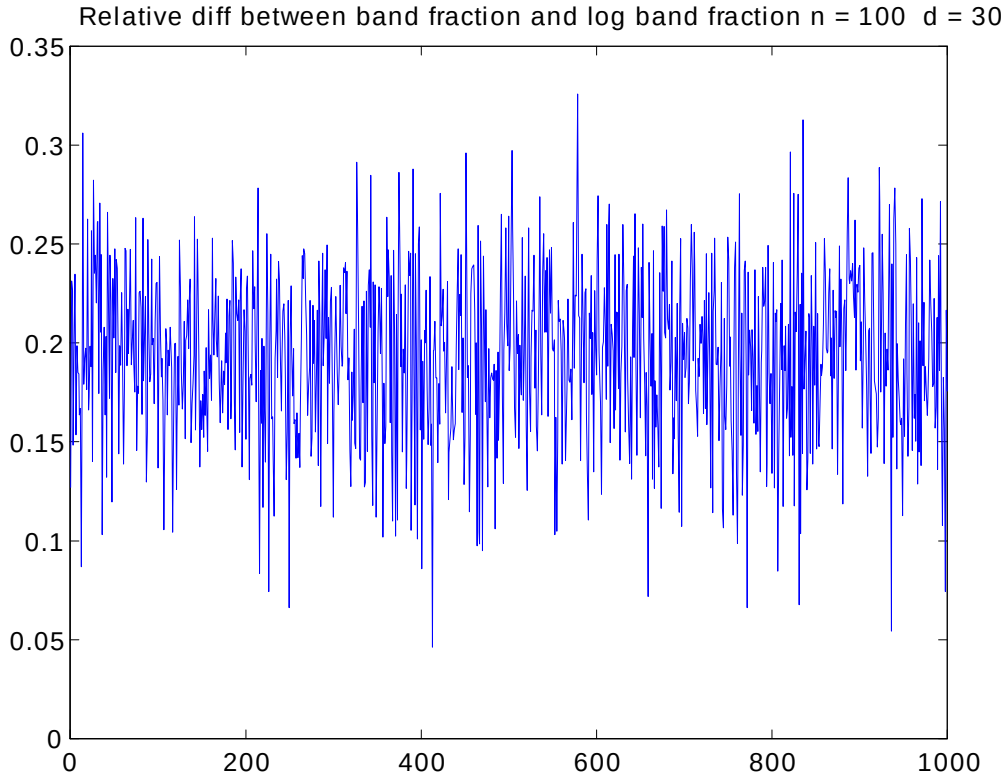

Figure 5: Relative difference of information distances of band fraction and log band fraction approximation ((distband-distlog band)/(distband)) for covariance matrix with factor structure

Similarly, we can compute the Hellinger distance. We can see that in this case log band fraction still outperform others in terms of Hellinger distance.

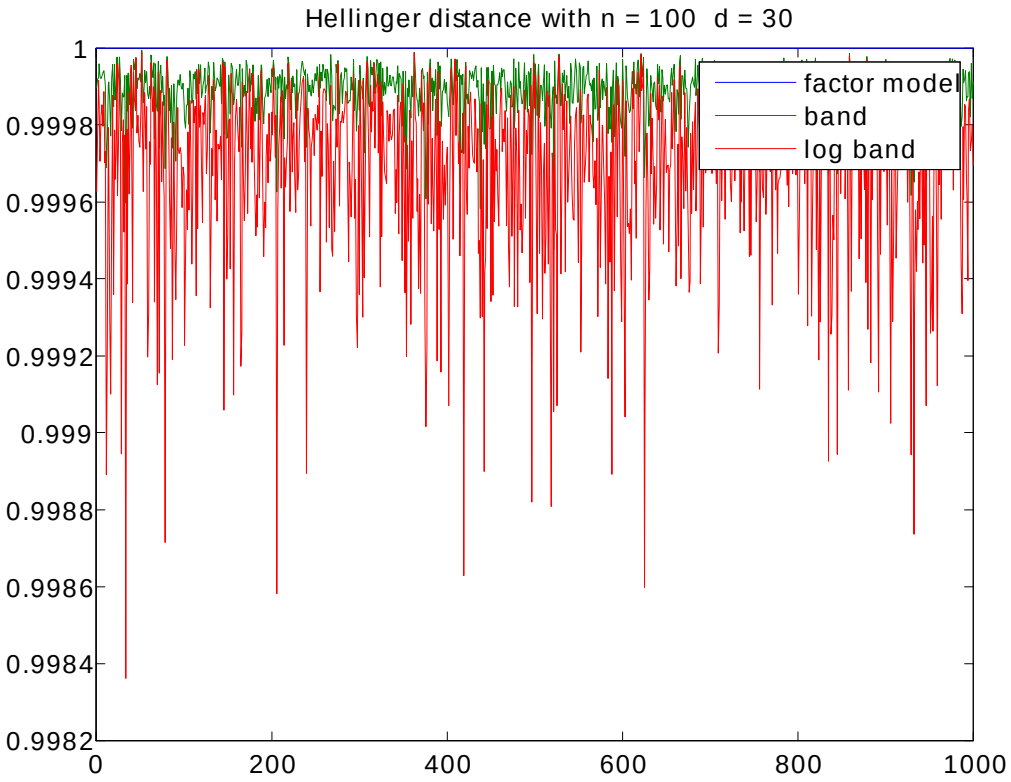

Figure 6: Hellinger distance for factor model, band fraction approximation, and log band fraction approximation for covariance matrix with factor structure with N=1000

5.8.3 Empirical sample covariance matrix

Here we compare the performance of the factor model, band fraction and log band fraction with the sample covariance matrix from the empirical data. We download the daily close prices of S&P 500 symbols from 01-01-2000 to 03-20-2015. We select the symbols with price history since 01-01-2000. There are totally 385 symbols and 3827 trading days. We compute the sample covariance of the 385 symbols with 3827 days of prices. Then we apply factor model, band fraction approximation, and log band fraction approximation to the sample covariance matrix with difference choices of d. Following table shows that as d increases, the information distance of log band fraction is getting smaller, and log band fraction is outperforming band fraction and factor model consistently.

Table 5.1: Information distance of factor model, band fraction, and log band fraction with different d

Information Distance	$d = 10$	$d = 20$	$d = 30$	$d = 40$	$d = 50$	$d = 60$	$d = 70$
Factor Model	109.0676	176.4427	152.5030	93.2462	110.2154	111.1381	37.8097
Band Fraction	63.5563	46.8955	37.0726	30.1355	24.3670	19.6291	15.6979
Log Band Fraction	61.4488	45.7505	35.9972	28.9177	23.2126	18.5553	14.5000

5.8.4 Portfolio Selection with Log Band Fraction

Here we compare the performance of the factor model, band fraction and log band fraction in portfolio optimization with empirical data. We use the same data set as in the previous section. That is, the daily close prices of S&P 500 symbols from 01-01-2000 to 03-20-2015. We will consider a simple Markowitz optimization problem as following

$$\min \frac{1}{2} w^T \Sigma w - r^T w$$

$$s.t\ w^T e = 1$$

We use tomorrow's realized log return as today's expected return r. Then we use the three approximation methods with $d = 30$. That is, we keep all parts the same except the covariance estimation. Also, we assume there is no transaction cost here. We run the simulation from 07-10-2013 to 03-20-2015 with a moving window of 2 years for covariance estimation. Following figure shows that log band fraction is outperforming factor model and band fraction.

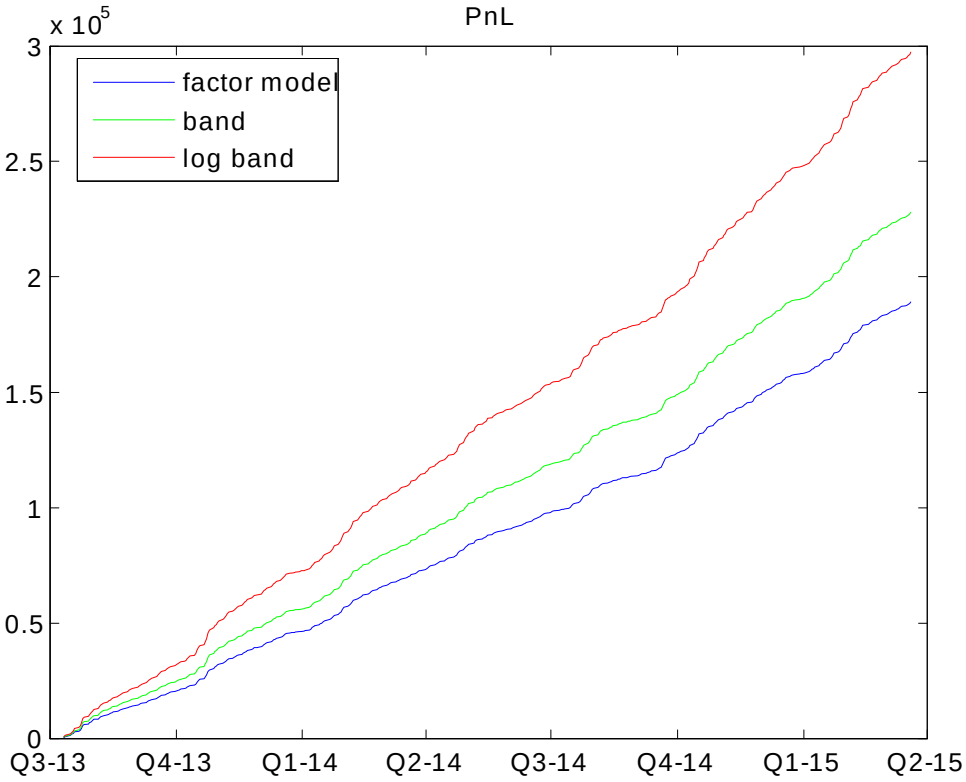

Figure 7: Equity curve with covariance estimation with factor model, band fraction, and log band fraction approximation.

Improved Factor Model Estimation Method with Log Band Fraction

6 Low Volatility Strategy with Log Band Fraction

In this section, we will propose our long-only strategy and 130-30 fund strategy. We optimize our portfolio based on our risk estimation with log band fraction approximation technic. We will compare our long-only strategy to SPY and SPLV, which are the ETFs for S&P 500 index and SP5LVI index. We will also compare our 130-30 fund to CSM, which is the ETF of CS13030 index. Moreover, we will apply the log band fraction to estimate the expected return of the stocks to further improve the performance of our strategies. We will show that structured covariance estimation helps reduces the noise in the expected return estimation. In the strategies we proposed, we assume five base point as the transaction cost and an evolving universe. Therefore, all the strategies we proposed are totally investable.

6.1 Long Only Strategy

Let us consider the following optimization problem

$$\min \frac{1}{2} w^T \Sigma w$$

$$s.t\ l \leq w \leq u$$

$$w^T e = 1$$

where Σ is the covariance matrix, l is the lower bound and u is the upper bound for of the portfolio weights.

In the long only portfolio, the lower bound of portfolio weights are all zeros and we use 5% as the upper bound for the portfolio weights. We assume 5 base point as the transaction cost. For the symbols in S&P 500, 5 base point is actually well above the average cost. We rebalance the portfolio quarterly, same as SPLV did, which makes it a fair comparison. We start the simulation since 05-05-2011, the launch day of the SPLV ETF. The Following figure shows that our strategy with log band fraction is outperforming others.

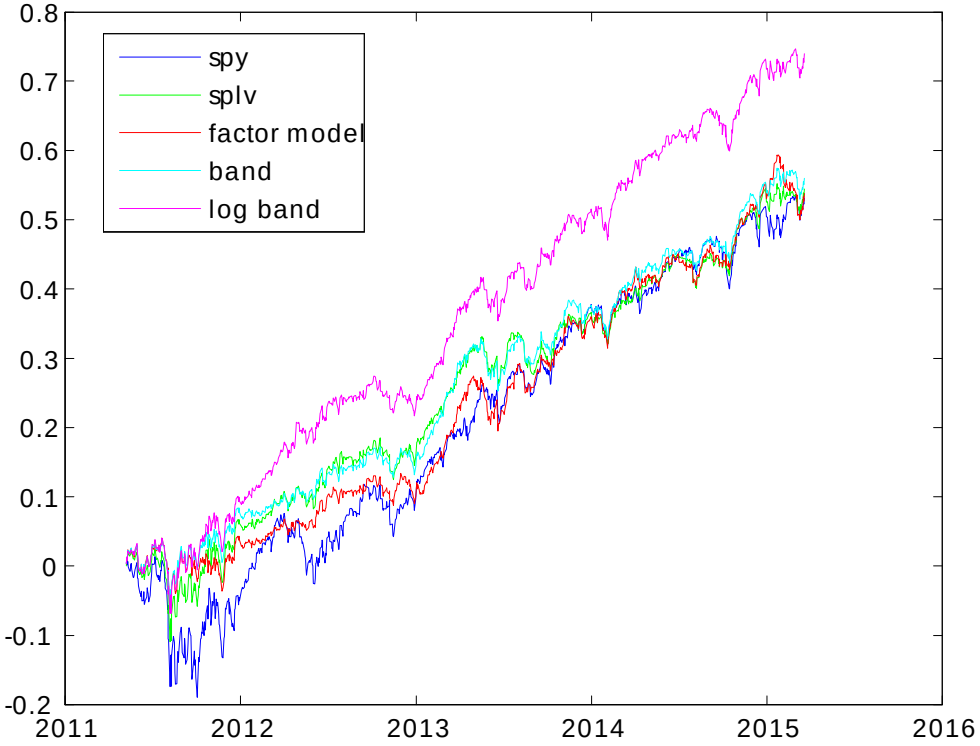

Figure 8: Equity curve of long only strategy by using factor model, band fraction, and log band fraction as the risk estimation.

Following table shows the Sharpe ratio of the strategies during different period. The strategy with log band fraction approximation has a Sharpe ratio of 1.6815 from 05-05-2011 to 03-20-2015.

Table 6.1: Sharpe ratio of long only portfolio

Sharpe Ratio	SPY	SPLV	Factor Model	Band Fraction	Log Band Fraction
whole history	0.8857	1.1762	1.2792	1.4131	1.6815
Last 1 year	1.2246	1.6023	1.3035	1.6909	1.7630
Last 2 years	1.5006	1.3582	1.6088	1.5985	1.8829
Last 3 years	1.2998	1.4883	1.6269	1.7075	1.9197

We also compute the correlation between various strategies. We can see that the strategies using factor model, band fraction, and log band fraction are all having higher correlation with SPLV than SPY, which is reasonable since they mainly focused on the minimization of the portfolio variance.

Table 6.2: Correlation between long only strategies

Correlation	Factor Model	Band	Log Band
SPY	0.7876	0.8490	0.8842
SPLV	0.8898	0.9340	0.9164

The maximum drawdown by using the log band fraction is much smaller than the benchmarks. During the whole history of simulation, SPY and SPLV have maximum drawdown of 20.29% and 13.34%, respectively, while band fraction and log band fraction only has a maximum drawdown of 9.74% and 10.86%.

Table 6.3: Maximum drawdown of long only strategies

	Maximum drawdown (100%)
SPY	0.2029
SPLV	0.1334
Factor Model	0.1021
Band Fraction	0.0974
Log Band Fraction	0.1086

Following figure shows the total long position and total short position over the simulation period. Notice that the portfolio is rebalanced quarterly. Therefore, we can observe quarterly rebalanced behavior through the total long position size.

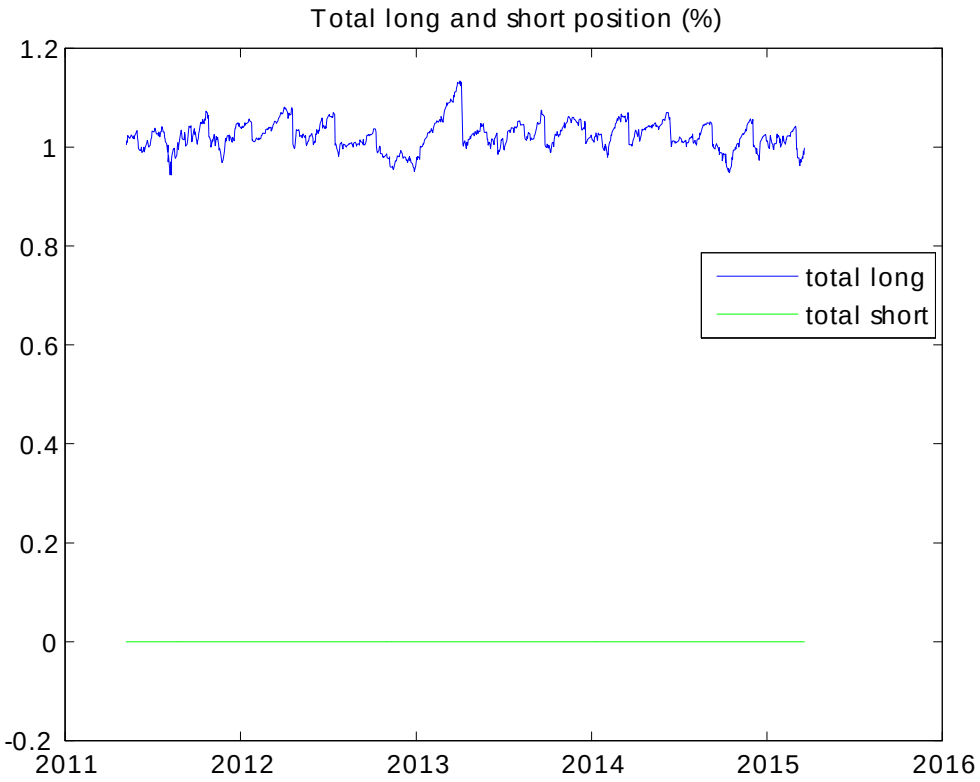

Figure 9: Total long and total short position of long only portfolio with log band fraction

6.2 130-30 fund strategy

Consider the following optimization problem

$$\min \frac{1}{2} w^T \Sigma w$$

$$s.t\ l \leq w \leq u$$

$$sum(w(w > 0)) = total\_long$$
$$sum(w(w < 0)) = -total\_short$$

We will solve the optimization problem by splitting the portfolio weights w into buy and sell positions

$$w = b - s$$

where

$$b \circ s = \begin{bmatrix} b_1 s_1 \\ b_2 s_2 \\ \vdots \\ b_n s_n \end{bmatrix} = 0$$

$$b \geq 0$$

$$s \geq 0$$

therefore, constraints of total long and short can be written as

$$\begin{aligned} b^T e &= total\_long \\ s^T e &= total\_short \end{aligned}$$

We would convert those two strict constraints to soft constraints by adding following terms to the objective function

$$\begin{aligned} & (b^T e - total\_long)(b^T e - total\_long)^T \\ = \quad & b^T e e^T b - 2 * total\_long * e^T b + (total\_long)^2 \end{aligned}$$

$$\begin{aligned} & (s^T e - total\_short)(s^T e - total\_short)^T \\ = \quad & s^T e e^T s - 2 * total\_short * e^T s + (total\_short)^2 \end{aligned}$$

The complementary condition

$$b \circ s = 0$$

can be achieved by adding the term

$$b^T I s$$

Thus, the objective function becomes

$$
\begin{aligned}
& \frac{1}{2}(b-s)^T\Sigma(b-s) + (b^Te - total\_long)^2 + (s^Te - total\_short)^2 + b^TIs \\
= \; & \frac{1}{2}\begin{bmatrix} b \\ s \end{bmatrix}^T \begin{bmatrix} I \\ -I \end{bmatrix} \Sigma \begin{bmatrix} I & -I \end{bmatrix} \begin{bmatrix} b \\ s \end{bmatrix} + \begin{bmatrix} b \\ s \end{bmatrix}^T \begin{bmatrix} ee^T & \\ & ee^T \end{bmatrix} \begin{bmatrix} b \\ s \end{bmatrix} + ... \\
& + \begin{bmatrix} -2 * total\_long * e \\ -2 * total\_short * e \end{bmatrix}^T \begin{bmatrix} b \\ s \end{bmatrix} + b^TIs + \text{constant} \\
= \; & \frac{1}{2}\begin{bmatrix} b \\ s \end{bmatrix}^T \begin{bmatrix} \Sigma + 2ee^T & -\Sigma + I \\ -\Sigma + I & \Sigma + 2ee^T \end{bmatrix} \begin{bmatrix} b \\ s \end{bmatrix} + \begin{bmatrix} -2 * total\_long * e \\ -2 * total\_short * e \end{bmatrix}^T \begin{bmatrix} b \\ s \end{bmatrix} + \text{constant}
\end{aligned}
$$

By setting $x = \begin{bmatrix} b \\ s \end{bmatrix}$, we can rewrite the optimization problem as

$$
\min \frac{1}{2}x^T \begin{bmatrix} \Sigma + 2ee^T & -\Sigma + I \\ -\Sigma + I & \Sigma + 2ee^T \end{bmatrix} x + \begin{bmatrix} -2 * total\_long * e \\ -2 * total\_short * e \end{bmatrix}^T x
$$

$$
s.t.\ 0 \le x \le u
$$

which is a quadratic optimization problem with box constraints.

We use $total\_long = 1.3$ and $total\_short - 0.3$ to get the 130-30 fund. The upper bound u is set to be 0.05, same as in the long only strategy. Following figure shows the quity curve of the 130-30 fund strategy with factor model, band fraction, and log band fraction estimation.

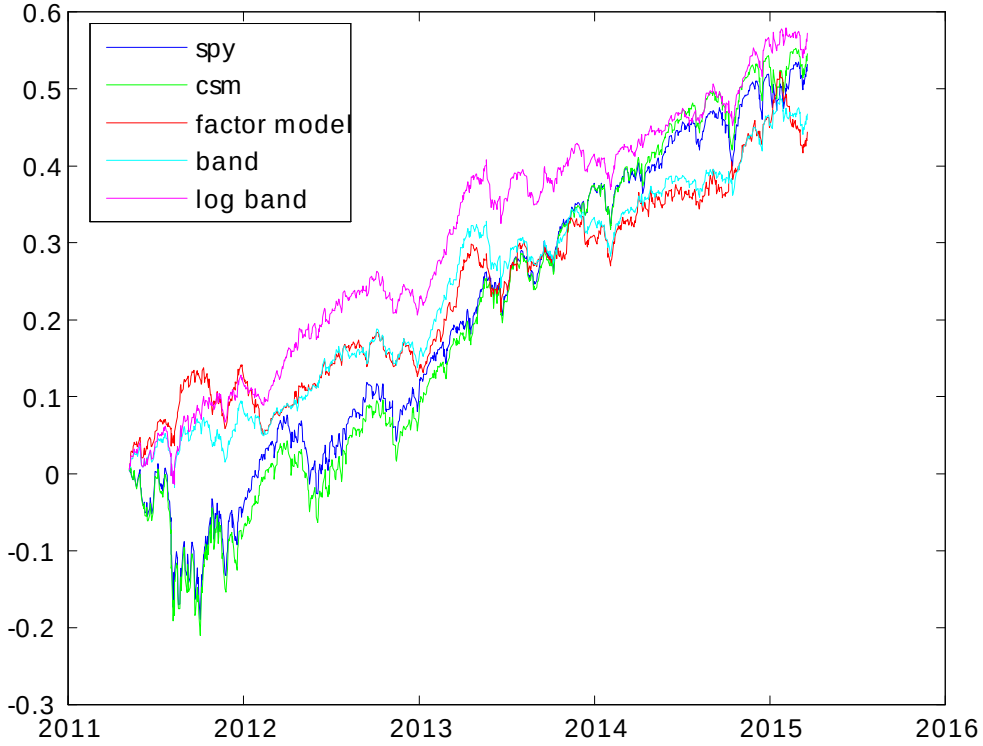

Figure 10: Equity curve of 130-30 strategy by using factor model, band fraction, and log band fraction as the risk estimation.

We can also compare Sharpe Ratio of these equity curves. Portfolio with log band fraction has the highest Sharpe ratio among these strategies. One observation is that the 130-30 fund ETF, CSM, is underperforming the long only ETF, SPLV, since 2011.

Table 6.4: Sharpe ratio of 130-30 portfolio

Sharpe Ratio	SPY	CSM	Factor Model	Band Fraction	Log Band Fraction
whole history	0.8857	0.9026	1.1853	1.4095	1.5894
Last 1 year	1.2246	1.2990	1.1652	1.5067	1.6770
Last 2 years	1.5006	1.6189	1.0956	1.2507	1.3700
Last 3 years	1.2998	1.4050	1.3149	1.6117	1.6812

Again, we can compare other portfolio statistics such as correlation as shown in the following table.

Table 6.5: Correlation between 130-30 strategies

Correlation	Factor Model	Band	Log Band
SPY	0.3981	0.6240	0.7261
CSM	0.3855	0.6035	0.7112

In terms of maximum drawdown, the log band fraction portfolio is having the smallest number as well. Also notice that CSM is suffering larger drawdown during 2011 than SPY, even thougth CSM is providing hedge using short position.

Table 6.6: Maximum drawdown of 130-30 strategies

	Maximum drawdown (100%)
SPY	0.2029
CSM	0.2225
Factor Model	0.1059
Band Fraction	0.0865
Log Band Fraction	0.0835

The log band fraction portfolio is rebalanced monthly, same as CSM did. Following figure shows the total long and short position.

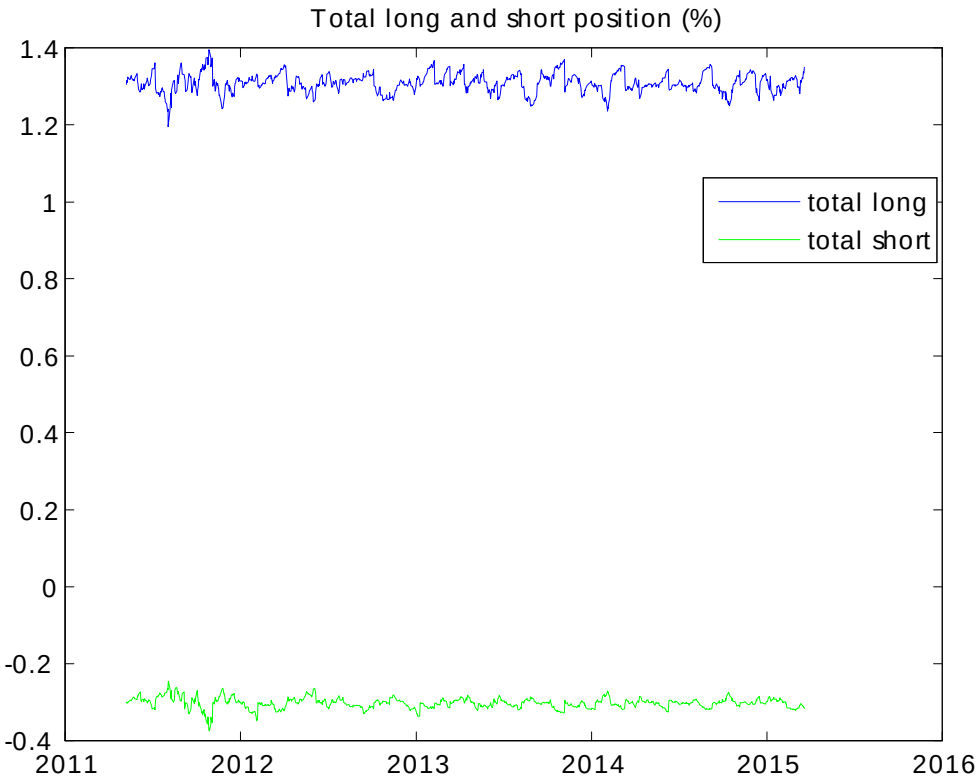

Figure 11: Total long and total short position of 130-30 portfolio with log band fraction

6.3 Expected Return From Structured Covariance Estimation

Here we will show that we can improve the sample mean estimation by using structured covariance estimation. Given the sample x from the normal distribution $N(\mu, C^2)$, we have the following equations

$$E(x) = \mu$$

$$E(xx^*) = C^2 + \mu\mu^*$$

$$E\left(\begin{bmatrix} 1 \\ x \end{bmatrix} \begin{bmatrix} 1 & x^T \end{bmatrix}\right) = \begin{bmatrix} 1 & \mu^* \\ \mu & C^2 + \mu\mu^* \end{bmatrix}$$

Thus, we can calculate the sample mean from the above equations by using structured covariance matrix estimation. Next, we will apply this to the long only strategy and 130-30 strategy, with expected return added in the optimization.

6.4 Long Only Strategy with Structured Expected Return

Let us add the expected return to the long only optimization problem

$$\min \frac{1}{2} w^T \Sigma w - r^T w$$

$$s.t\ l \leq w \leq u$$

$$w^T e = 1$$

Again, we rebalance our portfolio quarterly, same as SPLV did. The lower and upper bound are the same as the one in the long only strategy without expected return embedded. Following figure shows that the structured expected return could helps reduce the noise in the equity curve.

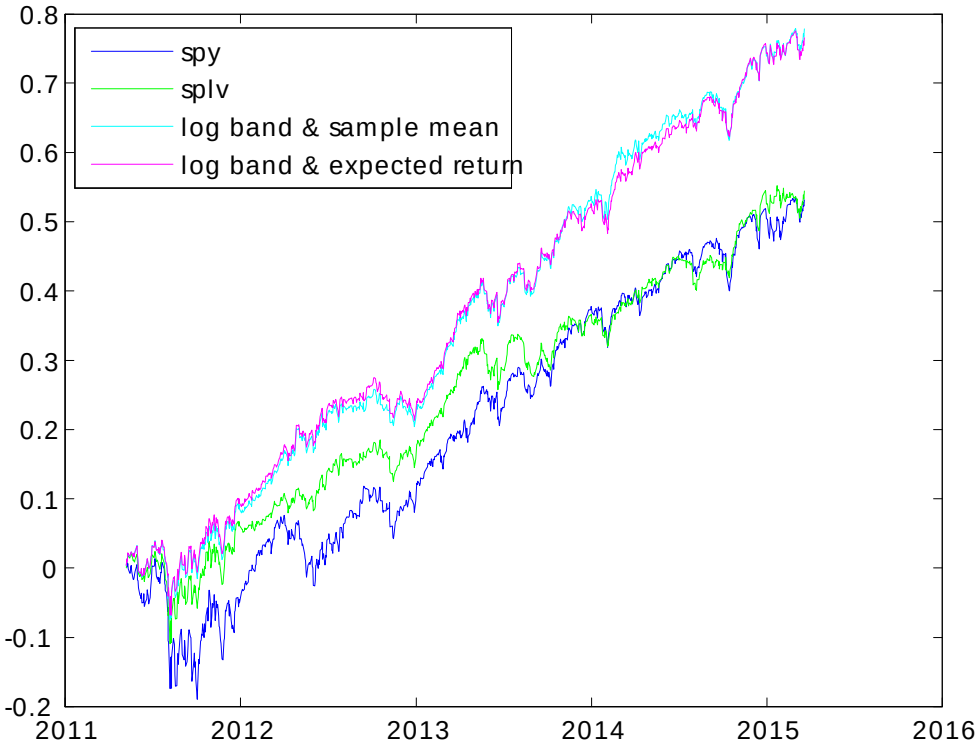

Figure 12: Equity curve of long only with expected return strategy by using log band fraction

We can see the effect of structured expected return more clearly by comparing the Sharpe ratio. The log band fraction embedded with structured expected return is outperforming other strategies.

Table 6.7: Sharpe ratio of long only portfolio with expected return embeded

Sharpe Ratio	SPY	SPLV	Log Band & sample mean	Log Band & structured r
whole history	0.8857	1.1762	1.6985	1.7298
Last 1 year	1.2246	1.6023	1.6354	1.8446
Last 2 years	1.5006	1.3582	1.9760	1.9946
Last 3 years	1.2998	1.4883	1.9588	1.9865

Following table shows the correlation of our strategies with the benchmarks. The correlation between SPY and log band fraction with structured expected return is 0.8847, while the correlation between SPLV and log band fraction with structured expected return is 0.9171.

Table 6.8: Correlation between long only strategies with expected return embeded

Correlation	Log Band_Sample Mean	Log Band_Structured r
SPY	0.8794	0.8847
SPLV	0.9069	0.9171

Also, let us compare the maximum drawdown of those strategies. In this case, log band fraction with structured expected return has the smallest maximum drawdown among these strategies.

Table 6.9: Maximum drawdown of long only strategies with expected return embeded

	Maximum drawdown (100%)
SPY	0.2029
SPLV	0.1334
Log Band_Sample Mean	0.1131
Log Band_Structured r	0.1083

The portfolio is rebalanced quarterly. Following figure shows the total long and short

position size of log band fraction portfolio through time.

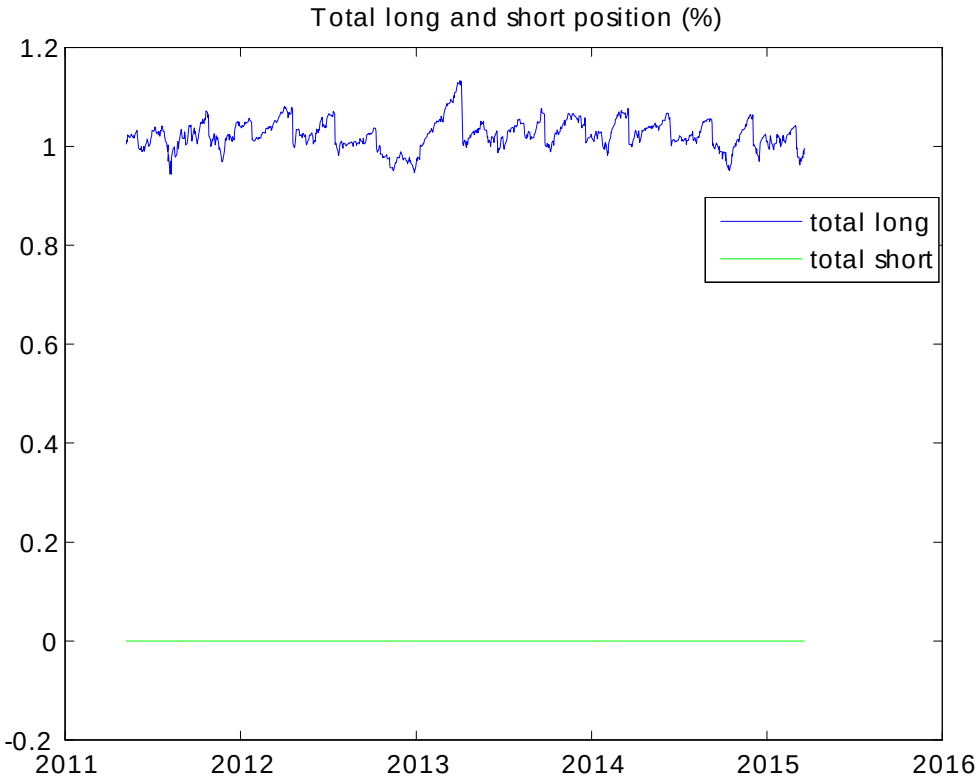

Figure 13: Total long and total short position of long only portfolio with log band fraction and structured expected return embedded

7 Conclusion

In this thesis, we compute the information matrix and metric tensor of band fraction representation of covariance matrix. We compute the first derivative of information distance of band fraction representation. We propose log band fraction structure to the covariance matrix. We compare the log band fraction approximation to factor model and band fraction approximation. We show that log band fraction is outperforming factor model and band fraction in terms of information distance and Hellinger distance.

We apply log band fraction to construct low volatility strategies. We propose our long only strategy and 130-30 strategy with log band fraction in the covariance estimation. Furthermore, we show that structured covariance estimation can be used to get the structured sample mean, which could used to improve the performance of strategies. Our strategies are shown to significantly outperform the widely used benchmarks, i.e., SPY, SPLV, CSM.

A Appendix: Computation of Kullback-Leibler divergence

In probability theory and information theory [107] [108] [109], the Kullback–Leibler divergence (KLD) measures the difference between two probability distributions P and Q. The Kullback-Leibler divergence from P to Q measures the information lost when we approximate P with Q.

Definition 3. *Kullback-Leibler divergence: For distributions P and Q of a continuous random variable, Kullback-Leibler divergence is defined as*

$$D_{KL}\left(P||Q\right)=\int_{-\infty}^{\infty}p\left(x\right)\log\left(\frac{p\left(x\right)}{q\left(x\right)}\right)dx$$

where p and q are the probability densities of P and Q.

Typically P is the real distribution of data or observations, while Q represents a model or approximation of P.

From Gibbs' inequality, it can be shown that the Kullback-Leibler divergence is always non-negative

$$D_{KL}\left(P||Q\right)\geq 0$$

with equality if and only if $P=Q$ almost everywhere.

Consider two multivariate Gaussian distributions $P_1\left(\mu_1,C_1^2\right)$ and $P_2\left(\mu_2,C_2^2\right)$, the probability densities are given as

$$\begin{aligned}p_1\left(x_1,\cdots,x_k\right) &= \frac{1}{\left(2\pi\right)^{\frac{k}{2}}\det\left(C_1^2\right)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(x-\mu_1\right)^T C_1^{-2}\left(x-\mu_1\right)\right)\\ p_2\left(x_1,\cdots,x_k\right) &= \frac{1}{\left(2\pi\right)^{\frac{k}{2}}\det\left(C_2^2\right)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\left(x-\mu_2\right)^T C_2^{-2}\left(x-\mu_2\right)\right)\end{aligned}$$

the Kullback-Leibler divergence from P_1 to P_2 is given as

$$\begin{aligned}
& D_{KL}\left(P_1\left(\mu_1, C_1^2\right)||P_2\left(\mu_2, C_2^2\right)\right) \\
= & \int p_1 \log \frac{p_1}{p_2} dx \\
= & \int p_1 \left[\log \frac{\det\left(C_2^2\right)^{\frac{1}{2}}}{\det\left(C_1^2\right)^{\frac{1}{2}}} + \frac{1}{2}\left(x-\mu_2\right)^T C_2^{-2}\left(x-\mu_2\right) - \frac{1}{2}\left(x-\mu_1\right)^T C_1^{-2}\left(x-\mu_1\right)\right] dx \\
= & \log \frac{\det C_2}{\det C_1} \int p_1 dx + \frac{1}{2}\int p_1 tr\left(\left(x-\mu_2\right)^T C_2^{-2}\left(x-\mu_2\right)\right) dx - \frac{1}{2}\int p_1 tr\left(\left(x-\mu_1\right)^T C_1^{-2}\left(x-\mu_1\right)\right) dx \\
= & \log \frac{\det C_2}{\det C_1} + \frac{1}{2}\int p_1 tr\left(C_2^{-2}\left(x-\mu_2\right)\left(x-\mu_2\right)^T\right) dx - \frac{1}{2}\int p_1 tr\left(C_1^{-2}\left(x-\mu_1\right)\left(x-\mu_1\right)^T\right) dx \\
= & \log \frac{\det C_2}{\det C_1} + \frac{1}{2} E\left[tr\left(C_2^{-2}\left(x-\mu_2\right)\left(x-\mu_2\right)^T\right)\right] - \frac{1}{2} E\left[tr\left(C_1^{-2}\left(x-\mu_1\right)\left(x-\mu_1\right)^T\right)\right] \\
= & \log \frac{\det C_2}{\det C_1} + \frac{1}{2} tr\left[C_2^{-2} E\left(\left(x-\mu_2\right)\left(x-\mu_2\right)^T\right)\right] - \frac{1}{2} tr\left[C_1^{-2} E\left(\left(x-\mu_1\right)\left(x-\mu_1\right)^T\right)\right] \\
= & \log \frac{\det C_2}{\det C_1} + \frac{1}{2} tr\left[C_2^{-2} E\left(\left(x-\mu_1+\mu_1-\mu_2\right)\left(x-\mu_1+\mu_1-\mu_2\right)^T\right)\right] - \frac{1}{2} tr\left(I\right) \\
= & \log \frac{\det C_2}{\det C_1} + \frac{1}{2} tr\left[C_2^{-2} E\left(\left(x-\mu_1\right)\left(x-\mu_1\right)^T + \left(\mu_1-\mu_2\right)\left(\mu_1-\mu_2\right)^T\right)\right] - \frac{1}{2} tr\left(I\right) \\
= & \log\det\left(C_1^{-1} C_2\right) - \frac{1}{2} trI + \frac{1}{2} tr\left(\left(C_2^{-1} C_1\right)^2\right) + \frac{1}{2}\left(\mu_1-\mu_2\right)^T C_2^{-2}\left(\mu_1-\mu_2\right)
\end{aligned}$$

For the special case $\mu_1 = \mu_2 = 0$, $C_2^2 = I$, that is, the Kullback-Leibler divergence between a zero-mean multivariate Gaussian distribution and a white noise is given as

$$\begin{aligned}
D_{KL}\left(P_1||P_2\right) &= \log\det\left(C_1^{-1}\right) - \frac{1}{2} trI + \frac{1}{2} tr\left(C_1^2\right) \\
&= \log\prod_{i=1}^{k}\sqrt{\frac{1}{\lambda_i}} + \frac{1}{2}\sum_{i=1}^{k}\lambda_i - \frac{1}{2}k \\
&= -\frac{1}{2}\sum_{i=1}^{k}\log\lambda_i + \frac{1}{2}\sum_{i=1}^{k}\lambda_i - \frac{1}{2}k
\end{aligned}$$

where λ_i are the eigenvalues of C_1^2.

B Appendix: Computation of Hellinger distance

The definition of Hellinger distance in Lehmann [105] is as follows

Definition 4. *Hellinger distance: Let P_1 and P_2 be probabilities. The Hellinger distance $H(P_1, P_2)$ between P_1 and P_2 is given by*

$$H^2\left(P_1, P_2\right) = \frac{1}{2}\int\left[\sqrt{p_1\left(x\right)} - \sqrt{p_2\left(x\right)}\right]^2 d\mu\left(x\right)$$

where p_i is the density of P_i with respect to any measure μ dominating P_1 and P_2.

The Hellinger distance defines a metric. Let $\rho(P_1, P_2)$ be the affinity between P_1 and P_2, defined as

$$\rho(P_1, P_2) = \int \sqrt{p_1(x) p_2(x)} d\mu(x)$$

then

$$H^2(P_1, P_2) = 1 - \rho(P_1, P_2)$$

By the Cauchy-Schwarz inequality,

$$0 \leq \rho(P_1, P_2) \leq 1$$

and

$$\rho(P_1, P_2) = 1$$

if and only if $P_1 = P_2$.

Furthermore,

$$\rho(P_1, P_2) = 0$$

if and only if P_0 and P_1 are mutually singular. It follows that $H(P_1, P_2) = 0$ if and only if $P_1 = P_2$.

The following theorem shows that Hellinger distance and total variation distance generate the same topology.

Theorem 2. *The Hellinger distance and total variation distance satisfy*

$$H^2(P_1, P_2) \leq \frac{1}{2} \|P_1 - P_2\|_1 \leq H(P_1, P_2) \left| 2 - H(P_1, P_2)^2 \right|^{1/2}$$

Consider two zero-mean multivariate Gaussian distributions $P_1\left(x, 0, C_1^2\right)$ and $P_2\left(x, 0, C_2^2\right)$, the affinity between P_1 and P_2 is

$$\begin{aligned}
\rho(P_1, P_2) &= \int \sqrt{p_1(x) p_2(x)} d\mu(x) \\
&= \int \frac{1}{(2\pi)^{\frac{k}{2}} \det(C_1)^{\frac{1}{2}} \det(C_2)^{\frac{1}{2}}} \exp\left(-\frac{1}{4} x^T \left(C_1^{-2} + C_2^{-2} \right) x \right) d\mu(x) \\
&= \frac{\det\left(\frac{C_1^{-2} + C_2^{-2}}{2} \right)^{-\frac{1}{2}}}{\det(C_1)^{\frac{1}{2}} \det(C_2)^{\frac{1}{2}}} \int \frac{1}{(2\pi)^{\frac{k}{2}} \det\left(\frac{C_1^{-2} + C_2^{-2}}{2} \right)^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} x^T \left(\frac{C_1^{-2} + C_2^{-2}}{2} \right) x \right) d\mu(x) \\
&= \det\left(\frac{C_1^{-2} + C_2^{-2}}{2} \right)^{-\frac{1}{2}} \det(C_1)^{-\frac{1}{2}} \det(C_2)^{-\frac{1}{2}}
\end{aligned}$$

therefore, the Helllinger distance between P_1 and P_2 is

$$H^2(P_1, P_2) = 1 - \det\left(\frac{C_1^{-2} + C_2^{-2}}{2}\right)^{-\frac{1}{2}} \det(C_1)^{-\frac{1}{2}} \det(C_2)^{-\frac{1}{2}}$$

For the special case $C_2 = I$, we can further simplify the formula

$$\begin{aligned} H^2(P_1, P_2) &= 1 - \det\left(\frac{C_1^{-2} + I}{2}\right)^{-\frac{1}{2}} \det(C_1)^{-\frac{1}{2}} \det(I)^{-\frac{1}{2}} \\ &= 1 - \prod_{i=1}^{k} \left(\frac{\lambda_i^{-1} + 1}{2}\right)^{-\frac{1}{2}} \lambda_i^{-\frac{1}{4}} \\ &= 1 - \prod_{i=1}^{k} \left(\frac{\lambda_i^{-\frac{1}{2}} + \lambda_i^{\frac{1}{2}}}{2}\right)^{-\frac{1}{2}} \\ &= 1 - \prod_{i=1}^{k} \left(\frac{2}{\lambda_i^{\frac{1}{2}} + \lambda_i^{-\frac{1}{2}}}\right)^{\frac{1}{2}} \end{aligned}$$

C Appendix: Total Variation Distance and Information Distance

C.1 Total Variation Distance

Definition 5. *Total Variation Distance: The total variation distance between two probability distributions P_1 and P_2 is defined as*

$$\|P_1 - P_2\|_1 = \int |p_1 - p_2| \, d\mu$$

and is independent of the dominating measure μ.

It is easy to see that total variation distance is a metric. Total variation distance controls hypothesis testing. Unfortunately, this quantity is often difficult to compute, but it relates to other distances which are easy to compute such as Hellinger distance.

C.2 Information Distance

Information distance corresponds to the Fisher information matrix. Given a transfer function $f(z)$ and $g(z)$

$$\log \frac{f(z)}{g(z)} = a_0 + a_1 z + a_2 z^2 + \cdots$$

the Information distance between $f(z)$ and $g(z)$ is

$$\left\|\log\frac{f(z)}{g(z)}\right\|^2_{H^2(D)} = |a_0|^2 + |a_1|^2 + |a_2|^2 + \cdots$$

For the case $g(z) = 1$, the Information distance from $f(z)$ to white noise is

$$\|\log f(z)\|^2_{H^2(D)} = f_0^2 + f_1^2 + f_2^2 + \cdots$$

Specifically, the Information distance between two Gaussian distributions $P_1\left(0, C_1^2\right)$ and $P_2\left(0, C_2^2\right)$ is

$$I^2 = \frac{1}{2}\sum_{k=1}^{n}(\log\lambda_k)^2$$

where λ_k are the eigenvalues of $C_1^{-2}C_2^2$.

D Appendix: Diagonal Iteration and Convergence Proof via KKT Equations

In this section, we will discuss the diagonal iteration, which is a fast algorithm for solving quadratic programming with box constraints. We will show the convergence proof via KKT Equations.

D.1 KKT conditions for Diagonal Iteration

The problem considered here is

$$\min\frac{1}{2}x^T\Sigma x + f^T x = \frac{1}{2}x^T Dx + (f + \frac{1}{2}VV^* x)^* x$$
$$subject\ to\ l \preceq x \preceq u$$

The KKT conditions for this original problem are

$$\begin{aligned}
l - x &\preceq 0\\
x - u &\preceq 0\\
\lambda &\succeq 0\\
v &\succeq 0\\
Dx + f + VV^* x - \lambda + v &= 0\\
\lambda_i(l_i - x_i) &= 0,\ i = 1, ...\\
v_i(x_i - u_i) &= 0,\ i = 1, ...
\end{aligned}$$

The approximate problem

$$\min \frac{1}{2}x^T Dx + (f + VV^* x_k)^* x$$

$$\textit{subject to } l \preceq x \preceq u$$

where x_k is fixed

$$x_{k+1} = \arg \min_{l \preceq x \preceq u} \frac{1}{2}x^T Dx + (f + VV^* x_k)^* x$$

The KKT conditions for the approximate problem are

$$\begin{aligned}
l - x_{k+1} &\preceq 0 \\
x_{k+1} - u &\preceq 0 \\
\lambda_{k+1} &\succeq 0 \\
v_{k+1} &\succeq 0 \\
Dx_{k+1} + f + VV^* x_k - \lambda_{k+1} + v_{k+1} &= 0 \\
\lambda_{(k+1)i}(l_i - x_{(k+1)i}) &= 0,\ i = 1, ... \\
v_{(k+1)i}(x_{(k+1)i} - u_i) &= 0,\ i = 1, ...
\end{aligned}$$

Here λ_{k+1} and v_{k+1} are the langrange multipliers correspond to x_{k+1}. $\lambda_{(k+1)i}$ means the ith component of λ_{k+1}.

The dual problem for solving λ_{k+1} and v_{k+1} is complicated and hard to get explicit expressions for λ and v for the approximate problem.

D.2 Convergence proof via comparison with central path

Now let us consider the Interior Point method for solving the original problem. The equivalent KKT conditions for the Interior Point method are

$$\begin{aligned} l - x &\preceq 0 \\ x - u &\preceq 0 \\ \lambda &\succeq 0 \\ v &\succeq 0 \\ Dx + f + VV^*x - \lambda + v &= 0 \\ -\lambda_i(l_i - x_i) &= \frac{1}{t},\ i = 1, ... \\ -v_i(x_i - u_i) &= \frac{1}{t},\ i = 1, ... \end{aligned}$$

Here $t > 0$. As $t \to \infty$, the point in the central point will approach the optimal point x^*.

For the approximate problem, if at each step we use Interior Point method to get x_{k+1} based on x_k (actually, we know that we do not need to use Interior Point method for the approximate problem since it is simple engough)

$$\begin{aligned} l - x_{k+1} &\preceq 0 \\ x_{k+1} - u &\preceq 0 \\ \lambda_{k+1} &\succeq 0 \\ v_{k+1} &\succeq 0 \\ Dx_{k+1} + f + VV^*x_k - \lambda_{k+1} + v_{k+1} &= 0 \\ -\lambda_{(k+1)i}(l_i - x_{(k+1)i}) &= \frac{1}{M},\ i = 1, ... \\ -v_{(k+1)i}(x_{(k+1)i} - u_i) &= \frac{1}{M},\ i = 1, ... \end{aligned}$$

As $M \to \infty$, the solution approaches x_{k+1}.

Remark: notice the difference between the two solutions obtained from Interior Point method here. For $t \to \infty$, the solution x^* is the final optimal point for the original problem. However, for $M \to \infty$, we get x_{k+1}, which is only the result of one step.

Let us compare the two equations

$$\begin{aligned} Dx_{k+1} + f + VV^*x_k - \lambda_{k+1} + v_{k+1} &= 0 \\ Dx + f + VV^*x - \lambda + v &= 0 \end{aligned}$$

use

$$\begin{aligned}
-\lambda_i(l_i - x_i) &= \frac{1}{t},\ i = 1, ... \\
-v_i(x_i - u_i) &= \frac{1}{t},\ i = 1, ... \\
-\lambda_{(k+1)i}(l_i - x_{(k+1)i}) &= \frac{1}{M},\ i = 1, ... \\
-v_{(k+1)i}(x_{(k+1)i} - u_i) &= \frac{1}{M},\ i = 1, ...
\end{aligned}$$

we get

$$\begin{aligned}
\lambda &= -\frac{1}{t}\begin{bmatrix} \frac{1}{l_1 - x_1} \\ . \\ . \\ . \\ \frac{1}{l_n - x_n} \end{bmatrix} \\
v &= -\frac{1}{t}\begin{bmatrix} \frac{1}{x_1 - u_1} \\ . \\ . \\ . \\ \frac{1}{x_n - u_n} \end{bmatrix} \\
\lambda_{k+1} &= -\frac{1}{M}\begin{bmatrix} \frac{1}{l_1 - x_{(k+1)1}} \\ . \\ . \\ . \\ \frac{1}{l_n - x_{(k+1)n}} \end{bmatrix} \\
v_{k+1} &= -\frac{1}{M}\begin{bmatrix} \frac{1}{x_{(k+1)1} - u_1} \\ . \\ . \\ . \\ \frac{1}{x_{(k+1)n} - u_n} \end{bmatrix}
\end{aligned}$$

then

$$\begin{aligned}
Dx_{k+1} + f - VV^* x_k - \lambda_{k+1} + v_{k+1} &= 0 \\
Dx + f - VV^* x - \lambda + v &= 0
\end{aligned}$$

becomes

$$Dx_{k+1} + f + VV^* x_k + \frac{1}{M}\begin{bmatrix} \frac{1}{l_1 - x_{(k+1)1}} \\ . \\ . \\ . \\ \frac{1}{l_n - x_{(k+1)n}} \end{bmatrix} - \frac{1}{M}\begin{bmatrix} \frac{1}{x_{(k+1)1} - u_1} \\ . \\ . \\ . \\ \frac{1}{x_{(k+1)n} - u_n} \end{bmatrix} = 0$$

$$Dx + f + VV^* x + \frac{1}{t}\begin{bmatrix} \frac{1}{l_1 - x_1} \\ . \\ . \\ . \\ \frac{1}{l_n - x_n} \end{bmatrix} - \frac{1}{t}\begin{bmatrix} \frac{1}{x_1 - u_1} \\ . \\ . \\ . \\ \frac{1}{x_n - u_n} \end{bmatrix} = 0$$

substract the first equation to the second equation

$$D(x_{k+1} - x) + VV^*(x_k - x) + \frac{1}{M}\begin{bmatrix} \frac{1}{l_1 - x_{(k+1)1}} \\ . \\ . \\ . \\ \frac{1}{l_n - x_{(k+1)n}} \end{bmatrix} - \frac{1}{M}\begin{bmatrix} \frac{1}{x_{(k+1)1} - u_1} \\ . \\ . \\ . \\ \frac{1}{x_{(k+1)n} - u_n} \end{bmatrix} - (\frac{1}{t}\begin{bmatrix} \frac{1}{l_1 - x_1} \\ . \\ . \\ . \\ \frac{1}{l_n - x_n} \end{bmatrix} - \frac{1}{t}\begin{bmatrix} \frac{1}{x_1 - u_1} \\ . \\ . \\ . \\ \frac{1}{x_n - u_n} \end{bmatrix}) = 0$$

use $t = M$ (t is very large), we have

$$\begin{aligned} \left\|D^{-1}VV^*(x_k - x)\right\| &= \left\| x_{k+1} - x + \frac{1}{t}D^{-1}\begin{bmatrix} \frac{x_{(k+1)1} - x_1}{\left(l_1 - x_{(k+1)1}\right)(l_1 - x_1)} \\ . \\ . \\ . \\ \frac{x_{(k+1)n} - x_n}{\left(l_n - x_{(k+1)n}\right)(l_n - x_n)} \end{bmatrix} + \frac{1}{t}D^{-1}\begin{bmatrix} \frac{x_{(k+1)1} - x_1}{\left(x_{(k+1)1} - u_1\right)(x_1 - u_1)} \\ . \\ . \\ . \\ \frac{x_{(k+1)n} - x_n}{\left(x_{(k+1)n} - u_n\right)(x_n - u_n)} \end{bmatrix} \right\| \\ &= \left\| (x_{k+1} - x)^T \left(I + \frac{1}{t}D^{-1}A\right) \right\| \\ &= \sqrt{(x_{k+1} - x)^T (I + \frac{1}{t}D^{-1}A)(I + \frac{1}{t}D^{-1}A)(x_{k+1} - x)} \\ &= \sqrt{(x_{k+1} - x)^T (I + \frac{2}{t}D^{-1}A + \frac{1}{t^2}D^{-1}AD^{-1}A)(x_{k+1} - x)} \\ &\geq \|x_{k+1} - x\| \end{aligned}$$

where

$$A = \begin{bmatrix} \frac{1}{(l_1 - x_{(k+1)1})(l_1 - x_1)} + \frac{1}{(x_{(k+1)1} - u_1)(x_1 - u_1)} & 0 & .. & 0 \\ 0 & . & . & 0 \\ ... & . & . & ... \\ 0 & 0 & . & \frac{1}{(l_n - x_{(k+1)n})(l_n - x_n)} + \frac{1}{(x_{(k+1)n} - u_n)(x_n - u_n)} \end{bmatrix} > 0$$

therefore the convergence rate is

$$\|x_k - x\| \leq \left\|D^{-1}VV^*\right\|^k \|x_0 - x\|$$

D.3 Choice of Perturbation Matrix to ensure convergence

We only need to perturb matrix D in the case when

$$\left\|D^{-1}VV^*\right\| \geq 1$$

where

$$\begin{aligned} D + VV^* &> 0 \\ D &= diag\{d_{11}, ..., d_{nn}\} > 0,\ VV^* > 0 \\ d_{11} &\geq d_{22} \geq ... \geq d_{nn} \end{aligned}$$

we want to choose a matrix M such that

$$\left\|(D + M)^{-1}(VV^* - M)\right\| < 1$$

since we want the approximate problem to be convex and as simple as possible, we need

$$D + M > 0$$

and M to be a diagonal matrix.

$$M = diag\{m_{11}, m_{22}, ..., m_{nn}\}$$

Let us now consider a simple choice of M

$$M = sI,\ s > 0$$

then

$$\begin{aligned}\left\|(D+sI)^{-1}(VV^*-sI)\right\| &\leq \left\|(D+sI)^{-1}\right\|\|VV^*-sI\| \\ &= \frac{1}{d_{nn}+s}\|VV^*-sI\|\end{aligned}$$

since

$$\Sigma = VV^* = UTU^*$$

where

$$\begin{aligned}T &= diag\{\lambda_1,...,\lambda_n\} > 0 \\ \lambda_1 &\geq ... \geq \lambda_n \text{ are eigenvalues} \\ UU^* &= I\end{aligned}$$

the above inequality becomes

$$\begin{aligned}\left\|(D+sI)^{-1}(VV^*-sI)\right\| &\leq \frac{1}{d_{nn}+s}\|UTU^*-sI\| \\ &\leq \frac{1}{d_{nn}+s}\|U\|\|T-sI\|\|U^*\| \\ &= \frac{1}{d_{nn}+s}\|T-sI\|\end{aligned}$$

By choosing s satisfying

$$\frac{1}{d_{nn}+s}\|T-sI\| < 1$$

we can ensure convergence.

D.4 Existence of s

First, from the condition

$$\left\|D^{-1}VV^*\right\| \geq 1$$

we have

$$\begin{aligned}1 &\leq \left\|D^{-1}VV^*\right\| = \left\|D^{-1}UTU^*\right\| \\ &\leq \left\|D^{-1}\right\|\|U\|\|T\|\|U^*\| \\ &= \frac{1}{d_{nn}}\lambda_1 \\ &\Rightarrow \lambda_1 \geq d_{nn}\end{aligned}$$

There are several cases depending on the value of s :

1) If $0 < s \leq \lambda_n$, then

$$\begin{aligned} \frac{1}{d_{nn}+s}\|T-sI\| &= \frac{1}{d_{nn}+s}(\lambda_1 - s) < 1 \\ &\Rightarrow \quad \lambda_1 - s < d_{nn} + s \\ &\Rightarrow \quad \lambda_1 - d_{nn} < 2s \end{aligned}$$

that is

$$0 \leq \frac{\lambda_1 - d_{nn}}{2} < s \leq \lambda_n$$

2) If $\lambda_n < s < \lambda_1$ and $\lambda_1 - s \geq -(\lambda_n - s)$, that is, $\lambda_n < s \leq \frac{\lambda_1+\lambda_n}{2}$

then

$$\begin{aligned} \frac{1}{d_{nn}+s}\|T-sI\| &= \frac{1}{d_{nn}+s}(\lambda_1 - s) < 1 \\ &\Rightarrow \quad \lambda_1 - s < d_{nn} + s \\ &\Rightarrow \quad \lambda_1 - d_{nn} < 2s \end{aligned}$$

that is

$$\max\left\{\lambda_n, \frac{\lambda_1 - d_{nn}}{2}\right\} < s \leq \frac{\lambda_1 + \lambda_n}{2}$$

3) If $\lambda_n < s < \lambda_1$ and $\lambda_1 - s < -(\lambda_n - s)$, that is $\frac{\lambda_1+\lambda_n}{2} < s < \lambda_1$, then

$$\begin{aligned} \frac{1}{d_{nn}+s}\|T-sI\| &= -\frac{1}{d_{nn}+s}(\lambda_n - s) < 1 \\ &\Rightarrow \quad s - \lambda_n < s + d_{nn} \\ &\Rightarrow \quad \lambda_n + d_{nn} > 0 \text{ (always satisfied)} \end{aligned}$$

then

$$\frac{\lambda_1 + \lambda_n}{2} < s < \lambda_1$$

4) If $s \geq \lambda_1$, then

$$\begin{aligned} \frac{1}{d_{nn}+s}\|T-sI\| &= -\frac{1}{d_{nn}+s}(\lambda_n - s) < 1 \\ &\Rightarrow \quad s - \lambda_n < s + d_{nn} \\ &\Rightarrow \quad \lambda_n + d_{nn} > 0 \text{ (always satisfied)} \end{aligned}$$

then

$$s \geq \lambda_1$$

Observe the above four cases, we can see that for case 3 and case 4, by choosing

$$s \geq \frac{\lambda_1 + \lambda_n}{2}$$

we can have

$$\left\|(D + sI)^{-1}(VV^* - sI)\right\| < 1$$

For case 1 and case 2, however, we should first check that

$$\begin{aligned} 0 &\leq \frac{\lambda_1 - d_{nn}}{2} < \lambda_n \text{ (case 1)} \\ \max\left\{\lambda_n, \frac{\lambda_1 - d_{nn}}{2}\right\} &< \frac{\lambda_1 + \lambda_n}{2} \text{ (case 2)} \end{aligned}$$

Since V is a thin and tall factor matrix, we have

$$\lambda_n(VV^*) = 0$$

therefore, we can choose

$$s = \frac{\lambda_1}{2}$$

to ensure the convergence of the algorithm

D.5 Upper bound for iteration steps

Assume ε is the precision we want and k the interation steps required by Diagonal Iteartion. From the convergence analysis, we have

$$\varepsilon = \|x_k - x^*\| \leq \left\|D^{-1}VV^*\right\|^k \|x_0 - x^*\|$$

Let $L = \|x_0 - x^*\|$. If we choose x_0 be the central point of the bound l and u. Then L is determined by the magnitude of the bound l and u.

Take logarithm of both sides of the inequality, we have

$$\log \varepsilon \leq k \log(\left\|D^{-1}VV^*\right\|) + \log L$$

then we can have a upper bound for the iteration steps k

$$k \leq \frac{\log \varepsilon - \log L}{\log(\|D^{-1}VV^*\|)}$$

The number of iteration steps needed is controled by the precision ε, the norm $\left\|D^{-1}VV^*\right\|$ and L, which is determined by the lower and upper bound.

References

[1] Ryan Takahashi. "Structured Matrices and the Algebra of Displacement Operators." *Harvey Mudd College, 2013.*

[2] V. Strassen. "Gaussian elimination is not optimal." *Numerishe Mathematik, 13:354-356, 1969.*

[3] D. Coppersmith and S. Winograd. "Matrix multiplication via arithmetic progressions." *Journal of Symbolic Computations, 9:251-280, 1990.*

[4] P. Burgisser, M. Clausen, and M.A. Shokrollahi. "Algebraia complexity theory." *Springer-Verlag, 1997.*

[5] V. Pan. "How to multiply matrices faster." *Lecture notes in computer science, volume 79. Springer-Verlag, 1985.*

[6] Victor Y. Pan. "On Parallel Computations with Banded Matrices." *Information and computation 120. 237-250, 1995.*

[7] H. Cohn and C. Umans. "A group-theoretic approach to fast matrix multiplication." *In Proc. of 44th FOCS, pages 438-449, 2003.*

[8] S. G. Mallat, Z. Zhang. "Matching Pursuits with Time-Frequency Dictionaries." *IEEE Transactions on Signal Processing, December 1993, pp. 3397–3415.*

[9] F. Bergeaud, S. Mallat. "Matching pursuit of images." *In Proc. International Conference on Image Processing, volume 1, pages 53–56 vol.1, 1995.*

[10] T. Tony Cai, Lei Wang. "Orthogonal Matching Pursuit for Sparse Signal Recovery." *IEEE Transactions On Information Theory, Vol. 57, No.7, July 2011.*

[11] Robert Tibshirani. "The LASSO method for variable selection in the cox model." *Statistics in medicine, Vol.16, 385-395, 1997.*

[12] Gilbert Strang. "Banded Matrices with Banded Inverses and $A = LPU$." *Proceedings Intl. Congress of Chinese Mathematicians: ICCM2010, Amer. Math. Soc. and Intl. Press, 2012.*

[13] Gregory Ammar, William Gragg, Lothar Reichel. "Constructing a Unitary Hessenberg Matrix from Spectral Data." *Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, NATO ASI Series Volume 70, 1991, pp 385-395.*

[14] G. S. Ammar, W. B. Gragg, L. Reichel. "Direct and Inverse Unitary Eigenproblems in Signal Processing: An Overview." *Linear Algebra for Large Scale and Real-Time Applications, NATO ASI Series Volume 232, 1993, pp 341-343.*

[15] Alfredo Remon, Enrique S. Quintana-Orti, Gregorio Quintana-Orti. "Cholesky Factorization of Band Matrices Using Multithreaded BLAS." *Applied Parallel Computing. State of the Art in Scientific Computing, Lecture Notes in Computer Science Volume 4699, 2007, pp 608-616.*

[16] R. S. Martin, J. H. Wilkinson. "Symmetric decomposition of positive definite band matrices." *Numerische Mathematik 1965, Volume 7, Issue 5, pp 355-361.*

[17] M. Lovric, M. Min-Oo, E. A. Ruh. "Multivariate normal distributions parametrized as a Riemannian symmetric space." *Journal of Multivariate Analysis, 74:36-48, 2000.*

[18] L. T. Skovgaard. "A Riemannian geometry of the multivariate normal model." *Scand. J. Statist. 11:211-223, 1984.*

[19] S.I.R. Costa, S.A. Santos, J.E.Straasson. "Fisher information distance: a geometrical reading." *CoRR, 2012.*

[20] C.K. Chui, G. Chen. "Discrete H^{∞} Optimization: with Applications in Signal Processing and Control Systems." *Information Science. ISBN: 978-3-540-61959-8.*

[21] Suliman Al-Homidan. "Solving Hankel matrix approximation problem using semidefinite programming." *Journal of Computational and Applied Mathematics, Volume 202, Issue 2, 15 May 2007, Pages 304–314.*

[22] Andrew W. Lo and Pankaj N. Patel. "130/30: The New Long-Only." Dec 11, 2007.

[23] H. Park, L. Zhang, J. B. Rosen. "Low Rank Approximation of a Hankel Matrix by Structured Total Least Norm" *BIT Numerical Mathematics 1999, Volume 39, Issue 4, pp 757-779.*

[24] Arun Tangirala. "Principles of System Identification: Theory and Practice." *Taylor & Francis, 2014.*

[25] Patrick Dewilde, Alle-Jan van der Veen. "Time-Varying Systems and Computations." Springer Science & Business Media, Jun 30, 1998.

[26] Anderson, T. W. "Statistical inference for covariance matrices with linear structure." *Multivariate Analysis, II (Proc. Second Internet. Sympos., Dayton, Ohio, 1968). 1969.*

[27] Mullhaupt, Andrew P., and Kurt S. Riedel. "Low grade matrices and matrix fraction representations." *Linear algebra and its applications 342.1 (2002): 187-201.*

[28] Mullhaupt, Andrew P. and Kurt S. Riedel. "Banded matrix fraction representation of triangular input normal pairs." *Automatic Control, IEEE Transactions on 46.12 (2001): 2018-2022.*

[29] Tengjie Jia, Andrew Mullhaupt. "Band Fraction Representation Estimation of Covariance Matrix"

[30] Chandrasekaran, Shiv, et al. "A superfast algorithm for Toeplitz systems of linear equations." *SIAM Journal on Matrix Analysis and Applications 29.4(2007): 1247-1266.*

[31] Chandrasekaran, Shiv, and Ming Gu. "Fast and stable algorithms for banded plus semiseparable systems of linear equations." *SIM Journal on Matrix Analysis and Applications 25.2 (2003): 373-384.*

[32] Chandrasekaran, Shiv, and Ming Gu. "A divide-and-conquer algorithm for the eigendecomposition of symmetric block-diagonal plus semiseparable matrices." *Numerische Mathematik 96.4 (2004): 723-731.*

[33] Chandrasekaran, Shiv, et al. "Some fast algorithms for sequentially semiseparable representations." *SIAM Journal on Matrix Analysis and Applications 27.2 (2005): 341-364.*

[34] Chandrasekaran, Shiv, M. Gu, and W. Lyons. "A fast and stable adaptive solver for hierarchically semi-separable representations." *Technical Report UCSB Math 2004-20, UC Santa Barbara, 2004.*

[35] Chandrasekaran, Shiv, Ming Gu, and T.Pals. "A fast ULV decomposition solver for hierarchically semiseparable representations." *SIAM Journal on Matrix Analysis and Applications 28.3 (2006): 603-622.*

[36] Christensen, Lars PB. "An Em-algorithm for band-teoplitz covariance matrix estimation." *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE*

[37] Ming Gu, Xiaoye S. Li, and Panayot S. Vassilevski. "Direction-Preserving and Schur-monotonic semiseparable approximations of symmetric positive definite matrices." *SIAM J. Matrix Anal. Appl. Vol. 31 No.5, pp. 2650-2664.*

[38] Shengguo Li, Ming Gu, et al. "New efficient and robust HSS Cholesky factorization of SPD matrices." *SIAM Journal on Matrix Analysis and Applications 07/2012; 33(3).*

[39] Jianlin Xia, Yuanzhe Xi, Ming Gu. "A superfast structured solver for Toeplitz linear systems via randomized sampling." *SIAM Journal on Matrix Analysis and Applications 07/2012; 33(3).*

[40] Ledoit, Olivier, and Michael Wolf. "A well-conditioned estimator for large dimensional covariance matrices." *Journal of multivariate analysis 88.2 (2004): 365-411.*

[41] Ledoit, Olivier, and Michael Wolf. "Nonlinear shrinkage estimation of large dimensional covariance matrices." *The Annals of Statistics 40.2 (2012): 1024-1060.*

[42] Jianlin Xia, et al. "Superfast multifrontal method for large structured linear systems of equations." *SIAM J. Matrix Analysis Applications. 01/2009; 31:1382-1411.*

[43] Colin Atkinson and Ann F. S. Mitchell. "Rao's Distance Measure". *Sankhya: The India Journal of Statistics. 1981, Volume 43, Series A, Pt. 3, pp. 345-365.*

[44] Rao, C. R. "Information and the accuracy attainable in the estimation of statistical parameters." *Bull. Calcutta math. Soc., 37, 81-91.*

[45] Rao, C. R. "Minimum variance and the estimation of several parameter." *Proc. Cambridge Philos. Soc., 43, 280-283.*

[46] Rao, C. R. "Sufficient statistics and minimum variance estimation." *Proc. Cambridge Philos. Soc., 45, 215-218.*

[47] Rao, C. R. "Asymptotic efficiency and limiting information." *Proceeding of 4th Berkeley Symposium on Mathematical Statistical and Probability, Vol. 1., University of California Press, pp. 531-546.*

[48] Rao, C. R. "Criteria of estimation in large samples." *Sankhya, 25, 189-206.*

[49] Stephen Boyd and Lieven Vandenberghe. "Convex Optimization." *Cambridge University Press.*

[50] Dempster, A., N.Laird, et al. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological) 39(1): 1-38.*

[51] Eaton, M.L, "Multivariate statistics: a vector space approach." *John Wiley and Sons. pp. 116-117. ISBN 0-471-02776-6(1983).*

[52] Fama, E. F. and French, K. R. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics, 33, 3C56(1993).*

[53] Maronna, R. "Robust M-estimators of multivariate location and scatter." *The annals of statistics: 51-67.*

[54] Rubin, D. and D. Thayer. "EM algorithms for ML factor analysis." *Psychometrika 47(1): 69-76.*

[55] Wilcox, R. "Introduction to robust estimation and hypothesis testing." *Academic Press.*

[56] Zha, J. H, Philip, L., and Jiang, Q. "ML estimation for factor analysis: EM or non-EM?" *Statistics and Computing, 18(2), 109-123.*

[57] Anderson, T. W. "Asymptotically efficient estimation of covariance matrices with linear structure." *The Annals of Statistics 1.1 (1973): 135-141.*

[58] Bai, Jushan and Shuzhong Shi. "Estimating high dimensional covariance matrices and its applications." (2001).

[59] Barnard, John, Robert McCulloch, and Xiao-Li Meng. "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage." *Statistica Sinica 10.4 (2000): 1281-1312.*

[60] Beran, Rudolf. "Minimum Hellinger distance estimates for parametric models." *The Annals of Statistics 5.3 (1977): 445-463.*

[61] Bickel, Peter J., and Yulia R. Gel. "Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series." *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.5 (2011): 711-728.*

[62] Bilmes, Jeff A. "Factored sparse inverse covariance matrices." *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. Vol. 2. IEEE, 2000.*

[63] Burg, John Parker, David G. Luenberger, and Daniel L. Wenger. "Estimation of structured covariance matrices." *Proceedings of the IEEE 70.9 (1982): 963-974.*

[64] Chang, Changgee, and Ruey S. Tsay. "Estimation of covariance matrix via sparse Cholesky factor with lasso." *Journal of Statistical Planning and Inference 140.12 (2010): 3858-3873.*

[65] Chaudhuri, Sanjay, Mathias Drton, and Thomas S. Richardson. "Estimation of a covariance matrix with zeros." *Biometrika 94.1 (2007): 199-216.*

[66] Chen, Yilun, Ami Wiesel, and Alfred O. Hero. "Shrinkage estimation of high dimensional covariance matrices." *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009.*

[67] Chiu, Tom YM, Tom Leonard, and Kam-Wah Tsui. "The matrix-logarithmic covariance model." *Journal of the American Statistical Association 91.433 (1996): 198-210.*

[68] Chu, Moody T., Robert E. Funderlic, and Robert J. Plemmons. "Structured low rank approximation." *Linear algebra and its applications 366 (2003): 157-172.*

[69] Daniels, Michael J., and Robert E. Kass. "Shrinkage estimators for covariance matrices." *Biometrics 57.4 (2001): 1173-1184.*

[70] Delvaus, Steven, and Marc Van Barel. "Structures preserved by Schur complementation." *SIAM journal on matrix analysis and applications 28.1 (2006): 229-252.*

[71] Delvaux, Steven and Marc Van Barel. "A Givens-weight representation for rank structured matrices." *SIAM Journal on Matrix Analysis and Applications 29.4 (2007): 1147-1170.*

[72] Dembo, Amir, Colin L. Mallows, and Lawrence A. Shepp. "Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation." *Information Theory, IEEE Transactions on 35.6 (1989): 1206-1212.*

[73] Deng, Xinwei, and Ming Yuan. "Large Gaussian covariance matrix estimation with Markov structures." *Journal of Computational and Graphical Statistics 18.3 (2009).*

[74] Dey, Dipak K. and C. Srinivasan. "Estimation of a covariance matrix under Stein's loss." *The Annals of Statistics (1985): 1581-1591.*

[75] Eidelman, Y., and I. Gohberg. "Fast inversion algorithms for diagonal plus semiseparable matrices." *Integral Equations and Operator Theory 27.2 (1997): 165-183.*

[76] Fan, Jianqing, Yingying Fan, and Jinchi Lv. "High dimensional covariance matrix estimation using a factor model." *Journal of Econometrics 147.1 (2008): 186-197.*

[77] Fasino, Dario, Nicola Mastronardi, and Marc Van Barel. "Fast and stable algorithms for reducing diagonal plus semiseparable matrices to tridiagonal and bidiagonal form." *Contemporary Mathematics 323 (2003): 105-118.*

[78] Frieze, Alan, Ravi Kannan, and Santosh Vempala. "Fast Monte-Carlo algorihms for finding low-rank approximations." Journal of the ACM (JACM) 51.6 (2004): 1025-1041.

[79] Friedman, Jerome, trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics 9.3 (2008): 432-441.*

[80] Gohberg, I., T. Kailath, and I. Koltracht. "Linear complexity algorithms for semiseparable matrices." *Integral Equations and Operator Theory 8.6 (1985): 780-804.*

[81] Huang, Jianhua Z., et al. "Covariance matrix selection and estimation via penalised normal likelihood." *Biometrika 93.1 (2006): 85-98.*

[82] Jansson, Magnus, and Bjorn Ottersten. "Structured covariance matrix estimation: a parametric approach." *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. Vol. 5. IEEE, 2000.*

[83] Karlis, Dimitris. "An EM type algorithm for maximum likelihood estimation of the normal-inverse Gaussian distribution." *Statistics & probability letters 57.1 (2002): 43-52.*

[84] Lawley, Derrick N. "The estimation of factor loadings by the method of maximum likelihood." *Proceedings of the Royal Society of Edinburgh 60.2 (1940): 64-82.*

[85] Markovsky, Ivan. "Structured low-rank approximation and its applications." *Automatica 44.4 (2008): 891-909.*

[86] Mastronardi, Nicola, Shivkumar Chandrasekaran, and Sabine Van Huffel. "Fast and stable two-way algorithm for diagonal plus semi-separable systems of linear equations." *Numerical linear algebra with applications 8.1 (2001): 7-12.*

[87] Miller, Michael I., and Donald L. Snyder. "The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariance." *Proceedings of the IEEE 75.7 (1987): 892-907.*

[88] Pourahmadi, Mohsen. "Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix." *Biometrika 87.2 (2000): 425-435.*

[89] Pourahmadi, Mohsen. "Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters." *Biometrika 94.4 (2007): 1006-1013.*

[90] Rajaratnam, Bala, Helene Massam, and Carlos M. Carvalho. "Flexible covariance estimation in graphical Gaussian models." *The Annals of Statistics 36.6 (2008): 2818-2849.*

[91] Rothman, Adam J., Elizaveta Levina, and Ji Zhu. "A new approach to Cholesky-based covariance regularization in high dimensions." *Biometrika 97.3 (2010): 539-550.*

[92] Rubin, Donal B., and Dorothy T. Thayer. "EM algorithms for ML factor analysis." *Psychometrika 47.1 (1982): 69-76.*

[93] Rubin, Donald B., and Ted H. Szatrowski. "Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm." *Biometrika 69.3 (1982): 657-660.*

[94] Simpson, Douglas G. "Minimum Hellinger distance estimation for the analysis of count data." *Journal of the American statistical Association 82.399 (1987): 802-807.*

[95] Tamura, Roy N., and Dennis D. Boss. "Minimum Hellinger distance estimation for multivariate location and covariance." *Journal of the American Statistical Association 81.393 (1986): 223-229.*

[96] Van der Veen, A. J. "Approximate inversion of a large semiseparable positive matrix." *Proc. 17th Int. Symp. on Mathematical Theory of Networks and Systems (MTNS-04), Brussels (BE). 2004.*

[97] Williams, Douglas B., and Don H. Johnson. "Robust estimation of structured covariance matrices." *Signal Processing, IEEE Transactions on 41.9 (1993): 2891-2906.*

[98] Wirfalt, P., and Magnus Jansson. "On Toeplitz and Kronecker structured covariance matrix estimation." *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2010 IEEE.*

[99] R.A.Robert, C.T.Mullis. "Digital Signal Processing." *Addison Wesley, MA, 1987.*

[100] T. Georgiou and A. Lindguist. "Kullback-Leibler approximation of spectral density functions," *IEEE trans.inf.Theory, vol 49, no.11, pp.2910-2917, Nov, 2003.*

[101] Burg, J,P. "Maximum entropy spectral analysis." *37th Annual international Meeting, Soc. of Explor. Geophys., Oklahoma City, Okla., Oct. 31, 1967.*

[102] Lacoss, R.T. "Data adaptive spectral analysis methods." *Geophysics, 36, 661-675, 1971.*

[103] Wiener, N. "Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications." *John Wiley and Sons, Inc., New York, 1950.*

[104] Ulrych, T.J. "Maximum entropy power spectrum of truncated sinusoids." *J. Geophys. Res., 77, 1396-1400, 1972.*

[105] Erich L. Lehmann, Joseph P. Romano. "Testing Statistical Hypotheses." *Springer, 2005.*

[106] U. Grenander and G. Szegö. "Toeplitz Forms and Their Applications." *University of Calif. Press, Berkeley and Los Angeles, 1958.*

[107] S. Kullback, R.A. Leibler. "On information and Sufficiency." *The Annals of Mathematical Statistics, Vol.22, No.1, 79-86, Mar.,1951.*

[108] S. Kullback. "Information theory and statistics." *John Wiley and Sons, NY, 1959.*

[109] S. Kullback. "Letter to the Editor: The Kullback-Leibler distance." *The American Statistician, Vol 44, No.4, 340-341, 1987.*

[110] Christopher I. Byrnes, Tryphon T. Georgiou, Anders Lindquist. "A new approach to Spectral Estimation: A Tunable High-Resolution Spectral Estimator." *IEEE Transactions on Signal Processing, Vol.48, No.11, November 2000.*

[111] Hsien Liang. "A Grenander and Szegö Limit Theorem for Toeplitz Operations on Locally Compact Abelian Groups." *Rocky Mountain Journal of Mathematics, Vol.22, No.4, Fall 1992.*

[112] Augusto Ferrante, Michele Pavon, Federico Ramponi. "Hellinger Versus Kullback-Leibler Multivariable Spectrum Approximation." *IEEE Transactions on automatic control, Vol 53, No.4, May 2008.*

[113] John Parker Burg, David G. Luenberger, Daniel L. Wenger. "Estimation of Structured Covariance Matrices." *Proceeding of The IEEE, Vol.70, No.9, September 1982.*

[114] Tad J. Ulrych, Thomas N. Bishop. "Maximum Entropy Spectral Analysis and Autoregressive Decomposition." *Reviews of Geophysics and Space Physics, Vol.13, No.1 February 1975.*

[115] N. Andersen. "On The Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis." *Geophysics, Vol.39, No.1, P.69-72, February 1974.*

[116] P.A.Regalia. "Adaptive IIR Filtering in Signal Processing and Control." *Marcel Dekker, 1995.*

[117] H.Dym. "Linear Algebra in Action." *American Mathematical Soc., 2007.*

[118] F.R. Gantmacher. "The Theory of Matrices." *Chelsea Publishing Co., NY 1960.*

[119] R. M. Gray. "On the asymptotic eigenvalue distribution of Toeplitz matrices." *IEEE Transactions on Information Theory, Vol. 18, November 1972, pp.725-730.*

[120] R. M. Gray. "On Unbounded Toeplitz Matrices and Nonstationary Time Series with an Application to Information Theory." *Information and Control, 24, pp. 181-196, 1971.*

[121] U. Grenander and G. Szegö. "Toeplitz Forms and Their Applications." *University of Calif. Press, Berkeley and Los Angeles, 1958.*

[122] P. Lancaster. "Theory of Matrices." *Academic Press, NY, 1969.*

[123] J. Pearl. "On Coding and Filtering Stationary Signals by Discrete Fourier Transform." *IEEE Trans. On Info. Theory, IT-198, pp.229-232, 1973.*

[124] W.F. Trench. "Absolute equal distribution of the spectra of Hermitian matrices." *Lin. Alg. Appl., 366 (2003), 417-431.*

[125] C. I. Byrnes and A. Lindquist. "The generalized moment problem with complexity constraint." *Integral Equations Operator Theory, Vol. 56, No. 2, pp.163-180, 2006.*

[126] A. Ferrante, M. Pavon, F. Ramponi. "Constrained Approximation in the Hellinger Distance." *Proc. Eurp. Control Conf. (ECC 2007), Kos, Greece, Jul., pp. 322-327.*

[127] T. Georgiou. "Realization of Power Spectra from Partial Covariance Sequences." *IEEE Trans. Acoust., Speech, Signal Process., Vol. ASSP-35, No.4, pp.438-449, Apr. 1987.*

[128] T. Georgiou. "Spectral analysis based on the state covariance: the maximum entropy spectrum and linear fractional parameterization." *IEEE Trans. Autom. Control, Vol. 47, No. 11, pp. 1811-1823, Nov. 2002.*

[129] T. T. Georgiou. "Distances and Riemannian metrics for spectral density functions." *IEEE Trans. Signal Process., Vol. 55, No. 8, pp. 3995-4003, Aug. 2007.*

[130] T. T. Georgiou. "An intrinsic metric for power spectral density functions." *IEEE Signal Process. Lett., Vol. 14, No. 8, pp. 561-563, Aug. 2007.*

[131] C. I. Byrnes, S. Gusev, A. Lindquist. "A convex optimization approach to the rational covariance extension problem." *SIAM J. Control and Optimization, 37:211-299, 1999.*

[132] T. Georgiou. "Relative entropy and the multivariable multidimensional moment problem." *IEEE Trans. Inform. Theory, 52:1052-1066, 2006.*

[133] S. Kullback. "Information Theory and Statistics." *2nd ed., Dover, Mineola NY, 1968.*

[134] M. Pavon, A. Ferrante. "On the Georgiou-Lindquist approach to constrained Kullback-Leibler approximation of spectral densities." *IEEE Trans. Aut. Control, 51:639-644, 2006.*

[135] P. Stoica, R. Moses. "Introduction to Spectral Analysis." *Prentice Hall, New York, 1997.*

[136] V. Vedral. "The role of relative entropy in quantum information theory." *Rev. Mod. Phys, 74:197-213, 2002.*