Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Linear Algebra and its Applications 364 (2003) 161–187
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

# Optimal low-rank approximation to a correlation matrix

Zhenyue Zhang [a], Lixin Wu [b,*]

[a] *Department of Mathematics, Zhejiang University, Yu-Quan Campus, China*
[b] *Department of Mathematics, University of Science and Technology, Hong Kong*

Received 19 October 2001; accepted 22 August 2002

Submitted by R.A. Brualdi

**Abstract**

Low-rank approximation of a correlation matrix is a constrained minimization problem that can be translated into a minimization–maximization problem by the method of Lagrange multiplier. In this paper, we solve the inner maximization problems with a single spectral decomposition, and the outer minimization problems with gradient-based descending methods. An in-depth analysis is done to characterize the solutions of the inner maximization problem for the case when they are non-unique. The well-posedness of the Lagrange multiplier problem and the convergence of the descending methods are rigorously justified. Numerical results are presented.


*Keywords:* Low-rank matrix approximation; Constrained minimization; Lagrange method; Matrix spectral decomposition and the method of steepest descend

## 1. Introduction

One of the major sources of low-rank approximation problem is image processing, where for storage and transmission purpose, images or data must be compressed in an efficient way. Mathematically, data compression can be interpreted into a low-rank matrix approximation problem which can be achieved by singular value decomposition (SVD). In the recent advance of financial engineering, risk managers

* Corresponding author. Present address: Department of Mathematics, Claremont Graduate University, 121 East Tenth Street, Claremont, CA 91711-6105, USA.
*E-mail address:* malwu@ust.hk (L. Wu).

0024-3795/03/$ - see front matter 
doi:10.1016/S0024-3795(02)00551-7

often need to seek for the low-rank approximation of correlation matrices, which is an approximation problem under constraints: the approximate matrices must be semi-positive definite with all diagonal entries equal to 1 (We call it "one-diagonal" condition for later reference). Due to the existence of constraints, we have to go beyond the SVD technique for the approximation, and this is the issue of the current paper.

The problem of approximating correlation matrices arose from the application of the so-called *market model* of interest rates [2,6]. In the market model, the state variables are the forward term rates of interest, which are assumed to follow the log-normal stochastic processes. The model is then used to price interest-rate derivative instruments. The forward term rates are observable variables in the market place. With historical data we are able to extract their correlation structure. Desirably, we want to implant the correlation structure into the stochastic processes for the forward rates, so that the model can appropriately describe the dynamics of forward rate evolutions, and, consequently, make the pricing more accurate. It is obvious that the rank of the model correlation matrix does not exceed the number of random factors the model takes. When the number of random factors is smaller than the rank of the historical correlation matrix, which is almost always the case, the model correlation matrix cannot be made identical to the historical one. In such circumstance, we will instead try to achieve the best approximation to the historical correlation matrix under, in particular, the Frobenius matrix norm. This gives rise to the problem of optimal low-rank approximation of a correlation matrix. Recently, Rebonato [8] proposed an elegant parameterization technique to ensure the one-diagonal condition, before getting into a brute force optimization procedure for the approximate matrices. The parameterization removes the constraints. The subsequent unconstraint optimization, however, has as many unknowns as the number of elements of a given correlation matrix, which can be very big and thus can be very time consuming to solve. In this paper, we take the approach of Lagrange multiplier (see [5], for example). This approach allows us to use spectral decomposition, iteratively, to solve the approximation problem. Numerically, our approach is both robust and efficient. Moreover, without any assumption, we can rigorously justify the convergence of the Lagrange multiplier method for the low-rank approximation of correlation matrices.

This paper is organized as follows. In Section 2, we will offer a simplified exposition of the market model for interest rate modeling, where we will describe the origin of the approximation problem for correlation matrices. In Section 3, we will describe the Lagrange multiplier methodology for the matrix approximation problem, impose extra regularization conditions for the case when the solutions are non-unique, and characterize the solutions. In Section 4, we will show that solution(s) produced by Lagrange multiplier method indeed solve the original constrained minimization problem, and prove the convergence of a gradient-based descending search algorithm. In Section 5 we present computational results with both hypothetical and practical correlation matrices. Finally in Section 6 we conclude the paper.

*Notation*. For a square matrix $A$, we denote by $\mathrm{diag}(A)$ the column vector with the diagonal entries of $A$. Conversely, if $d$ is a (column) vector, we define $\mathrm{diag}(d)$ the diagonal matrix with the diagonal entries as the components of $d$. We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix and $\|\cdot\|_2$ for the spectrum norm of a matrix or the 2-norm of a vector.

## 2. Background of the approximation problem in finance

In the interest-rate derivative market of London Inter-Bank Offer Rates (LIBOR), financial professionals use mathematical models to price and hedge a variety of derivative products. Among these models, the market model is perhaps the most popular one. With an interest-rate model, people try to describe the probability distribution of future interest rates. In the market model, the forward term rates of interest are taken as the state variables, and are assumed to follow lognormal stochastic processes. Under the log-normal forward rate processes, liquid interest-rate derivatives like caps, floors and swaptions (see [8] for description) are priced with the famous Black's formula [1]. The existence of the closed-form formula facilitates efficient determination, so-called *calibration* in the market place, of model parameters. We refer readers to [1] for the pricing theory of commodity options. For a comprehensive discussion of the market model, we refer readers to [2,6].

The market model was built upon the lognormal dynamics of forward LIBOR rates. Let $f_j(t) = f(t; T_j, T_{j+1})$ be the arbitrage-free forward lending rate seen at time $t$ for the period $(T_j, T_{j+1})$ in the future. The rate $f_j(t)$ is assumed to follow the so-called lognormal process:

$$\mathrm{d}f_j(t) = f_j(t)(\gamma_j(t))^{\mathrm{T}}[\sigma_{j+1}(t)\,\mathrm{d}t + \mathrm{d}\mathbf{Z}(t)], \tag{1}$$

where $\mathbf{Z}(t)$ is the vector of $n_t$-dimensional independent Brownian motions, $\gamma_j(t)$ is the vector of the instantaneous volatility coefficients for forward rate, and $\sigma_{j+1}(t)$ is the vector of instantaneous volatility coefficients of zero-coupon bond of maturity $T_{j+1}$. We consider a collection of $N$ forward rates, $f_j$, $j = 1, 2, \ldots, N$. Due to the strong correlation amongst the forward rates, financial engineers do not use as many random factors as the number of forward rates. Instead, the number of factors $n_t$ is often taken to be three or four, which is typically much smaller than $N$, the number of forward rates. The volatility of the forward rates and the volatility of the zero-coupon bond are not independent. The no-arbitrage condition [2] gives rise to the relation between the two volatilities:

$$\sigma_j(t) = \sum_{k=0}^{j-1} \frac{\Delta T_k f_k(t)}{1 + \Delta T_k f_k(t)} \gamma_k(t), \tag{2}$$

where $\Delta T_j = T_{j+1} - T_j$ and $\gamma_j(t) = 0$ for $t \geqslant T_j$. As the convention we label the time of today by $t = T_0 = 0$. The stochastic evolution of the $N$ forward rates is fully described by the quantities of correlation defined by

$$C_{jk}^i = \frac{\int_{T_{i-1}}^{T_i} \gamma_j(t) \cdot \gamma_k(t)\,\mathrm{d}t}{\sqrt{\int_{T_{i-1}}^{T_i} \|\gamma_j(t)\|_2^2\,\mathrm{d}t} \cdot \sqrt{\int_{T_{i-1}}^{T_i} \|\gamma_k(t)\|_2^2\,\mathrm{d}t}}, \quad 1 \leqslant i \leqslant N. \tag{3}$$

Note that $C_{jk}^i = 0$ for either $j < i$ or $k < i$ since either $f_j$ or $f_k$ has been known by the time $T_i$. Hence for fixed $i$, $C_{jk}^i$ constitutes the elements of an $(N - i + 1)$ by $(N - i + 1)$ non-negative symmetric matrix:

$$\mathbf{C}^i = \begin{pmatrix} C_{i,i}^i & C_{i,i+1}^i & \cdots & C_{i,N}^i \\ C_{i+1,i}^i & C_{i+1,i+1}^i & \cdots & C_{i,N}^i \\ \vdots & \vdots & \ddots & \vdots \\ C_{N,i}^i & C_{N,i+1}^i & \cdots & C_{N,N}^i \end{pmatrix}, \quad i = 1, 2, \ldots, N. \tag{4}$$

Under the lognormal forward rate processes, one can derive Black's formula [2] for caplets. A caplet of maturity $T_j$ is an interest-rate contract which, for some pre-specified interest-rate level $K$, delivers at time $T_{j+1}$ a cash flow in the amount of $L\Delta T_j(f_j(T_j) - K)$ if the term rate $f_j(T_j)$ is above $K$, or zero if otherwise. Here $L$ is called the notional amount of the contract. The pricing formula of the caplet is

$$C_{\text{let}} = L\Delta T_j[f_j(t)N(d_1) - KN(d_2)], \tag{5}$$

where $N(\cdot)$ is the normal accumulative function,

$$d_1 = \frac{\ln \frac{f_j(t)}{K} + \frac{1}{2}\zeta_j^2 T_j}{\zeta_j\sqrt{T_j}}, \quad d_2 = d_1 - \zeta_j, \tag{6}$$

and

$$\zeta_j^2 = \frac{1}{T_j}\int_0^{T_j} \|\gamma_j(t)\|_2^2\,\mathrm{d}t. \tag{7}$$

Quantity $\zeta_j$ is called the caplet volatility. Black's formula establishes a one-to-one correspondence between the caplet price and the caplet volatility.

In reality, the prices of liquid instruments like caplets are determined by supply-and-demand. Other more complex instruments, for example, Bermudian options and cancelation swaps (see for example, [8]), are priced in a way consistent with the prices of the liquid instruments. A major financial engineering problem in the market place is to find $\gamma_j(t)$ subject to a given set of caplet prices, or equivalently, a set of caplet volatilities $\{\zeta_j\}$, as well as the historical correlation matrices $\{\mathbf{C}^i\}$ of the forward rates involved.

In the so-called non-parametric approach, the problem boils down to the determination of the piece-wise constant vector function of volatilities

$$\gamma_j(t) = \gamma_j^i = s_j\left(a_{j,1}^i, a_{j,2}^i, \ldots, a_{j,n_t}^i\right) \equiv s_j\mathbf{a}_j^i, \quad \text{for } T_{i-1} \leqslant t \leqslant T_i, \tag{8}$$

with

$$s_j = \|\gamma_j^i\|_2 = \zeta_j \quad \text{and} \quad \|\mathbf{a}_j^i\|_2 = 1. \tag{9}$$

Apparently, $s_j$ is uniquely determined by the caplet volatility $\zeta_j$. The forward rate volatility components $\mathbf{a}_j^i$, meanwhile, arise from the SVD [1] of the correlation matrix $\mathbf{C}^i$. To see this, suppose the rank of $\mathbf{C}^i$ is less than or equal to $n_t$. Perform SVD on $\mathbf{C}^i$:

$$\mathbf{C}^i = U\Lambda U^{\mathrm{T}}, \tag{10}$$

where $\Lambda$ is an $n_t$ by $n_t$ diagonal matrix with non-negative diagonal elements, and define $\mathbf{a}_j^i$ as the $j$th row of $U\Lambda^{1/2}$, i.e.,

$$\mathbf{C}^i = \begin{pmatrix} \mathbf{a}_i^i \\ \vdots \\ \mathbf{a}_N^i \end{pmatrix} \big((\mathbf{a}_i^i)^{\mathrm{T}}, \ldots, (\mathbf{a}_N^i)^{\mathrm{T}}\big). \tag{11}$$

Then the model correlation so obtained is

$$\begin{aligned} \mathrm{Corr}(\Delta f_j(t_i), \Delta f_k(t_i)) &= \frac{\Delta T \sum_{l=1}^{n_t} a_{l,j}^i a_{l,k}^i}{\sqrt{\Delta T \sum_{l=1}^{n_t} (a_{l,j}^i)^2} \cdot \sqrt{\Delta T \sum_{l=1}^{n_t} (a_{l,k}^i)^2}} \\ &= C_{jk}^i \end{aligned} \tag{12}$$

by property (11), where

$$\Delta f_j(t_i) = f_j(t_i) s_j (\mathbf{a}_j^i)^{\mathrm{T}} [\sigma_{j+1} \Delta t + \Delta Z(t)]. \tag{13}$$

For later reference, we call the columns of the matrix $U\Lambda^{1/2}$ *principle components* of the matrix $\mathbf{C}^i$.

The complication in the determination of $\mathbf{a}_j^i$, $j = i, \ldots, N$, is that the rank of $\mathbf{C}^i$ is in general equal to $N - i + 1$, the number of forward rate "alive", which is typically much bigger than $n_t$, the number of random factors being taken (see [7] for financial background). In such case the above procedure for calculating $\mathbf{a}_j^i$, $j = i, \ldots, N$, breaks down. Therefore, a preprocessing is in general required to reduce the ranks of the given correlation matrices. For a given correlation matrix $\mathbf{C}^i$ for the period, preprocessing is naturally formulated as the following minimization problem with constraints:

$$\begin{aligned} &\min_{\hat{\mathbf{C}}^i} \|\mathbf{C}^i - \hat{\mathbf{C}}^i\|_{\mathrm{F}}, \\ &\text{s.t.} \quad \hat{\mathbf{C}}^i \geqslant 0, \quad \mathrm{rank}(\hat{\mathbf{C}}^i) \leqslant n_t, \quad \hat{\mathbf{C}}_{kk}^i = 1, \\ &\qquad k = i, \ldots, N, \quad i = 1, \ldots, N. \end{aligned} \tag{14}$$

Here, $\hat{\mathbf{C}}^i \geqslant 0$ means the non-negative definiteness of $\hat{\mathbf{C}}^i$. In the following section, we consider the solution of such constraint minimization problem.

[1] It is called Principle Component Analysis (PCA) by financial professionals.

## 3. Optimal constrained lower-rank approximation

Given a real symmetric matrix $C$ of order $N$ and an integer $n < N$, we consider the following constrained minimization problem:

$$\begin{aligned} &\min_X \|C - X\|_{\mathrm{F}}, \\ &\text{s.t.} \quad \operatorname{rank}(X) \leqslant n, \quad \operatorname{diag}(X) = \operatorname{diag}(C). \end{aligned} \tag{15}$$

We denote the optimal solution by $C^*$, and the feasible set of solutions by $\mathscr{S}$:

$$\mathscr{S} = \left\{ X \in \mathscr{R}^{N \times N} \mid \operatorname{rank}(X) \leqslant n, \operatorname{diag}(X) = \operatorname{diag}(C) \right\}.$$

For the market model introduced in the last section, the optimal solution $C^*$ serves as a correlation matrix and hence must be non-negative definite. It thus appears that the feasible set of the optimal problem should instead be $\mathscr{S}^+$, the subset of $\mathscr{S}$ that consists of only positive semi-definite matrices. Adding explicitly such constraint will inevitably increase the difficulty of the problem. Fortunately, as we will see later, the solution $C^*$ to (15) will automatically be positive semi-definite, given $C$ a positive semi-definite matrix. Hence the explicit imposition of the extra constraint becomes unnecessary.

Our approach for solving the constrained optimal approximation problem begins with transforming it into an equivalently min–max problem by the method of Lagrange multiplier. Let $\mathscr{R}_n$ be the set of $N \times N$ matrices with rank less or equal to $n$. The Lagrange multiplier problem corresponding to (15) is defined as the min–max problem

$$\min_d \max_{X \in \mathscr{R}_n} L(X, d) \tag{16}$$

with the Lagrange function

$$L(X, d) = -\|C - X\|_{\mathrm{F}}^2 - 2d^{\mathrm{T}} \operatorname{diag}(C - X), \tag{17}$$

where $d$ is the vector of the multipliers. Note that $L(X, d)$ is linear in $d$ in the sense

$$L(X, td + (1 - t)\hat{d}) = tL(X, d) + (1 - t)L(X, \hat{d}). \tag{18}$$

We will justify later that problem (16) is equivalent to the original constrained problem (15).

Numerically the min–max problem (16) will be handled as a maximization problem

$$V(d) = \max_{X \in \mathscr{R}_n} L(X, d) \tag{19}$$

nested in another minimization problem

$$\min_d V(d). \tag{20}$$

It is a matter then to look for efficient methods separately for the maximization problem (19) and minimization problem (20). For the problem of low-rank approximation, it is crucial to observe that the Lagrange function can be written into

$$L(X, d) = -\|C + D - X\|_F^2 + \|d\|_2^2, \tag{21}$$

where $D$ is the diagonalised matrix of $d$: $D = \operatorname{diag}(d)$. Hence for fixed $d$, the maximizer to (19) can be obtained by the spectral decomposition of matrix $C + D$ (which is symmetric). Let

$$C + D = U\Lambda U^T \tag{22}$$

be the spectral decomposition with orthogonal matrix $U$ and eigenvalue matrix

$$\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N),$$

where, for our interest, the diagonal elements are put in non-increasing order in magnitude:

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_N|.$$

Note that both $U$ and $\Lambda$ depend on the multiplier vector $d$, so they will be denoted by $U(d)$ and $\Lambda(d)$ when we need to highlight the dependence. A solution to the problem (19), the best rank-$n$ approximation of $C + D$, is given by

$$C(d) \equiv C_n(d) = U_n \Lambda_n U_n^T, \tag{23}$$

where $U_n$ is the matrix consisting of the first $n$ columns of $U$, and $\Lambda_n = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is the principle submatrix of $\Lambda$ of degree $n$. For the objective function it then follows that

$$V(d) = -\sum_{j=n+1}^{N} \lambda_j^2 + \|d\|_2^2.$$

Obviously, we have the solution uniqueness only for $|\lambda_n| > |\lambda_{n+1}|$. When $|\lambda_n| = |\lambda_{n+1}|$, we need to add more constraints to narrow down a solution. The additional constraint we adopt to (19) is

$$\min_{X \in \mathscr{S}(d)} \|\operatorname{diag}(C - X)\|_2, \tag{24}$$

where $\mathscr{S}(d)$ is the set of optimal solutions for multipliers $d$. That is, we look for the minimizer(s) that best observe the original constraints in (15).

To figure out the structure of solutions to (19), (24), we consider, without loss of generality, the following case:

$$|\lambda_\alpha| > |\lambda_{\alpha+1},\ \lambda_{\alpha+1} = \cdots = \lambda_\beta|, \quad |\lambda_\beta| > |\lambda_{\beta+1}|$$

with integers $\alpha$ and $\beta$ such that $\alpha < n \leqslant \beta$. The set of solutions to (19) then are expressed as

$$\mathscr{S}(d) \equiv \left\{C_\alpha + \lambda_n Q_1 Y Y^T Q_1^T : Y^T Y = I_{n-\alpha}\right\}, \tag{25}$$

where $C_\alpha$ is defined as $C_n$ for $n=\alpha$, and $Q_1 = U(:, \alpha+1:\beta)$ is the submatrix consisting of the columns $\alpha+1, \dots, \beta$ of the unitary matrix $U$. Since for $X \in \mathscr{S}(d)$,

$$C - X = C - C_\alpha - \lambda_n Q_1 Y Y^{\mathrm{T}} Q_1^{\mathrm{T}}, \tag{26}$$

problem (24) is in fact a minimization problem over the Stiefel manifold

$$\mathscr{V} = \left\{ Y \in \mathscr{R}^{(\beta-\alpha)\times(n-\alpha)} : Y^{\mathrm{T}} Y = I_{n-\alpha} \right\}.$$

One can solve such problem using Newton's method or conjugate gradient method [3]. Nevertheless, solutions to (19), (24) may still be non-unique. The diagonal vectors $x = \mathrm{diag}(X)$ corresponding to the solutions $X$, however, can be shown to be unique under very weak conditions. Let us define by $\Gamma$ the set of the diagonal vectors of matrices $Q_1 Y Y^{\mathrm{T}} Q_1^{\mathrm{T}}$:

$$\Gamma = \{z = \mathrm{diag}(Q_1 Y Y^{\mathrm{T}} Q_1^{\mathrm{T}}) : Y \in \mathscr{V}\}.$$

It is trivial to verify that all the vectors in $\Gamma$ belong to the superplane

$$z(1) + z(2) + \cdots + z(N) = n - \alpha.$$

In the case when $\alpha = n-1$ and $\beta = n+1$, $\Gamma$ is a segment of an ellipsoid of lower dimension. To see that we let $Y = [\cos(\theta), \sin(\theta)]^{\mathrm{T}}$ and $Q_1 = [u, v]$. Then for any $z \in \Gamma$, we have

$$\begin{aligned} z(i) &= (u(i)\cos(\theta) + v(i)\sin(\theta))^2 &(27)\\ &= \frac{u(i)^2 + v(i)^2}{2} + \frac{u(i)^2 - v(i)^2}{2}\cos(2\theta) + u(i)v(i)\sin(2\theta). &(28)\end{aligned}$$

Denoting the coefficients by

$$p_0(i) = \frac{u(i)^2 + v(i)^2}{2}, \quad p_1(i) = \frac{u(i)^2 - v(i)^2}{2}, \quad p_2(i) = u(i)v(i), \tag{29}$$

which are independent on $Y \in \mathscr{V}$, we can write the vector $z \in \Gamma$ as

$$z = p_0 + p_1\cos(2\theta) + p_2\sin(2\theta).$$

Let

$$[p_1, p_2] = G_1 \varDelta W^{\mathrm{T}}$$

be the SVD of $[p_1, p_2]$ with diagonal matrix $\varDelta$ of order $r = \mathrm{rank}([p_1, p_2]) \leqslant 2$, and orthonormal matrices $G_1 \in \mathscr{R}^{N\times 2}$ and $W \in \mathscr{R}^{2\times r}$. We rewrite $z \in \Gamma$ into

$$z = p_0 + G_1 \varDelta y \tag{30}$$

with $y = W^{\mathrm{T}}[\cos(2\theta), \sin(2\theta)]^{\mathrm{T}}$. Clearly, when $r = 2$, $y$ is a 2-dimensional unit vector and $\Gamma$ is an ellipsoid such that

$$(z - p_0)^{\mathrm{T}} S (z - p_0) = 1$$

for a rank-2 positive semi-definite matrix

$$S = G_1 \Delta^{-2} G_1^{\mathrm{T}}.$$

For the case when $r = 1$, $y$ is a bounded scalar variable, $|y| \leqslant 1$. Hence $\Gamma$ is a line segment, a degenerate ellipsoid.

The following theorem shows that, when $\lambda_n = \lambda_{n+1}$, the simplest case of multiple eigenvalues, optimal solutions to (19), (24) have the same diagonal elements under very mild conditions.

**Theorem 3.1.** *Assume* $\alpha = n - 1$ *and* $\beta = n + 1$. *Denote* $f = G_1^{\mathrm{T}}(\mathrm{diag}(C - C_\alpha) - p_0)$. *If*

(1) $\mathrm{rank}([p_1, p_2]) = 1$, *or*
(2) $\mathrm{rank}([p_1, p_2]) = 2$ *and* $f^{\mathrm{T}} \Delta^{-2} f > 1$, *or*
(3) $\mathrm{rank}([p_1, p_2]) = 2$, $f^{\mathrm{T}} \Delta^{-2} f \leqslant 1$, *and* $f(2) \neq 0$,

*then for any two solutions* $X_i$ $(i = 1, 2)$ *of* (19) *and* (24), *we have*

$$\mathrm{diag}(X_1) = \mathrm{diag}(X_2).$$

**Proof.** Recalling (24), (25) and (26), we consider the minimization problem

$$\min \left\{ \|\mathrm{diag}(C) - x\|_2 : \ x \in \mathscr{D}_{\mathscr{S}} \right\}, \tag{31}$$

where

$$\mathscr{D}_{\mathscr{S}} = \left\{ x = \mathrm{diag}(X) : X \in \mathscr{S}(d) \right\},$$

i.e., the set of diagonal vectors of the matrices in $\mathscr{S}(d)$. It is clear that if $\mathrm{rank}([p_1, p_2]) = 1$, $\mathscr{D}_{\mathscr{S}}$ is a line segment and the minimizer of the problem (31) is unique.

When $\mathrm{rank}([p_1, p_2]) = 2$, we rewrite $\mathscr{D}_{\mathscr{S}}$ as

$$\mathscr{D}_{\mathscr{S}} = \left\{ x = \mathrm{diag}(C_\alpha) + p_0 + G_1 \Delta y : y^{\mathrm{T}} y = 1 \right\}.$$

Denoting

$$h = \mathrm{diag}(C - C_\alpha) - p_0,$$

we have that for $x \in \mathscr{D}_{\mathscr{S}}$,

$$\|x - \mathrm{diag}(C)\|_2^2 = \|h - G_1 \Delta y\|_2^2 = \left\| G_1^{\mathrm{T}} h - \Delta y \right\|_2^2 + \left\| G_2^{\mathrm{T}} h \right\|_2^2, \tag{32}$$

where $G_2$ is the orthonormal matrix by itself while being orthogonal to $G_1$. Hence problem (31) is equivalent to minimizing the distance from $f = G_1^{\mathrm{T}} h$ to the ellipse

$$w^{\mathrm{T}} \Delta^{-2} w = 1$$

with $w = \Delta y$. It can be argued easily that the minimizer is unique under the conditions of the theorem, which completes the proof. □

**Remark.** Following the above proof, we can show that all the solutions of problem (19), (24) have totally two different diagonal vectors in the case when $\mathrm{rank}([p_1, p_2]) = 2$, $f(2) \neq 0$, and $|f(1)| \leqslant \|[p_1, p_2]\|_2$. One may construct such an example.

In order to generalize Theorem 3.1 to the case $k \equiv n - \alpha \geqslant 1$ and $m \equiv \beta - \alpha > 2$, we need to characterize the structure of set $\Gamma$.

**Lemma 3.1.** *Let $Q \in \mathscr{R}^{N\times m}$ be an orthonormal matrix and $k \leqslant m$. Then there exist integer $r \leqslant (m^2 + m)/2 - 1$, constant vector $p_0 \in \mathscr{R}^N$, constant matrix $P \in \mathscr{R}^{N\times r}$, and unit column vector $\phi(Y) \in \mathscr{R}^r$ with r independent functional components defined over the Stiefel manifold*

$$\mathscr{V} = \left\{Y \in \mathscr{R}^{m\times k} : Y^{\mathrm{T}}Y = I_k\right\},$$

*such that for each $Y \in \mathscr{V}$,*

$$\mathrm{diag}(QYY^{\mathrm{T}}Q^{\mathrm{T}}) = p_0 + P\phi(Y).$$

**Proof.** Let us first consider the case when $k = 1$. Define

$$H = H(Y) \equiv mYY^{\mathrm{T}} - I_m$$

and denote $h = \mathrm{vect}(H)$ as the column vector consisting of the entries of $H$, i.e.,

$$h(Y) = \mathrm{vect}(H) = \left[h_{11}, \ldots, h_{m1}, h_{12}, \ldots, h_{m2}, \ldots, h_{1m}, \ldots, h_{mm}\right]^{\mathrm{T}}.$$

Clearly, $h(Y)$ has a constant norm for any $Y$,

$$\|h(Y)\|_2^2 = \|H\|_{\mathrm{F}}^2 = \mathrm{tr}(H^2) = m^2 - m.$$

Substituting $YY^{\mathrm{T}} = (I_m + H)/m$ into $z(Y) = \mathrm{diag}(QYY^{\mathrm{T}}Q^{\mathrm{T}})$ yields

$$z(Y) = \frac{1}{m}\mathrm{diag}(QQ^{\mathrm{T}}) + \frac{1}{m}\mathrm{diag}(QHQ^{\mathrm{T}}) = p_0 + \hat{P}h(Y),$$

where $p_0 = \mathrm{diag}(QQ^{\mathrm{T}})/m$ and $\hat{P}$ with the rows

$$\hat{P}(i,:) = \frac{1}{m}\left[Q(i,1)Q(i,:), \ldots, Q(i,n)Q(i,:)\right].$$

Here $X(i,:)$ denotes the $i$th row vector of matrix $X$. (We will also use $X(:,j)$ later to denote the $j$th column vector of $X$.) Obviously, the sequence of $m^2$ functions $h_{ij}(Y)$ are not linearly independent for $Y \in \mathscr{V}$ because $h_{ij}(Y) \equiv h_{ji}(Y)$ and

$$h_{11}(Y) + h_{22}(Y) + \cdots + h_{mm}(Y) = \mathrm{tr}(H(Y)) \equiv 0.$$

Let $r$ be the number of independent component functions of $h(Y)$ (so there must be $r \leqslant (m^2 + m)/2 - 1$). We can write $h(Y)$ as

$$h(Y) = F\hat{\phi}(Y),$$

where $F$ is a constant matrix, and $\hat{\phi}(Y)$ is a $r$-dimension column vector consisting of independent component functions. We can further assume that $F$ is orthonormal. (Otherwise we can take SVD for $F$, $F = \hat{U}\hat{S}\hat{V}^{\mathrm{T}}$ and replace $F$ and $\hat{\phi}$ by $\hat{U}$ and $\hat{S}\hat{V}^{\mathrm{T}}\hat{\phi}$, respectively.) Clearly we also have $\|\hat{\phi}(Y)\|_2 = \sqrt{m^2 - m}$. Define

$$P = \sqrt{m^2 - m}\ \hat{P}F, \quad \phi(Y) = \hat{\phi}(Y)/\sqrt{m^2 - m},$$

we then have

$$z(Y) = p_0 + P\phi(Y)$$

with unit vector $\phi(Y)$.

Next we consider the case $k > 1$, i.e., $Y$ has multiple columns. From the proof above we have that for each column $Y(:, j)$ of $Y \in \mathscr{V}$,

$$\mathrm{diag}\big(QY(:, j)Y(:, j)^{\mathrm{T}}Q^{\mathrm{T}}\big) = p_0 + P\phi(Y(:, j)),$$

which renders

$$\mathrm{diag}(QYY^{\mathrm{T}}Q^{\mathrm{T}}) = kp_0 + P\sum_{j=1}^{k}\phi(Y(:, j)) = kp_0 + P\tilde{\phi}(Y)$$

with $\tilde{\phi}(Y) = \sum_{j=1}^{k}\phi(Y(:, j))$. Next we want to prove that the $r$ component functions of $\tilde{\phi}(Y)$ are also linearly independent and the norm of $\tilde{\phi}(Y)$ is a constant for $Y \in \mathscr{V}$, so that it can be normalized without affecting the linear independence of the component functions.

To show the linear independence, let $a$ be a constant vector such that $a^{\mathrm{T}}\tilde{\phi}(Y) \equiv 0$ for $Y \in \mathscr{V}$ and let $b = Fa$. Note that for $i = 1, \ldots, r$,

$$a(i) = F(:, i)^{\mathrm{T}}b = \mathrm{tr}\big(\mathrm{vect}^{-1}(b)^{\mathrm{T}}\mathrm{vect}^{-1}(F(:, i))\big).$$

(Here $\mathrm{vect}^{-1}(x)$ denoted the unique matrix $X$ of order $m$ satisfying $\mathrm{vect}(X) = x$ for an $m^2$-dimensional vector $x$.) Then $a = 0$ and the linear independence follows immediately if we prove that: (1) the $m$-by-$m$ matrix $B = \mathrm{vect}^{-1}(b)$ is skew-symmetric, and (2) for each $i = 1, \ldots, r$, the matrix $\mathrm{vect}^{-1}(F(:, i))$ is symmetric. The key of proof is the formula

$$h(y) = \sqrt{m^2 - m}F\phi(y) \tag{33}$$

for unit vector $y$.

To show that $B$ is skew-symmetric, we first use (33) with $y = Y(:, j)$ to obtain that

$$0 = \sqrt{m^2 - m}\, a^{\mathrm{T}}\tilde{\phi}(Y) = b^{\mathrm{T}} \sum_{j=1}^{k} h(Y(:, j)) = m\,\mathrm{tr}\big(B(YY^{\mathrm{T}} - I_m)\big),$$

i.e., there must be $\mathrm{tr}(B(YY^{\mathrm{T}} - I_m)) = 0$. For any orthonormal matrix $G \in \mathscr{R}^{m\times(m-k)}$, one can write $GG^{\mathrm{T}} = YY^{\mathrm{T}} - I_m$ for a $Y \in \mathscr{V}$, so we have

$$\mathrm{tr}(G^{\mathrm{T}}BG) = 0$$

for any orthonormal matrix $G \in \mathscr{R}^{m\times(m-k)}$, which implies that $B$ has zero diagonal entries. Based on these properties we have that for any orthogonal matrix $[Y \quad G]$ with $Y$ of $k$ columns and $G$ of $(m-k)$ columns,

$$\mathrm{tr}(Y^{\mathrm{T}}BY) = 0 \quad \text{and} \quad \mathrm{tr}(G^{\mathrm{T}}BG) = 0. \tag{34}$$

Take the spectral decomposition of the symmetric matrix

$$B + B^{\mathrm{T}} = \tilde{Q}\tilde{D}\tilde{Q}^{\mathrm{T}} = \tilde{Q}_1\tilde{D}_1\tilde{Q}_1^{\mathrm{T}} + \tilde{Q}_2\tilde{D}_2\tilde{Q}_2^{\mathrm{T}}$$

for a orthogonal matrix $\tilde{Q} = [\tilde{Q}_1, \tilde{Q}_2]$ and a diagonal matrix $\tilde{D} = \mathrm{diag}(\tilde{D}_1, \tilde{D}_2)$ with decreasing diagonals. We then apply (34) with $Y = \tilde{Q}_1$ and $G = \tilde{Q}_2$, and obtain $\mathrm{tr}(\tilde{D}_i) = 0$ for $i = 1, 2$. It then follows that $B$ is skew-symmetric: $B^{\mathrm{T}} = -B$.

To show that $\mathrm{vect}^{-1}(F(:, i))$ is symmetric, we let $y_1, \ldots, y_r$ be $r$ unit vectors such that the matrix $\Phi = [\phi(y_1), \ldots, \phi(y_r)]$ is nonsingular. Furthermore, we let $\eta = [\eta(1), \ldots, \eta(r)]^{\mathrm{T}}$ be the $i$th column of $\Phi^{-1}$ (the index $i$ is omitted for simplicity), i.e., $\sum_{j=1}^{r} \eta(j)\phi(y_j) = e_i$, the $i$th column of identity matrix $I_r$, $i = 1, \ldots, r$. Again we use (33) with $y = y_j$, $j = 1, \ldots, r$, and obtain

$$\begin{aligned} F(:, i) = F\sum_{j=1}^{r}\eta(j)\phi(y_j) &= \frac{1}{\sqrt{m^2 - m}} F \sum_{j=1}^{r} \eta(j)h(y_j) \\ &= \frac{1}{\sqrt{m^2 - m}} \mathrm{vect}\left(\sum_{j=1}^{r}\eta(j)\big(my_jy_j^{\mathrm{T}} - I_m\big)\right). \end{aligned}$$

Clearly,

$$\mathrm{vect}^{-1}(F(:, i)) = \frac{1}{\sqrt{m^2 - m}} \sum_{j=1}^{r} \eta(j)\big(my_jy_j^{\mathrm{T}} - I_m\big)$$

is symmetric.

Finally, to normalize $\tilde{\phi}(Y)$, we again use (33) with $y = Y(:, j)$, $j = 1, \ldots, k$, and have for $i \neq j$,

$$\begin{aligned} (m^2 - m)\phi(Y(:, i))^{\mathrm{T}}\phi(Y(:, j)) &= h(Y(:, i))^{\mathrm{T}}h(Y(:, j)) \\ &= \mathrm{tr}\big(H(Y(:, i))H(Y(:, j))\big) = -m. \end{aligned}$$

It follows that

$$\|\tilde{\phi}(Y)\|_2^2 = k - \frac{k^2 - k}{m - 1} = \frac{k(m - k)}{m - 1}.$$

Redefine

$$p_0 := kp_0, \quad P := \sqrt{k(m-k)/(m-1)}P, \quad \text{and}$$
$$\phi(Y) := \sqrt{(m-1)/(k(m-k))}\tilde{\phi}(Y),$$

we again end up with

$$z = p_0 + P\phi(Y).$$

This completes the proof. □

Lemma 3.1 shows that $\Gamma$ is an ellipsoid of lower dimension for the general cases $k = n - \alpha \geqslant 1$ and $m = \beta - \alpha > 2$. Following similar arguments, we can extend Theorem 3.1 to this general case. The proof for the next theorem is omitted due to similarity.

**Theorem 3.2.** *Let $P = G_1 \Delta W$ be the SVD of $P$. Denote $f = G_1^{\mathrm{T}}(\mathrm{diag}(C - C_\alpha) - p_0)$. If $P$ is full rank in column and $f^{\mathrm{T}}\Delta^{-2}f > 1$, then all solutions to problems* (19) *and* (24) *have same diagonal vectors.*

Although (19), (24) may not have unique solution, the resulted objective function $V(d)$ is unique and is given by

$$V(d) = -\sum_{j=n+1}^{N} \lambda_j^2(d) + \|d\|_2^2. \tag{35}$$

Since $\{\lambda_j\}$'s depend continuously on $d$, so does $V(d)$. Yet in the case of multiple eigenvalues, the eigenvalues and $V(d)$ are generally not differentiable with respect to $d$. Moreover, in the case of multiple eigenvalues, matrix $C(d)$ may be non-continuous. Given below is an example with positive definite matrix $C$.

**Example**

Construct matrix $C$ of order $N = 3$ as

$$C = \begin{pmatrix} 7 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 7 \end{pmatrix}.$$

For $n = 2$, the minimizer $d^*$ and the optimal solution $C(d^*)$ are, respectively,

$$d^* = \begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix} \quad \text{and} \quad C(d^*) = \begin{pmatrix} 7 & 4 & 1 \\ 4 & 4 & 4 \\ 1 & 4 & 7 \end{pmatrix}.$$

Let us define $d(t) = (1, 4 + 2t, 1)^{\mathrm{T}}$. One can verify that the eigenvalues of $C + \operatorname{diag}(d(t))$ are given by

$$\lambda_1(t) = 9 + t + \sqrt{(t-1)^2 + 8}, \quad \lambda_2(t) = 6,$$
$$\lambda_3(t) = 9 + t - \sqrt{(t-1)^2 + 8}.$$

Note that at $t = 0$ we have a double eigenvalue $\lambda_2(0) = \lambda_3(0) = 6$. For $t \neq 0$, we have the corresponding (non-normalized) eigenvectors as

$$u_1(t) = \begin{pmatrix} 1 - t + \sqrt{(t-1)^2 + 8} \\ 4 \\ 1 - t + \sqrt{(t-1)^2 + 8} \end{pmatrix}, \quad u_2(t) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$
$$u_3(t) = \begin{pmatrix} 1 - t - \sqrt{(t-1)^2 + 8} \\ 4 \\ 1 - t - \sqrt{(t-1)^2 + 8} \end{pmatrix}.$$

If $t > 0$, then $\lambda_1(t) > \lambda_3(t) > \lambda_2(t)$ and

$$C(d(t)) = \lambda_1(t) \frac{u_1(t)u_1(t)^{\mathrm{T}}}{\|u_1(t)\|_2^2} + \lambda_3(t) \frac{u_3(t)u_3(t)^{\mathrm{T}}}{\|u_3(t)\|_2^2} \longrightarrow \begin{pmatrix} 5 & 2 & 5 \\ 2 & 8 & 2 \\ 5 & 2 & 5 \end{pmatrix}$$

as $t \to 0$, while for $t < 0$, $C(d(t)) \to C(d^*)$. The one-sided derivatives of $V(d(t))$ are quite different:

$$\frac{\mathrm{d}}{\mathrm{d}t} V(d(t))\bigg|_{t=+0} = 16, \quad \frac{\mathrm{d}}{\mathrm{d}t} V(d(t))\bigg|_{t=-0} = 0.$$

The example shows that $C(d(t))$ is discontinuous and $V(d(t))$ is not differentiable at $t = 0$.

## 4. Analysis for the equivalence

In this section, we will give a justification to the equivalence between the constrained low-rank approximation problem (15) and the min–max problem (16). Such equivalence would have been easy to show, should there be certain degree of analyticity for the solution(s), which is however not the case here: we do not have the differentiability for the objective function $V(d)$, the uniqueness and even the continuity of $C(d)$.

First of all, we prove the existence of solution(s) of the min–max problem (16) or the minimization problem (20). Our analysis will make extensive use of the linear convexity property of $V(d)$.

**Theorem 4.1.** *There exists at least one solution to* (20), *and any local minimum of* $V(d)$ *is also a global minimum.*

**Proof.** To prove existence we use the method of contradiction. Suppose there is no solution to (16), then there exist a sequence of multiplier vectors $d^{(j)} \to \infty$ such that $V(d^{(j)})$ decreases. Write $D^{(j)} = D_1^{(j)} + D_2^{(j)}$ as a direct sum of two diagonal matrices with $\operatorname{rank}(D_1^{(j)}) \leqslant n$ and $\|D_1^{(j)}\|_\infty = \|D^{(j)}\|_\infty \longrightarrow +\infty$. Since

$$V(d^{(j)}) = \max_{X\in\mathscr{R}_n} L(X, d^{(j)}) \geqslant L\big(D_1^{(j)}, d^{(j)}\big),$$

we have, using (21),

$$\begin{aligned} V(d^{(j)}) &\geqslant -\big\|C + D_2^{(j)}\big\|_{\mathrm{F}}^2 + \|D^{(j)}\|_{\mathrm{F}}^2 \\ &= -\|C\|_{\mathrm{F}}^2 - 2\operatorname{tr}\big(CD_2^{(j)}\big) + \big\|D_1^{(j)}\big\|_{\mathrm{F}}^2 \\ &\geqslant -\|C\|_{\mathrm{F}}^2 + \big(\|D^{(j)}\|_\infty - 2\operatorname{tr}(|C|)\big)\|D^{(j)}\|_\infty \longrightarrow +\infty. \end{aligned} \tag{36}$$

This however contradicts to the assumption that $V(d^{(j)})$ decreases. Here we use $|C|$ to denote the matrix with entries of $C$ in absolute values. The existence of solution(s) is hence obtained.

The property that any local minimum must be at the same time a global minimum stems from the convexity of $V(d)$. Consider any two vectors $d$ and $\hat{d}$. For any $t \in (0, 1)$, the linear-convexity property of $L(d)$ yields

$$\begin{aligned} V(td + (1-t)\hat{d}) &= \max_{X\in\mathscr{R}_n} \big(tL(X, d) + (1-t)L(X, \hat{d})\big) \\ &\leqslant t\max_{X\in\mathscr{R}_n} L(X, d) + (1-t)\max_{X\in\mathscr{R}_n} L(X, \hat{d}) \\ &= tV(d) + (1-t)V(\hat{d}) \\ &\leqslant \max(V(d), V(\hat{d})). \end{aligned} \tag{37}$$

Inequality (37) implies that all local minimums have the same values, so any local minimum must at the same time be a global minimum. □

We remark that if $C$ can be permuted to a block-diagonal matrix, say,

$$P^{\mathrm{T}}CP = \begin{pmatrix} C_1 & \\ & C_2 \end{pmatrix},$$

where $P$ is a permutation matrix, then the optimal solution $C^*$ to the problem (15) is given by

$$C^* = P\begin{pmatrix} C_1^* & \\ & C_2^* \end{pmatrix}P^{\mathrm{T}},$$

with $C_i^*$ $(i = 1, 2)$ being an optimal rank-$n_i$ approximation to $C_i$ (in the sense of (15) for certain splitting $n = n_1 + n_2$. We are now ready to show the uniqueness of the minimizer $d^*$.

**Theorem 4.2.** *If C cannot be permuted into a block diagonal matrix, then the minimizer $d^*$ of the optimal problem* (19) *is unique.*

**Proof.** Assume to the contrary that there exists $d^{**} \neq d^*$ such that $V(d^{**}) = V(d^*)$. Denote $d(t) = d^* + t(d^{**} - d^*)$ for $t \in [0, 1]$, then we have

$$\begin{aligned} V(d(t)) &= L(C(d(t)), d(t)) \\ &= (1-t)L(C(d(t)), d^*) + tL(C(d(t)), d^{**}) \\ &\leqslant (1-t)V(d^*) + tV(d^{**}) = V(d^*), \end{aligned} \tag{38}$$

by the linear convexity property (18). This means that $V(d(t)) = V(d^*)$ for all $t \in [0, 1]$. Thus all $d(t)$, $t \in [0, 1]$, are minimizers as well. From the above equalities we also get that for each $X \in \mathscr{S}(d(t))$, $L(X, d^*) = V(d^*)$, which means that $X$ is also an optimal rank-$n$ approximation to matrix $C + D^*$, so we conclude that

$$\mathscr{S}(d(t)) = \mathscr{S}(d^*).$$

We can write, therefore,

$$C + D^* = X + E, \quad C + D(t) = X + F(t),$$

for fixed $X \in \mathscr{B}(d^*)$, and $E$ and $F(t)$ are orthogonal to $X$. Subtracting the above two equalities we end up with

$$t(D^{**} - D^*) = D(t) - D^* = F(t) - E.$$

This implies that for $t \in (0, 1]$,

$$(D^{**} - D^*)X = 0$$

because of the orthogonality between $X$ and $E$ or $F(t)$. Without loss of generality we can write

$$D^{**} - D^* = \begin{pmatrix} 0^* & \\ & D_2 \end{pmatrix}$$

with non-singular diagonal matrix $D_2$. It follows that

$$X = \begin{pmatrix} X_1 & \\ & 0 \end{pmatrix}, \quad E = \begin{pmatrix} 0 & \\ & E_2 \end{pmatrix},$$

which gives, for the partition $D^* = \mathrm{diag}(D_1^*,\ D_2^*)$ that

$$C = X + E - D^* = \begin{pmatrix} X_1 - D_1^* & \\ & E_2 - D_2^* \end{pmatrix}.$$

A contradiction thus arises: $C$ is block-diagonal. □

To prove the equivalence between the constrained minimization problem (15) and the mini–max problem (19), (20), we need to establish some properties of $V(d)$ to substitute the differentiability. The following inequality, although simple, places a crucial role in the subsequent analysis.

**Theorem 4.3.** *For any two vectors* $d,\ \hat{d}$, *and any* $X \in \mathscr{S}(d)$,

$$V(\hat{d}) - V(d) \geqslant 2(d - \hat{d})^{\mathrm{T}}\mathrm{diag}(C - X).$$

**Proof.** By definition,

$$V(\hat{d}) \geqslant L(X, \hat{d}) = -\|C - X\|_{\mathrm{F}}^2 - 2\hat{d}^{\mathrm{T}}\mathrm{diag}(C - X). \tag{39}$$

Note that

$$V(d) = -\|C - X\|_{\mathrm{F}}^2 - 2d^{\mathrm{T}}\mathrm{diag}(C - X).$$

The inequality of the theorem follows immediately. □

An immediate consequence of the above theorem is that a vector $d^*$ satisfying $\mathrm{diag}(C(d^*)) = \mathrm{diag}(C)$ must be the minimizer of $V(d)$. We will show next that it is true vice-versa provided the $n$th eigenvalue of $C + \mathrm{diag}(d^*)$ is simple.

**Theorem 4.4.** *Let* $d^*$ *be a minimizer of* $V(d)$. *If* $\lambda_n(d^*)$ *is simple*, *then*

$$\mathrm{diag}(C(d^*)) = \mathrm{diag}(C),$$

*and* $C(d^*)$ *also solves the constrained problem* (15).

**Proof.** Let $d^*$ be a minimizer of $V(d)$ and $X = C(d^*)$ the unique matrix in $\mathscr{S}(d^*)$. Applying Theorem 4.3 for $\hat{d} = d^*$, $d = d(t) \equiv d^* + t \cdot \mathrm{diag}(C - X)$, and $Y(t) \in \mathscr{S}(d(t))$ with positive $t$, we obtain

$$0 \geqslant V(d^*) - V(d(t)) \geqslant t \cdot \mathrm{diag}(C - X)^{\mathrm{T}}\mathrm{diag}(C - Y(t)).$$

It follows that for all positive $t$,

$$\mathrm{diag}(C - X)^{\mathrm{T}}\mathrm{diag}(C - Y(t)) \leqslant 0.$$

Because all the eigenvalues of a matrix are continuously dependent on the entries of the matrix, we can conclude that for sufficiently small $t$, the $n$th eigenvalue of

$C + \mathrm{diag}(d(t))$ is also simple and then $Y(t)$ tends to $X$ as $t \to 0$. Therefore let $t \to 0$, we obtain that

$$\mathrm{diag}(C - X)^{\mathrm{T}}\mathrm{diag}(C - X) \leqslant 0,$$

which implies that $\mathrm{diag}(X) = \mathrm{diag}(C)$.

The property that $\mathrm{diag}(C(d^*)) = \mathrm{diag}(C)$ results in simple form of $V(d^*)$:

$$V(d^*) = -\|C - C(d^*)\|_{\mathrm{F}}^2.$$

For any $X \in \mathscr{S}$, we have

$$\|C - X\|_{\mathrm{F}}^2 = -L(X, d^*) \geqslant -V(d^*) = \|C - C(d^*)\|_{\mathrm{F}}^2.$$

Therefore $C(d^*)$ is an optimal solution of the constrained problem (15). □

As was shown in Section 2, in the application of market model calibration we must have positive semi-definite lower-rank approximations to the given correlation matrices. This constraint has not been imposed in our solution procedure. Yet, as we will show next, the positive semi-definiteness of the optimal approximations holds automatically.

**Theorem 4.5.** *Assume matrix $C$ is positive semi-definite. If the* $n$th *eigenvalue of matrix $C + \mathrm{diag}(d^*)$ is simple for the optimal multiplier $d^*$, then* (1) *$d^*$ is nonnegative, and* (2) *the solution $C^* = C(d^*)$ is also positive semi-definite.*

**Proof.** The positive semi-definiteness of $C^*$ follows directly from $d^* \geqslant 0$, since $C + \mathrm{diag}(d^*)$ is positive semi-definite. We prove the non-negativeness of $d^*$ using the method of contradiction. Without loss of generalities, we can assume that $C$ has no zero rows/columns. (Otherwise we can apply the theorem to a smaller problem for $C_1$ to show the non-negativeness of the corresponding minimizer $d_1^*$ when $C$ is permuted as

$$C = P\begin{pmatrix} C_1 & \\ & 0 \end{pmatrix}P^{\mathrm{T}},$$

where $C_1$ has no zero rows/columns. Then the minimizer $d^* = P[(d_1^*)^{\mathrm{T}}, 0]^{\mathrm{T}}$.)

Assume on the contrary that $d^*$ has a negative component. Write $C + D^* = C(d^*) + E$ with $E$ orthogonal to $C(d^*)$. Furthermore, without loss of generality, we assume that the diagonal entries of $D^*$ are nondecreasing so that the first diagonal entry $d^*(1)$ of $D^* = \mathrm{diag}(d^*)$ is the smallest one. By Theorem 4.4, $\mathrm{diag}(C) = \mathrm{diag}(C(d^*))$. So we have $d^* = \mathrm{diag}(E)$ and the matrix $E$ must have at least one negative eigenvalue. Let $\mu$ be the smallest eigenvalue of $E$ and $x$ be the corresponding eigenvector with unit 2-norm. It is easy to verify by the orthogonality that $C(d^*)x = 0$ because $\mu \neq 0$. Hence we have that

$$d^*(1) \leqslant x^{\mathrm{T}}D^*x = \mu - x^{\mathrm{T}}Cx \leqslant \mu.$$

On the other hand, recalling that the diagonals of a symmetric matrix are bounded by its 2-norm [4], we have that $\mu \leqslant d^*(1)$. Therefore

$$d^*(1) = \mu$$

and the first few of diagonals of $E - \mu I$ must be zeros. Note that $E - \mu I$ is a positive semi-definite matrix. Thus the first few rows and columns of $E - \mu I$ should be zero, too, i.e., $E$ is a block diagonal matrix,

$$E = \begin{pmatrix} \mu I & \\ & E_0 \end{pmatrix}.$$

By the orthogonality between $C(d^*)$ and $E$, we have also

$$C(d^*) = \begin{pmatrix} 0 & \\ & C_0 \end{pmatrix}.$$

Hence the first row/column of matrix $C = C(d^*) + E - \mathrm{diag}(d^*)$ must be zero, which contradicts the initial assumption. □

Finally, we remark that due to the convexity of $V(d)$, its minimization can be easily subdued by gradient-based minimization approaches. As an example, we show that min $V(d)$ can be solved with the method of descending, whose details are given in the following algorithm.

**Algorithm.** Take $d^{(0)}$ to be an initial guess for $d^*$, and repeat the following steps:

1. Compute the SVD of $C + D^{(k)}$ for $C(d^{(k)})$, solve the additional problem (24) if necessary.
2. If $\|\mathrm{diag}(C - C(d^{(k)}))\|_2 \leqslant \mathrm{tol}$, terminate iteration.
3. Define $d(t) = d^{(k)} + t \cdot \mathrm{diag}(C - C(d^{(k)}))$, and solve the one-dimensional subproblem $\min_{t \geqslant 0} V(d(t))$ to get the optimal $t = t^{(k)}$. [2]
4. Set $d^{(k+1)} = d^{(k)} + t^{(k)} \cdot \mathrm{diag}(C - C(d^{(k)}))$, and go back to step 1.

For the convergence of the above algorithm we have

**Proposition 1.** *The sequence* $\{d^{(k)}\}$ *has the following properties*:

(1) *The sequence* $\{d^{(k)}\}$ *is bounded*.
(2) *For any accumulation point* $d^*$ *of* $\{d^{(k)}\}$, *if the* $n$th *singular value of* $C + \mathrm{diag}(d^*)$ *is single*, *then* $d^*$ *must be a global minimizer*.
(3) *Furthermore*, *if* $C$ *cannot be permuted to be a block diagonal matrix*, $\{d_i^{(k)}\}$ *converges to the unique global minimizer* $d^*$.

[2] By Theorem 4.3, $V(d(t)) \geqslant V(d^{(k)})$ for $t < 0$.

**Proof.** As was shown in the proof of Theorem 4.1, $V(d^{(k)})$ will be unbounded from above if $\{d^{(k)}\}$ are unbounded. The monotonic decreasing of the function $V(d^{(k)})$ guarantees the boundedness of the sequence $\{d^{(k)}\}$. For bounded $\{d^{(k)}\}$, there must be at least one accumulation point.

Let $d^*$ be an accumulation point. If $d^*$ is not a minimizer, we must have, by Theorem 4.4, $\operatorname{diag}(C - C(d^*)) \neq 0$. Consider $V(d(t))$ with $d(t) = d^* + t \cdot \operatorname{diag}(C - C(d^*))$. Applying Theorem 4.3 with $d = d(t)$, $\hat{d} = d^*$, and $X = C(d(t))$ we have that

$$V(d^*) - V(d(t)) \geqslant t \cdot \operatorname{diag}(C - C(d^*))^{\mathrm{T}} \operatorname{diag}(C - C(d(t))).$$

Note that $C(d(t)) \to C(d^*)$ as $t \to 0$. We have that for some sufficiently small $t > 0$, $\operatorname{diag}(C - C(d^*))^{\mathrm{T}} \operatorname{diag}(C - C(d(t))) > 0$ which gives $V(d^*) - V(d(t)) > 0$. Therefore we have

$$V(d(t^*)) = \min_{t>0} V(d(t)) < V(d^*).$$

Denote $d^{**} = d(t^*)$. For any subsequence $\{d^{(n_k)}\}$ of $\{d^{(k)}\}$, which converges to the accumulation point $d^*$, we have $d^{(n_k)} + t^* \operatorname{diag}(C - C(d^{(n_k)})) \to d^{**}$ by definition. Hence we have that for sufficiently large $n_k$,

$$|V(d^{(n_k)} + t^* \operatorname{diag}(C - C(d^{(n_k)}))) - V(d^{**})| < \tfrac{1}{2}(V(d^*) - V(d^{**})).$$

It follows that

$$V(d^{(n_k)} + t^* \operatorname{diag}(C - C(d^{(n_k)}))) < V(d^{**}) + \tfrac{1}{2}(V(d^*) - V(d^{**})) < V(d^*).$$

Minimizing $V(d^{(n_k)} + t \cdot \operatorname{diag}(C - C(d^{(n_k)})))$ for the variable $t$ with fixed $n_k$ leads to

$$V(d^{(n_k+1)}) < V(d^*),$$

which contradicts to the property of monotonic decreasing of $V(d^{(k)})$. □

We conclude this section with a remark. The above method can be easily extended to the Frobenius norms with "weights". In some applications, the correlation between some forward rates are more relevant to the problem than the rest of the correlations. So we want to ensure in the calibration process that the relevance is properly emphasized. To this purpose we consider the Frobenius norm with "weights":

$$\|A\|_{\mathrm{W,F}}^2 = \left\|\sqrt{W} A \sqrt{W}\right\|_{\mathrm{F}}^2 \tag{40}$$

with

$$W = \operatorname{diag}(w_1, w_2, \ldots, w_n) \tag{41}$$

a diagonal matrix with positive entries. If we think the correlation of the first $i_0$ forward rates are more important than the correlations between other rates, we can take $w_i = 1, i = 1, \ldots, i_0$, while keeping $w_i < 1, i > i_0$. In the computations with the weighted norm we simply replace $\sqrt{W} C \sqrt{W}$ for $C$.

## 5. Numerical results

The first example is taken from [8]. We consider a collection of twelve 12-month forward rates with hypothetical historical (market) correlation matrix $C$ given by

$$c_{ij}^{\text{market}} = LongCorr + (1 - LongCorr)\exp[\beta|t_i - t_j|],$$
$$\beta = d_1 - d_2 \max(t_i, t_j),$$

where $LongCorr = 0.3$, $d_1 = -0.12$, $d_2 = 0.005$. In our computation we take $t_i = i \times \left(\frac{1}{2}\right)$. We will compare the correlation surfaces (or matrices) and curves (or matrix columns) between the models and the market. Also, we will display the principle components of the market and model correlation matrices. Due to the obvious superiority of our method to Rebonato's [8] we will not compare the performance of the two methods. Instead, we attach the method of Rebonato in Appendix A for readers' reference.

Without loss of generality we consider three-factor approximation. The results are given in Table 1 and Figs. 1 and 2. Fig. 1 shows the market correlation surface (on the left) and the model correlation surface (on the right). Apparently, the latter is smooth and in good agreement with the market correlation, except the extent of convexity near the diagonal. Table 1 and Fig. 2 display (the difference of) principle components. It can be seen that, while the first principle component of the model is very close to that of the market, there exist visible difference in the second and the third principle components of the market and model correlation matrices, respectively. This is actually a desirable property because the first component describes the proportional parallel shift of the rates, and it is the most important component that characterizes the correlation between the different rates. The market and model correlation between (a) the first, (b) third, (c) the sixth, and (d) the tenth forward rate and the rest of the forward rates are displayed in Fig. 3. The overall error of model correlation is very small.

Table 1
Principle components of the rank-one correction

| $U_1$ | $U_2$ | $U_3$ | $U_{a,1}$ | $U_{a,2}$ | $U_{a,3}$ |
|---|---|---|---|---|---|
| 0.86 | −0.39 | 0.22 | 0.87 | −0.42 | 0.27 |
| 0.89 | −0.38 | 0.17 | 0.90 | −0.40 | 0.20 |
| 0.92 | −0.34 | 0.08 | 0.92 | −0.37 | 0.11 |
| 0.93 | −0.27 | −0.04 | 0.95 | −0.31 | −0.04 |
| 0.95 | −0.18 | −0.15 | 0.96 | −0.20 | −0.20 |
| 0.95 | −0.07 | −0.22 | 0.96 | −0.07 | −0.26 |
| 0.95 | 0.05 | −0.23 | 0.96 | 0.06 | −0.27 |
| 0.95 | 0.17 | −0.17 | 0.96 | 0.19 | −0.22 |
| 0.93 | 0.27 | −0.07 | 0.95 | 0.31 | −0.08 |
| 0.91 | 0.35 | 0.06 | 0.92 | 0.38 | 0.10 |
| 0.88 | 0.40 | 0.18 | 0.89 | 0.41 | 0.21 |
| 0.85 | 0.41 | 0.24 | 0.85 | 0.43 | 0.29 |

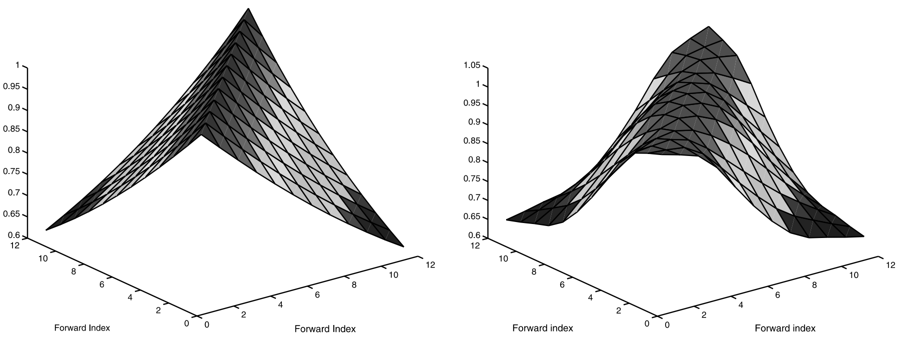


Fig. 1. Market correlation surface (left) and correlation surface of rank-one correction method (right).

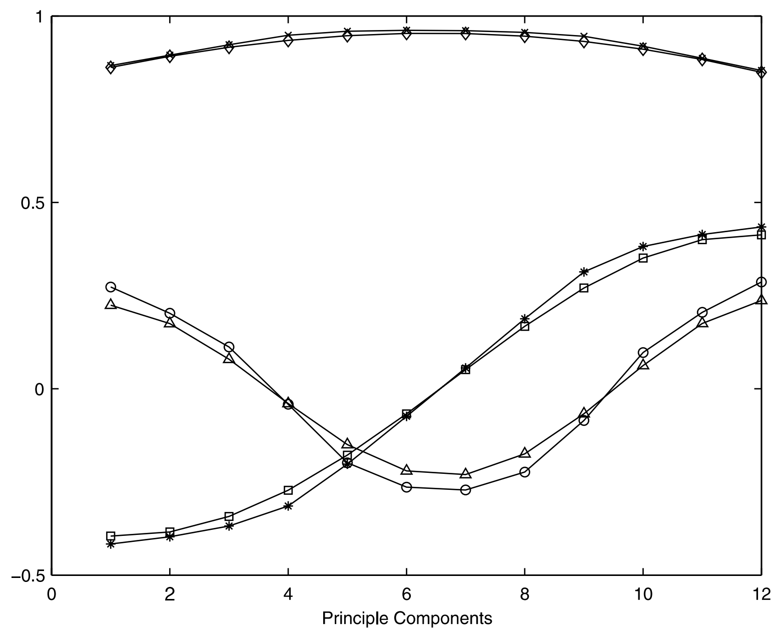


Fig. 2. The first three principle components of the market and model correlation matrixes.

The second is a practical example taken from [2]. In this example, the historical correlation matrix of Sterling pound is listed in Table 2, where the first column and row are the maturities of the forward rates. The correlation matrix is visualized by the surface plot in Fig. 4. For this matrix, we calculate its rank one, two, three and six approximations and plot the results in Fig. 5. The computation time for all approximations is in the magnitude of seconds. The trend of convergence with respect to the rank increase is given in Fig. 6.

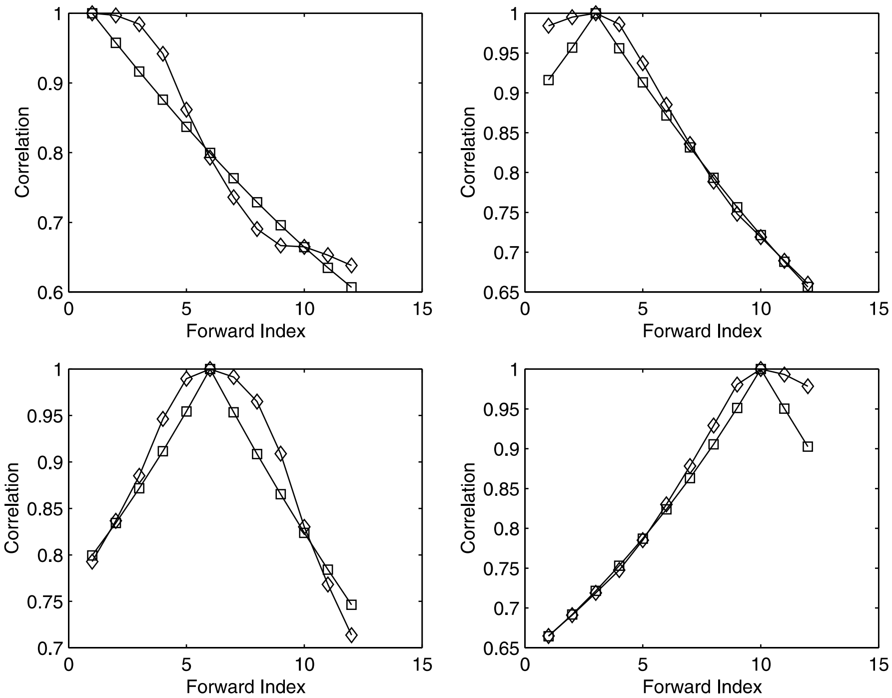


Fig. 3. The market and model correlations between (a) the first, (b) the third, (c) the sixth and (d) the tenth forward rates and the rest of the forward rates obtained using three-factor iterative model.

Table 2
Historical correlation matrix for the GBP rates

| | 0.25 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 4 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.25 | 1.0000 | 0.8415 | 0.6246 | 0.6231 | 0.5330 | 0.4287 | 0.3274 | 0.4463 | 0.2439 | 0.3326 | 0.2625 |
| 0.5 | 0.8415 | 1.0000 | 0.7903 | 0.7844 | 0.7320 | 0.6346 | 0.4521 | 0.5812 | 0.3439 | 0.4533 | 0.3661 |
| 1 | 0.6246 | 0.7903 | 1.0000 | 0.9967 | 0.8108 | 0.7239 | 0.5429 | 0.6121 | 0.4426 | 0.5189 | 0.4251 |
| 1.5 | 0.6231 | 0.7844 | 0.9967 | 1.0000 | 0.8149 | 0.7286 | 0.5384 | 0.6169 | 0.4464 | 0.5233 | 0.4299 |
| 2 | 0.5330 | 0.7320 | 0.8108 | 0.8149 | 1.0000 | 0.9756 | 0.5676 | 0.6860 | 0.4969 | 0.5734 | 0.4771 |
| 2.5 | 0.4287 | 0.6346 | 0.7239 | 0.7286 | 0.9756 | 1.0000 | 0.5457 | 0.6583 | 0.4921 | 0.5510 | 0.4581 |
| 3 | 0.3274 | 0.4521 | 0.5429 | 0.5384 | 0.5676 | 0.5457 | 1.0000 | 0.5942 | 0.6078 | 0.6751 | 0.6017 |
| 4 | 0.4463 | 0.5812 | 0.6121 | 0.6169 | 0.6860 | 0.6583 | 0.5942 | 1.0000 | 0.4845 | 0.6452 | 0.5673 |
| 5 | 0.2439 | 0.3439 | 0.4426 | 0.4464 | 0.4969 | 0.4921 | 0.6078 | 0.4845 | 1.0000 | 0.6015 | 0.5200 |
| 7 | 0.3326 | 0.4533 | 0.5189 | 0.5233 | 0.5734 | 0.5510 | 0.6751 | 0.6452 | 0.6015 | 1.0000 | 0.9889 |
| 9 | 0.2625 | 0.3661 | 0.4251 | 0.4299 | 0.4771 | 0.4581 | 0.6017 | 0.5673 | 0.5200 | 0.9889 | 1.0000 |

## 6. Conclusion

In this paper we have developed a very efficient method to find the low-rank approximating of correlation matrices. With the method of Lagrange multiplier we to

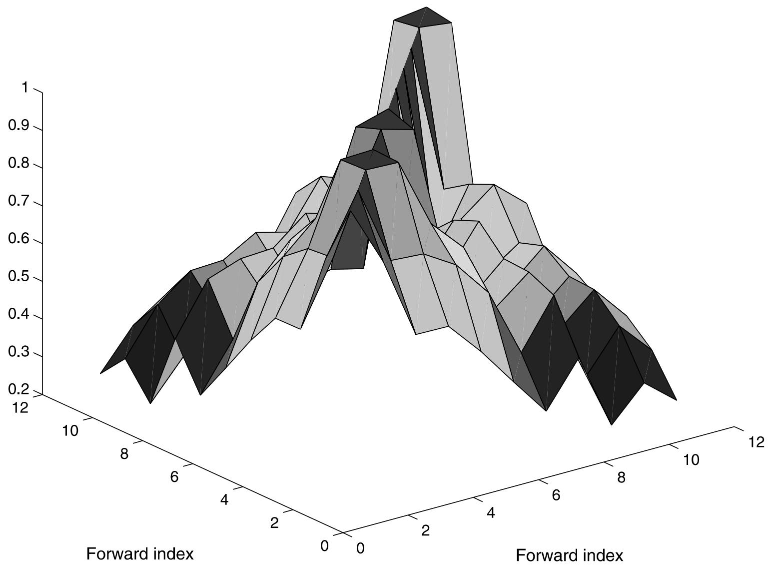


Fig. 4. Market correlation surface.

turn the constrained minimization into an min–max problem without constraint. The inner maximization problem is solved with a single spectral decomposition, while the outer minimization problem is solve iteratively with gradient-based methods. The convergence of the iteration (for the outer minimization problem) is guaranteed due to the convexity of the objective function. This technique has a direct application in calibrating the market model in financial engineering. The method developed in this paper extends naturally to the matrix approximation problem with other kinds of constraints.

## Acknowledgements

This project is partially supported by the Special Funds for Major State Basic Research Projects (project G19990328), and Foundation for University Key Teacher by the Ministry of Education, China; and Direct Allocation Grand DAG99/00.SC24, Hong Kong. We want to thank the anonymous referee for helpful comments.

## Appendix A. Method of Rebonato

To solve for (15), Rebonato [8] considers solution of the form

$$X = BB^{\mathrm{T}}, \tag{A.1}$$

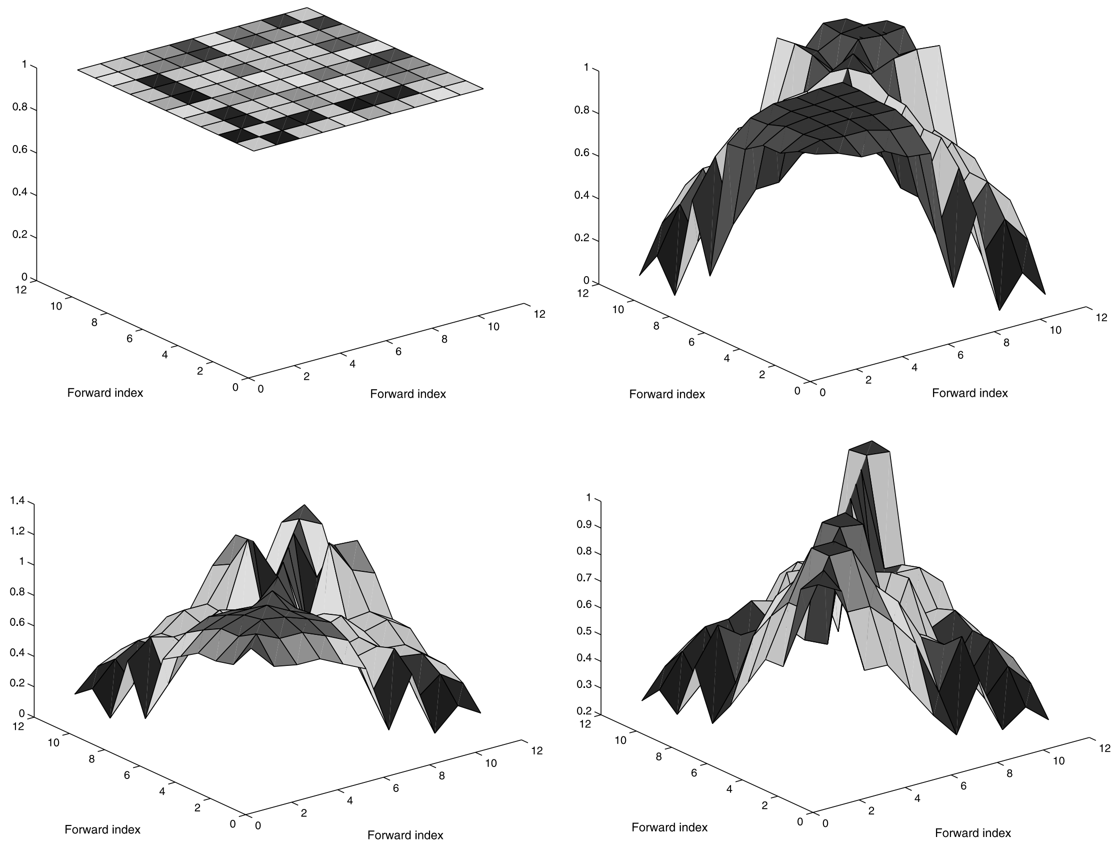


Fig. 5. Model correlation surfaces: one factor (1,1); two factors (1,2); three factors (2,1); six factors (2,2).

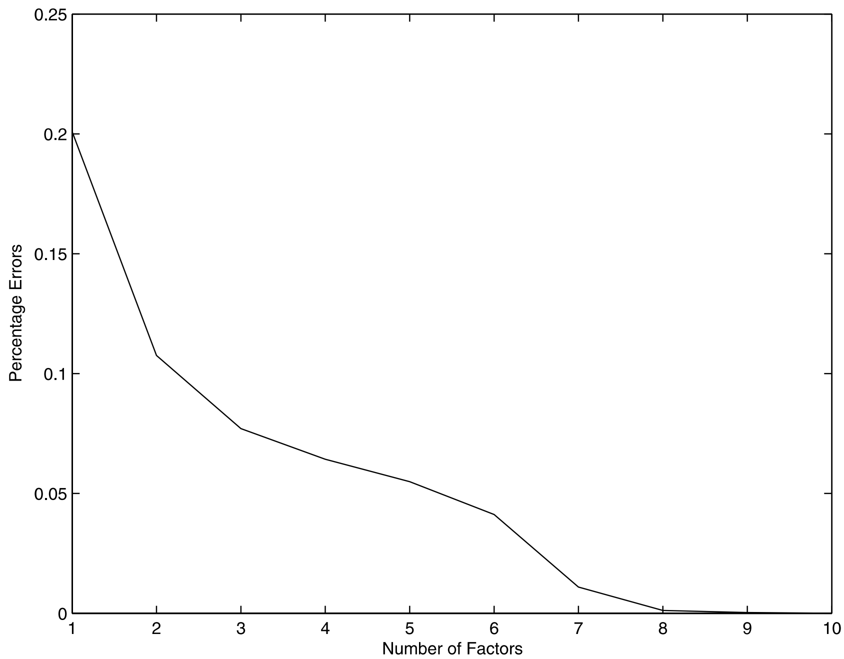


Fig. 6. Trend of convergence with increasing rank.

where $B$ is an $N$ by $n_t$ matrix whose elements are of the parametric form

$$
\begin{aligned}
b_{jk} &= \cos(\theta_{jk}) \prod_{l=1}^{k-1} \sin(\theta_{jl}), \quad k = 1, \ldots, n_t - 1, \\
b_{jn_t} &= \prod_{l=1}^{n_t-1} \sin(\theta_{jl}).
\end{aligned}
\tag{A.2}
$$

Note that representation (A.1) guarantees the rank of $X$ to be less or equal to $n_t$, while the parameterization (A.2) ensures the "one-diagonal" condition as we have

$$
\sum_{k=1}^{n_t} b_{jk}^2 = 1, \quad j = 1, 2, \ldots, N, \tag{A.3}
$$

for any angles $\{\theta_{jk}\}$. Given the representation and parameterization which effectively remove the constraint, Rebonato proceeds to solve the unconstraint problem

$$
\min_{\{\theta_{jk}\}} \left\| C - B(\{\theta_{jk}\}) B^{\mathrm{T}}(\{\theta_{jk}\}) \right\|_{\mathrm{F}} \tag{A.4}
$$

with standard unconstraint minimization methodologies. This is a nonlinear optimization problem with $N \times n_t$ unknowns. In financial applications, this number can go as high as $80 \times 4 = 320$, which then poses a horrendous challenge to any existing methodologies.

## References


[1] F. Black, The pricing of commodity contract, J. Financial Econom. 3 (1976) 167–179.
[2] A. Brace, D. Gatarek, M. Musiela, The Market model of interest rate dynamics, Math. Finance 7 (2) (1997) 127–155.
[3] A. Edelman, T.A. Arias, S.T. Smith, The geometry of algorithm with orthogonality constraints, SIAM J. Matrix Anal. Appl. 20 (1998) 303–353.
[4] G. Golub, C. von Loan, Matrix Computations, 3rd ed., Johns Hokkins University Press, Baltimore, MD, 1996.
[5] D.G. Luenberger, Optimization by Vector Spaces Methods, John Wiley, 1969.
[6] F. Jamshidian, LIBOR and swap market models and measures, Finance Stoch. 1 (1997) 293–330.
[7] K. Miltersen, K. Sandmann, D. Sondermann, Closed-form solutions for term structure derivatives with lognormal interest rates, J. Finance (1997) 409–430.
[8] R. Rebonato, Calibrating the BGM model, RISK March (1999) 74–79.