The Matrix Ridge Approximation: Algorithms and Applications

Zhihua Zhang
MOE-Microsoft Key Lab for Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai, China 200240
zhihua@sjtu.edu.cn

Revised, November 27, 2013

Abstract
We are concerned with an approximation problem for a symmetric positive semidefinite matrix due to motivation from a class of nonlinear machine learning methods. We discuss an approximation approach that we call matrix ridge approximation. In particular, we define the matrix ridge approximation as an incomplete matrix factorization plus a ridge term. Moreover, we present probabilistic interpretations using a normal latent variable model and a Wishart model for this approximation approach. The idea behind the latent variable model in turn leads us to an efficient EM iterative method for handling the matrix ridge approximation problem. Finally, we illustrate the applications of the approximation approach in multivariate data analysis. Empirical studies in spectral clustering and Gaussian process regression show that the matrix ridge approximation with the EM iteration is potentially useful.

Keywords: Positive semidefinite matrices; Matrix ridge approximation; Incomplete matrix factorization; Expectation maximization algorithms; Probabilistic models.

1. Introduction
Symmetric positive semidefinite matrices play an important role in multivariate statistical analysis and machine learning. Especially, the low-rank approximation of a positive semidefinite matrix has been widely applied to multivariate data analysis. In this paper we study the low-rank approximation problem of a positive semidefinite matrix as well as its applications in machine learning. Moreover, we always assume that the positive semidefinite matrix in question is symmetric.

Some machine learning methods require computing the inverse of a positive definite matrix or the spectral decomposition of a positive semidefinite matrix. For example, the kernel PCA (principal component analysis) (Schölkopf and Smola, 2002), classical multidimensional scaling (also called principal coordinate analysis, PCO) (Mardia et al., 1979) and spectral clustering algorithms (Zhang and Jordan, 2008) require solving an eigenvalue problem with linear constraints on an $m \times m$ inner-product matrix (Golub, 1973), and Gaussian processes (GPs) (Rasmussen and Williams, 2006) need to invert $m \times m$ covariance matrices.
Typically, these methods take $O(m^3)$ operations where $m$ denotes the number of training instances. This scaling is unfavorable for applications in massive datasets.

Several approaches have been also proposed to address this computational challenge, such as randomized techniques (Achlioptas et al., 2001), sparse greedy approximation (Smola and Schölkopf, 2000), and the Nyström method (Williams and Seeger, 2001, Yang et al., 2012). All these approaches are based on sampling techniques. Similar ideas include random Fourier features (Rahimi and Recht, 2008, Quiñonero-Candela et al., 2007, Lázaro-Gredilla et al., 2010, Le et al., 2013) and hashing features (Shi et al., 2009). Specifically, the random feature method avoids inversion of a matrix by solving a linear system of equations instead. Another widely used approach is to employ the incomplete Cholesky decomposition method (Golub and Loan, 1996, Fine et al., 2001). The approach is deterministic. Although these approaches can be efficient, their range of applications might be limited; e.g., these approaches are always infeasible in handling the eigenvalue decomposition problem with linear constraints.

In this paper we present a new deterministic low-rank approximation approach. Roughly speaking, the approach is to approximate a positive semidefinite matrix as an incomplete matrix decomposition plus a ridge term. We refer to such an approximation method as the matrix ridge approximation due to its direct motivation from the ridge regression model (Hoerl and Kennard, 1970). The approximation is built on an optimization problem with linear constraints. This problem can be in turn solved by using the conventional spectral decomposition technique or an efficient iterative method.

Although the idea behind the matrix ridge approximation is simple, our method is attractive. Firstly, it yields an approximation tighter than the incomplete Cholesky decomposition and the incomplete spectral decomposition do. Secondly, it yields an approximate matrix, whose condition number is not higher than that of the original matrix. This can make numerical computations involved more stable. More importantly, it can widen the application range of the low-rank approximation approach. Particularly, we show that our method can be applied to the approximate computation of the inverse and spectral decomposition of a positive (semi)definite matrix. We illustrate the application of the matrix ridge approximation in spectral clustering and Gaussian process regression.

We also discuss two statistical counterparts for the ridge approximation. The first counterpart is in the spirit of probabilistic interpretations of some machine learning methods, including probabilistic PCA (Tipping and Bishop, 1999, Roweis, 1998, Ahn and Oh, 2003), probabilistic nonlinear component analysis (Rosipal and Girolami, 2001), Gaussian process latent variable models (Lawrence, 2004), and factor analysis (Magnus and Neudecker, 1999). In particular, we define a normal latent variable model in which we impose the linear constraints. Based on the latent variable model, we devise an iterative method, i.e., the expectation-maximization (EM) algorithm (Dempster et al., 1977), for solving the matrix ridge approximation problem.

The second counterpart is a Wishart model, which is derived from the normal latent variable model by using the relationship between Wishart distributions and Gaussian distributions (Gupta and Nagar, 2000, Zhang et al., 2006). These two statistical counterparts in turn define probabilistic matrix ridge approximation models. Moreover, we show that the maximum likelihood approach to estimating the parameters of the probabilistic models results in the same solution as that based on the standard spectral decomposition technique.
The remainder of the paper is organized as follows. We first give the notation in Section 2. We present the matrix ridge approximation in Section 3 and illustrate its applications in Section 4. We reformulate the matrix ridge approximation by using a normal latent variable model and a Wishart model in Section 5. Consequently, we develop probabilistic ridge approximation and an EM iterative algorithm. Section 6 conducts the empirical analysis, and Section 7 concludes our work. Note that we put all proofs to the appendices.

2. Notation and Terminology

We let $I_m$ denote the $m \times m$ identity matrix, and $1_m$ denote the $m \times 1$ vector of ones. For a matrix $Y$, we denote its rank, Frobenius norm and condition number by $\text{rk}(Y)$, $\|Y\|_F$ and $\kappa(Y)$, respectively. When $Y$ is square, we denote its determinant and trace by $|Y|$ and $\text{tr}(Y)$. Additionally, $A \otimes B$ denotes the Kronecker product of $A$ and $B$.

For an $s \times t$ random matrix $Z$, $Z \sim N_{s,t}(M, A \otimes B)$ means that $Z = [z_{ij}]$ $(s \times t)$ follows a matrix-variate normal distribution with mean matrix $M = [m_{ij}]$ $(s \times t)$ and covariance matrix $A \otimes B$, where $A$ $(s \times s)$ and $B$ $(t \times t)$ are symmetric positive definite. Note that a matrix variate normal distribution is defined through a multivariate normal distribution (Gupta and Nagar, 2000). In particular, let $\text{vec}(Z^T) = (z_{11}, \ldots, z_{1t}, z_{21}, \ldots, z_{st})^T$ $(st \times 1)$ and $\text{vec}(M^T) = (m_{11}, \ldots, m_{1t}, m_{21}, \ldots, m_{st})^T$ $(st \times 1)$. Then, $Z \sim N_{s,t}(M, A \otimes B)$ if and only if $\text{vec}(Z^T) \sim N_{st}(\text{vec}(M^T), A \otimes B)$. We also use the notation in Gupta and Nagar (2000) for Wishart distributions. That is, for an $m \times m$ positive definite random $Y$, $Y \sim W_m(r, \Sigma)$ represents that $Y$ follows a Wishart distribution with degree of freedom $r$.

Finally, in Table 1 we list some notations that will be used throughout this paper. It is clear that $H_b H_b^T = H_b$ and $P P = P$; i.e., they are idempotent. Moreover, we have $H_b 1_m = 0$, $b^T H_b = 0$, $P b = 0$ and $b^T P = 0$. A typical nonzero case for $b$ is $b = \frac{1}{\sqrt{m}}1_m$. This case implies that $H_b = P = I_m - \frac{1}{m} 1_m 1_m^T$ and $A^T 1_m = 0$ (that is, the mean of the rows of $A$ is zero). In addition, let us keep in mind that $H_b = P = I_m$ when $b = 0$ for notational simplicity. In this case we always have $M = S = T$.

| $b \in \mathbb{R}_+^m$ | a $m$-dimensional nonnegative vector |
| $M \in S_+^{m \times m}$ | a positive semidefinite matrix of rank $p$ $(1 < p \leq m)$ |
| $H_b = I_m - \frac{1}{m} b b^T$ $(b \neq 0)$ | centering matrix |
| $P = I_m - b b^T$ | projection matrix |
| $S = H_b M H_b^T$ | positive semidefinite matrix |
| $T = P M P$ | positive semidefinite matrix |

3. The Matrix Ridge Approximation

We are given a nonnegative vector $b \in \mathbb{R}_+^m$ and a positive semidefinite matrix $M \in S_+^{m \times m}$ of rank $p$ $(1 < p \leq m)$. The ridge approximation of $M$ is defined as

$$M \approx A A^T + \delta I_m,$$
where $\delta > 0$ is called a ridge term, and $A \in \mathbb{R}^{m \times q}$ is a matrix of full column rank $q (< p)$ and satisfies $A^T b = 0$. The idea behind the matrix ridge approximation is simple, and the terminology is motivated by the ridge regression model (Hoerl and Kennard, 1970). Note that when $b = 0$, $A^T b = 0$ is always true. This implies no constraints. In this paper we consider both the cases with and without the linear constraints. Since $A^T b = 0$ is equivalent to $cA^T b = 0$ for any nonzero constant $c$, we assume that $b^T b = 1$ whenever $b \neq 0$ to make the constraint identifiable.

The constraint $A^T b = 0$ for $b \neq 0$ is often met in machine learning methods such as the classical multidimensional scaling (Gower and Legendre, 1986), kernel PCA (Schölkopf and Smola, 2002), spectral clustering (Zhang and Jordan, 2008), etc. If $b = 0$ and $\delta = 0$, we obtain the incomplete factorization $M \approx AA^T$ straightforwardly by using the spectral decomposition of $M$ (Magnus and Neudecker, 1999). In this setting, the ridge approximation is also closely related to the incomplete Cholesky factorization (Golub and Loan, 1996). Furthermore, if $q = p$ it is feasible to obtain an exact expression $M = AA^T$ via the spectral (or Cholesky) decomposition. In this paper we concentrate on the case that $q < p$ and $\delta > 0$, so we have a sparse plus low-rank approximation of $M$ ($\delta I_m$ is sparse and $AA^T$ is low-rank).

In order to estimate $A$ and $\delta$, we exploit two loss functions which were developed for estimation of covariance matrices (Anderson, 1984). In particular, the first loss function is a least-squares error:

$$F(A, \delta) = \text{tr}[(S - AA^T - \delta I_m)^2]$$

while the second loss is derived from the likelihood function; namely,

$$G(A, \delta) = \log |AA^T + \delta I_m| + \text{tr}[(AA^T + \delta I_m)^{-1}S].$$

**Theorem 1** Let $\gamma_1 \geq \cdots \geq \gamma_q \geq \cdots \geq \gamma_m$ ($\geq 0$) be the eigenvalues of $S = H_bM^2 H_b^T$, $V$ be an arbitrary $q \times q$ orthogonal matrix, $\Gamma_q$ be a $q \times q$ diagonal matrix containing the first $q$ principal (largest) eigenvalues $\gamma_i$, and $U_q$ be an $n \times q$ column-orthonormal matrix in which the $q$ column vectors are the principal eigenvectors corresponding to $\Gamma_q$. Assume that $\delta > 0$ and that $A \in \mathbb{R}^{m \times q}$ ($q < \min(m, p)$) is of full column rank and satisfies $A^T b = 0$. If there exists a $j \in \{q + 1, \ldots, m\}$ such that $\gamma_q > \gamma_j > 0$, then the strict local minimum of $F(A, \delta)$ and of $G(A, \delta)$ with respect to (w.r.t.) $(A, \delta)$ is obtained when

$$\hat{A} = U_q(\Gamma_q - \delta I_q)^{1/2}V \quad \text{and} \quad \hat{\delta} = \frac{1}{m-q} \sum_{j=q+1}^{m} \gamma_j.$$

Theorem 1 is a direct corollary of Theorem 7 in Appendix B. Theorem 1 also shows that the minimizer $(\hat{A}, \hat{\delta})$ of $F(A, \delta)$ is the same to that of $G(A, \delta)$. We consider the case that $b = 0$. In this case, the condition number of $M$ ($= S$) is $\kappa(M) = \frac{\gamma_1}{\gamma_m}$. It follows from Theorem 1 that $\kappa(\hat{A} \hat{A}^T + \hat{\delta} I_m) = \frac{\gamma_1}{m-q} \sum_{j=q+1}^{m} \gamma_j \leq \kappa(M)$. This implies that $\hat{A} \hat{A}^T + \hat{\delta} I_m$ is well-conditioned more than $M$ (Golub and Loan, 1996). In other words, if $M$ is well-conditioned, so is $\hat{A} \hat{A}^T + \hat{\delta} I_m$.

In addition, it is easily calculated that

$$F(\hat{A}, \hat{\delta}) = \sum_{i=q+1}^{m} \gamma_i^2 - \frac{1}{m-q} \left( \sum_{i=q+1}^{m} \gamma_i \right)^2.$$
It is well known that
\[
\inf_{\mathbf{B} \in \mathbb{R}^{m \times m} \, \text{rk}(\mathbf{B}) \leq q} \| \mathbf{S} - \mathbf{B} \|_F^2 = \inf_{\mathbf{Y} \in \mathbb{R}^{m \times q} \, \text{rk}(\mathbf{Y}) \leq q} \| \mathbf{S} - \mathbf{YY}^T \|_F^2 = \sum_{i=q+1}^{m} \gamma_i^2.
\]

Thus, when comparing the ridge approximation of \( \mathbf{S} \) with the incomplete Cholesky decomposition of \( \mathbf{S} \), we have
\[
\inf_{\mathbf{L} \in \mathcal{L} \, \text{rk}(\mathbf{L}) = q} \| \mathbf{S} - \mathbf{LL}^T \|_F^2 \geq \inf_{\mathbf{Y} \in \mathbb{R}^{m \times q} \, \text{rk}(\mathbf{Y}) = q} \| \mathbf{S} - \mathbf{YY}^T \|_F^2 \geq \inf_{\delta \geq 0, \mathbf{A} \in \mathbb{R}^{m \times q} \, \text{rk}(\mathbf{A}) = q} \| \mathbf{S} - \mathbf{AA}^T - \delta \mathbf{I}_m \|_F^2,
\]

where \( \mathcal{L} = \{ \mathbf{L} \in \mathbb{R}^{m \times q} : \mathbf{L} \text{ is lower triangular} \} \). This shows that the ridge approximation yields a tighter approximation of \( \mathbf{S} \) than both the incomplete Cholesky decomposition and the incomplete spectral decomposition do.

As we mentioned, \( G(\mathbf{A}, \delta) \) is associated with a likelihood function. In Section 5.1 we will show that \( G \) is derived from a normal latent variable model. Thus, the solution in Theorem 1 is in fact the conventional maximum likelihood (ML) estimate. Furthermore, the ML estimation method is based on the direct spectral decomposition of the \( m \times m \) matrix \( \mathbf{M} \) or \( \mathbf{S} \), which takes \( O(m^3) \) operations. Thus, the method is inefficient when \( m \) is very large. Based on the idea behind the latent variable model, we develop an iterative method for solving the matrix ridge approximation.

In particular, given the \( t \)-th estimates \( \mathbf{A}(t) \) and \( \delta(t) \) of \( \mathbf{A} \) and \( \delta \), the next estimates of \( \mathbf{A} \) and \( \delta \) in our iterative method are given as:
\[
\mathbf{A}_{(t+1)} = \mathbf{SA}_{(t)} \left( \delta_{(t)} \mathbf{I}_q + \mathbf{A}_{(t)}^T \mathbf{A}_{(t)}^{-1} \right)^{-1}, \quad (1)
\]
\[
\delta_{(t+1)} = \frac{1}{m} \left[ \text{tr}(\mathbf{S}) - \text{tr}(\mathbf{A}_{(t)}^T \mathbf{A}_{(t)}) \right], \quad (2)
\]

where \( \mathbf{\Sigma}_{(t)} = \delta_{(t)} \mathbf{I}_q + \mathbf{A}_{(t)}^T \mathbf{A}_{(t)} \). Derivation of the algorithm is given in Section 5.1 and Appendix E. This procedure involves multiplication of \( m \times m \) matrices by \( m \times q \) matrices and inversion of \( q \times q \) matrices. Inverting a \( q \times q \) matrix takes \( O(q^3) \) operations, and multiplying an \( m \times m \) matrix by an \( m \times q \) matrix runs in \( m^2 q \) flops. Thus, this method takes time \( O(Tm^2 q) \), where \( T \) is the maximum iterative number. The method is efficient because \( T \) is usually far smaller than \( m \) (even smaller than \( \sqrt{m} \)), especially when \( m \) is very large. In the following experiment, we will see that in most cases the EM iterations get convergence after about 20 steps. Moreover, the matrix multiplication can be easily implemented in parallel. Additionally, the EM method does not necessarily load whole \( m \times m \) matrix \( \mathbf{S} \) during the iterations, which can significantly reduce the storage space.

Given an initial matrix \( \mathbf{A}_{(0)} \) such that \( \text{range}(\mathbf{A}_{(0)}) \subseteq \text{range}(\mathbf{S}) \) where \( \text{range}(\mathbf{Z}) \) represents the space spanned by the columns of \( \mathbf{Z} \), we have the following lemma.

**Lemma 2** Assume that the matrices \( \{ \mathbf{A}_{(t)} \} \) are generated by (1) and (2). If \( \text{range}(\mathbf{A}_{(0)}) \subseteq \text{range}(\mathbf{S}) \) and \( \text{rk}(\mathbf{A}_{(0)}) = q \), then for all \( t > 1 \), the \( \mathbf{A}_{(t)} \) are of full column rank.
In Section 5.1 we will show that the iterative method given in (1) and (2) is a standard EM iterative procedure (Dempster et al., 1977). Consequently, its convergence has been well established (Wu, 1983). The following theorem proves that the constraints $A^T(0)b = 0$ always hold during the iteration procedure and the EM estimates converge to the corresponding ML estimates. In other words, the EM iteration converges to the strict local minimizer.

**Theorem 3** Given initial values $\delta_0$ and $A(0)$ subject to $\delta(0) > 0$ and $A^T(0)b = 0$, the values of $A(t)$ and $\delta(t)$ calculated via (1) and (2) always satisfy $A^T(t)b = 0$ and $\delta(t) > 0$. Moreover, the EM estimates of $A$ and $\delta$ converge to the conventional ML solutions given in Theorem 1.

The EM algorithm provides an efficient iterative method for computing the matrix ridge approximation. This iterative method is related to the power method and the Lanczos method (Golub and Loan, 1996), which typically serve for solving matrix eigenvector problems numerically. Specifically, this EM algorithm is similar to the QR orthogonal iteration, which is a straightforward generalization of the power method to find a higher-dimensional invariant subspace (Golub and Loan, 1996).

Intuitively, it seems interesting that we consider a two-step procedure to solve the matrix ridge approximation as follows. Specifically, we first apply the QR orthogonal iteration to obtain an $m \times q$ column-orthonormal matrix $Q$ and set $\hat{A} = Q(Q^TQ)^{1/2}$. We then calculate $\delta = \text{tr}(S) - \text{tr}(Q^TQ)$ based on the minimization of $\text{tr}((S - \hat{A}\hat{A}^T - \delta I_m)^2)$ w.r.t. $\delta$. Assume that $Q^TQ = \Gamma_q$. Then $\delta = \frac{\sum_{i=1}^{m} \gamma_i}{m}$. It is directly computed that

$$\text{tr}((S - \hat{A}\hat{A}^T - \delta I_m)^2) = \sum_{i=q+1}^{m} \gamma_i^2 - \frac{1}{m} \left( \sum_{i=q+1}^{m} \gamma_i \right)^2 > \sum_{i=q+1}^{m} \gamma_i^2 - \frac{1}{m - q} \left( \sum_{i=q+1}^{m} \gamma_i \right)^2.$$  

This implies that the two-step procedure can not find the optimum solution of the matrix ridge approximation problem. Moreover, we have $\kappa(\hat{A}\hat{A}^T + \delta I_m) = 1 + \frac{m \gamma_1^2}{\sum_{i=q+1}^{m} \gamma_i}$. Compared with our method, this two-step method results in the approximation with higher condition number. Moreover, the method can not keep the well-conditionedness of the original matrix (if it is well-conditioned). We will conduct simulation on a toy data in Section 6.1, which shows that the two-step method fails to solve the matrix ridge approximation problem.

It is worth noting that the nonnegativity on $b$ is not necessary in our derivation for the estimation methods. In fact, we are able to extend the constraints $A^Tb = 0$ to $A^TE = 0$ where $E$ is an $m \times k$ matrix of full column rank ($k + p \leq m$). In this case, letting $P = I_m - E(E^TE)^{-1}E^T$ and $S = PMP$, we alternatively use $\text{tr}(S - AA^T - \delta I_m)^2$ as the loss function. The resulting solution is also similar to that in Theorem 1.

4. Applications of the Matrix Ridge Approximation

The matrix ridge approximation has potential applications in multivariate analysis and machine learning. In this section we present two important examples to illustrate its applications.

Let $M$ be an $m \times m$ symmetric positive (semi)definite matrix. It is well known that the computational complexities of calculating the inverse of $M$ and the spectral decomposition of $M$ are $O(m^3)$. Thus, the computational costs are high when $m$ is large. We now address
these two computational issues via the matrix ridge approximation. First of all, assume we obtain that $M \approx \delta I_m + AA^T$ where $A \in \mathbb{R}^{m \times q}$ and $q \ll m$ using the EM iteration.

In the first example we consider the computation of $M^{-1}$ where $M$ is positive definite. We approximate $M^{-1}$ by $(\delta I_m + AA^T)^{-1}$ which is then calculated by using the Sherman-Morrison-Woodbury formula; i.e.,

$$(\delta I_m + AA^T)^{-1} = \delta^{-1} I_m - \delta^{-1} A(\delta I_q + A^T A)^{-1} A^T.$$ (3)

Clearly, the current complexity is $O(mq^2)$. Thus, the computational cost will become much lower when $q$ is far less than $m$.

Recall that the incomplete Cholesky decomposition is widely used in the literature. For the $m \times m$ positive definite matrix $M$, we can consider its approximation by using the incomplete Cholesky decomposition, that is, $M \approx LL^T$ where $L$ is an $m \times q$ lower triangular matrix. Since $LL^T$ is singular, this decomposition can not directly provide us an approach to the approximation of $M^{-1}$. Also, the Nyström method could not be directly used for the approximation of $M^{-1}$. We can employ the two-step procedure as discussed in the previous section. However, we have also shown that this two-step procedure can not find the optimum solution, which will be empirically illustrated in Section 6.1.

If $M$ has an explicit form of

$$M = \Phi + \alpha^2 I_m$$ (4)

where $\Phi$ is an available $m \times m$ positive semidefinite matrix and $\alpha \neq 0$ is prespecified, both the incomplete Cholesky decomposition and the Nyström method work. Specifically, one first implements either the incomplete Cholesky decomposition or the Nyström method on $\Phi$ to obtain $L$ and then uses the Sherman-Morrison-Woodbury formula. Since our method directly applies to $M$ (rather than $\Phi$), our method can obtain a tighter approximation to $M$. Consider that the ridge term in our method $\delta$ is $\alpha^2 + \frac{1}{m-q} \sum_{j=q+1}^{m} \lambda_j$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ are the eigenvalues of $\Phi$. The condition number of the approximate matrix with our method is

$$\kappa(AA^T + \delta I_m) = \frac{\alpha^2 + \lambda_1}{\alpha^2 + \frac{1}{m-q} \sum_{j=q+1}^{m} \lambda_j},$$

while the condition number with the incomplete Cholesky decomposition is

$$\kappa(LL^T + \alpha^2 I_m) = \frac{\alpha^2 + \lambda_1}{\alpha^2}.$$ 

Therefore, our method is more stable numerically especially when $\alpha^2$ takes a very small value. Our simulation in Section 6.1 further illustrates the issues. We will see that when $\alpha^2$ takes a very small value, the incomplete Cholesky decomposition fails to approximate the inversion of $M$.

We note that any strictly positive definite matrix $M$ can be expressed as in (4). For example, we take $\alpha^2$ as the smallest eigenvalue of $M$. In this case, it is required to estimate the smallest eigenvalue prior to the implementation of the incomplete Cholesky decomposition (or the Nyström method). Thus, the method becomes inefficient. Moreover, the previous issues still exist for the incomplete Cholesky decomposition and the Nyström method in comparison with our method.
In the second example, we are concerned with the symmetric eigenvector problem, which plays an important role in multivariate statistical analysis and machine learning. The eigenvector problem is defined by

$$\max_{X \in \mathbb{R}^{m \times q}} \frac{1}{2} \text{tr}(X^T MX)$$

subject to $X^T X = I_q$ and $X^T b = 0$.

If $b = 0$, Problem (5) becomes the standard Rayleigh quotient problem (Golub and Loan, 1996). Furthermore, if viewing $M$ as a sample covariance matrix, it is equivalent to the PCA problem (Jolliffe, 2002).

If $b^T b = 1$, the problem in (5) is a symmetric eigenvector problem with linear constraints (Golub, 1973). It defines a spectral clustering problem when $M$ is set as a kernel matrix (Zhang and Jordan, 2008).

Consider the spectral decomposition (or singular value decomposition, SVD) of $T = PMP$ as $T = U^T \Gamma U$ where $U \in \mathbb{R}^{m \times m}$ is orthogonal and $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_m)$ is arranged in descending order. Let $\hat{X} = U_q V$, where $U_q$ is the $m \times q$ matrix containing the first $q$ columns of $U$ and $V$ is an arbitrary $q \times q$ orthogonal matrix. Then the matrix $\hat{X}$ is the maximizer of the eigenvector problem in (5) (see, Golub, 1973). On the other hand, it follows from Theorem 1 that $\hat{A}(\hat{A}^T \hat{A})^{-\frac{1}{2}} = U_q V = \hat{X}$. This implies that we can obtain the solution of (5) via the matrix ridge approximation.

Note that if $b = 0$, $\hat{A}(\hat{A}^T \hat{A})^{-\frac{1}{2}}$ spans the same subspace as that spanned by the first $q$ principal eigenvectors of $M$ ($= T$). When $q = 1$, $\hat{A}(\hat{A}^T \hat{A})^{-\frac{1}{2}}$ is the top eigenvector of $M$. In this case, the EM iteration bears resemblance to the power method (Golub and Loan, 1996).

Naturally and intuitively, the incomplete Cholesky decomposition method might be used to approximate the solution of the problem in (5). Specifically, one first finds the incomplete Cholesky decomposition of $T$ as $T \approx LL^T$ and then treats $L(L^T L)^{-\frac{1}{2}}$ as the solution of Problem (5). However, to our knowledge, in the existing literature there is no theoretical guarantee that $L(L^T L)^{-\frac{1}{2}}$ is a solution of Problem (5). In fact, our experimental results in Section 6.1 show that the incomplete Cholesky decomposition method is not appropriate to approximate the solution of the problem in (5).

5. Probabilistic Matrix Ridge Approximation Models

In this section we consider two probabilistic models for the matrix ridge approximation. We thus show that the ML estimation approach for the parameters of the probabilistic models results in the same solution as that based on the standard spectral decomposition technique. The probabilistic formulation also gives rise to the EM iterative method defined in (1) and (2).

5.1 The Normal Latent Variable Model

In order to derive the EM iteration, we consider a probabilistic formulation of the matrix ridge approximation. Our work is directly motivated by existing probabilistic interpretations of dimensionality reduction methods, such as probabilistic PCA (Tipping and Bishop,
where \( \delta > 0 \).

Typically, only \( \mathbf{M} \) is available while both \( r \) and \( \mathbf{F} \) are unknown in our case. Fortunately, we will see that our model can work via some matrix tricks to yield an estimation procedure for the unknown parameters \( \mathbf{A} \) and \( \delta \), which does not explicitly depend on \( r \) and \( \mathbf{F} \).

It is clear that \( \mathbf{F} \sim N_{m,r}(\mathbf{1}_m \mathbf{u}^T, (\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m) \otimes \mathbf{I}_r / r) \). Thus, the log-likelihood is

\[
\ell(\mathbf{A}, \delta, \mathbf{u}) = -\frac{mr}{2} \log(2\pi) + \frac{mr}{2} \log r - \frac{r}{2} \log |\Omega| - \frac{r}{2} \text{tr}((\mathbf{F} - \mathbf{1}_m \mathbf{u}^T)^T \Omega^{-1} (\mathbf{F} - \mathbf{1}_m \mathbf{u}^T))
\]

\[
\propto -\log |\Omega| - \text{tr}((\mathbf{F} - \mathbf{1}_m \mathbf{u}^T)^T \Omega^{-1} (\mathbf{F} - \mathbf{1}_m \mathbf{u}^T))
\]

where \( \Omega = \mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m \).

We consider two setups for the mean vector \( \mathbf{u} \). In the first setup we let \( \mathbf{u} = \mathbf{0} \). We then see that maximizing \( \ell(\mathbf{A}, \delta, \mathbf{0}) \) is equivalent to minimizing \( G_1(\mathbf{A}, \delta) = \log |\Omega| + \text{tr}(\Omega^{-1} \mathbf{M}) \) where \( \mathbf{M} = \mathbf{F} \mathbf{F}^T \), w.r.t. \( (\mathbf{A}, \delta) \) under the constraint \( \mathbf{A}^T \mathbf{b} = \mathbf{0} \). In the second setup we let \( \mathbf{u} = \frac{1}{\mathbf{I}_r \mathbf{b}} \mathbf{F}^T \mathbf{b} \). Substituting such a \( \mathbf{u} \) in \( \ell(\mathbf{A}, \delta, \mathbf{u}) \) leads to the conclusion that the maximum likelihood estimate is equivalent to minimizing \( G(\mathbf{A}, \delta) = \log |\Omega| + \text{tr}(\Omega^{-1} \mathbf{S}) \) w.r.t. \( (\mathbf{A}, \delta) \) under the constraint \( \mathbf{A}^T \mathbf{b} = \mathbf{0} \). Thus, the matrix ridge approximation can also be solved from the probabilistic formulation.

Since our probabilistic model defined by (6) and (7) is a latent variable model, this encourages us to develop an EM algorithm for the parameter estimation. In particular, considering \( \mathbf{W} \) as the missing data, \( \{\mathbf{W}, \mathbf{F}\} \) as the complete data, and \( \mathbf{A} \) and \( \delta \) as the model parameters, we now have the EM algorithm for the matrix ridge approximation, which is given in (1) and (2). The derivation is then given in Appendix E. The algorithm is related to the EM algorithm derived in the literature (Roweis, 1998, Tipping and Bishop, 1999). However, we impose the constraint \( \mathbf{A}^T \mathbf{b} = \mathbf{0} \) in our model.

### 5.2 The Wishart Model

In this subsection we further explore the statistical properties of the matrix ridge approximation. In particular, we establish a Wishart model, corresponding to the treatments in the maximum likelihood estimation method.

First, we assume \( \mathbf{u} = \mathbf{0} \). We then have \( \mathbf{F} \sim N_{m,r}(\mathbf{0}, (\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m) \otimes \mathbf{I}_r / r) \). Consequently, \( \mathbf{M} = \mathbf{F} \mathbf{F}^T \) follows Wishart distribution \( W_m(r, (\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m) \otimes \mathbf{I}_r / r) \). Second, it follows from (6) that \( \mathbf{F} - \mathbf{1}_m \mathbf{u}^T | \mathbf{W} \sim N_{m,r}(\mathbf{A} \mathbf{W}, \delta \mathbf{I}_m \otimes \mathbf{I}_r / r) \). Hence,

\[
\mathbf{F} - \mathbf{1}_m \mathbf{u}^T \sim N_{m,r}(\mathbf{0}, (\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m) \otimes \mathbf{I}_r / r).
\]
Subsequently, \((F - 1_m u^T)(F - 1_m u^T)^T\) is distributed according to \(W_m(r, (AA^T + \delta I_m)/r)\). When \(u = \frac{1}{m} F^T b\), we thus have \((F - 1_m u^T)(F - 1_m u^T)^T = S \sim W_m(r, (AA^T + \delta I_m)/r)\).

Conversely, let \(M\) or \(S\) follow a Wishart distribution with an integral degree of freedom \(r\). According to the equivalence between Gaussian and Wishart distributions (Gupta and Nagar, 2000, Zhang et al., 2006), we can also obtain an \(m \times r\) matrix \(F\) which follows a matrix-variate normal distribution.

In the normal latent variable and Wishart models, we assume that \(r\), the dimensionality of \(F\), is finite. In the reproducing kernel literature (Schölkopf and Smola, 2002), \(r\) is the dimensionality of the feature space that can be infinite. For example, the dimensionality of the feature space induced by the Gaussian RBF kernel is infinite. In this case, we study the asymptotic distribution of \(S\). Specifically, the asymptotic distribution of \(\frac{1}{\sqrt{r}}(S - (AA^T + \delta I_m))\), as \(r \to \infty\), is a symmetric matrix-variate normal distribution (Gupta and Nagar, 2000).

It is worth pointing out that the latent variable models provide a probabilistic formulation for PCO. That is, it defines a probabilistic PCO model, which is dual to the probabilistic PCA model (Tipping and Bishop, 1999).

In the existing statistical approaches to multidimensional scaling (Ramsay, 1982, Groenen et al., 1995, Oh and Raftery, 2001), an error structure of \(\delta_{ij}^2\) is incorporated so that \(\delta_{ij}^2\), conditioned on \(d_{ij}^2 = \|y_i - y_j\|^2\), has p.d.f. \(p(\delta_{ij}^2|d_{ij}^2)\). Since \(\delta_{ij}\) must be nonnegative, \(\delta_{ij}^2\) is usually modeled as a truncated normal or log-normal distribution, with parameters \(d_{ij}^2\). Moreover, the \(\delta_{ij}^2\) are assumed to be independent. This provides an approach to the ML estimates of the \(y_i\). Some numerical methods such as gradient methods and Bayesian sampling methods such as MCMC are then used to calculate the \(y_i\).

However, these statistical approaches are not appropriate for probabilistic modeling of PCO. Since the dissimilarity matrix \(\Delta = [\delta_{ij}^2]\) in PCO is Euclidean, the triangle inequality

\[
\delta_{ij} + \delta_{ik} \geq \delta_{jk}
\]

should be satisfied. On one hand, this makes a conflict with the assumption that the \(\delta_{ij}^2\) are independent. On the other hand, for the \(\delta_{ij}^2\) generated from a truncated normal or log-normal distribution, the triangle inequality is no longer guaranteed. Accordingly, \(\Delta\) is not Euclidean. In our Wishart model the interactions among the \(\delta_{ij}^2\) are explored, because we treat the similarity matrix \(Q = -\frac{1}{4} H_b \Delta H_b^T\) as a Wishart matrix. Moreover, the positive semidefiniteness of \(Q\) implies the Euclideanarity of \(\Delta\) (e.g. see, Gower and Legendre, 1986).

6. Experiments

As we see from Theorem 1, the conventional ML estimation approach gives the same solution as the corresponding least squares counterpart. Moreover, the ML estimate is obtained by using the standard direct spectral decomposition (SD) technique. Our analysis has also provided an EM iterative algorithm. Thus, it is of interest to compare the performance of the EM algorithm with the direct SD method. All algorithms have been implemented in Matlab.
6.1 Performance Analysis on Toy Datasets

In Section 3 we show that the EM algorithm is more efficient than the SD method when \( m \) is large. Moreover, the solution of the EM algorithm converges to that of the conventional ML estimate based on the SD method. We performed our experimental analysis based on a toy dataset by studying the two applications of the matrix ridge approximation presented in Section 4.

In the simulation we used a 10×10 positive definite matrix \( M \), which is given by

\[
M = \begin{bmatrix}
1.8147 & 0.8650 & 0.8781 & 0.8106 & 0.9900 & 0.8270 & 0.8737 & 0.9851 & 0.6538 & 0.8958 \\
0.8650 & 1.9058 & 0.9560 & 0.9465 & 0.8311 & 0.5516 & 0.8781 & 0.9139 & 0.8781 & 0.9851 \\
0.8781 & 0.9560 & 1.1270 & 0.9704 & 0.8781 & 0.5543 & 0.9656 & 0.9185 & 0.9094 & 0.9512 \\
0.8106 & 0.9465 & 0.9704 & 1.9134 & 0.8106 & 0.5066 & 0.9465 & 0.8825 & 0.9560 & 0.9512 \\
0.9900 & 0.8311 & 0.8781 & 0.8106 & 1.6324 & 0.8270 & 0.9003 & 0.9753 & 0.6538 & 0.8694 \\
0.8270 & 0.5516 & 0.5543 & 0.5066 & 0.8270 & 1.0975 & 0.6096 & 0.7827 & 0.3447 & 0.6005 \\
0.8737 & 0.8781 & 0.9656 & 0.9465 & 0.9003 & 0.6096 & 1.2785 & 0.9139 & 0.8607 & 0.8914 \\
0.9851 & 0.9139 & 0.9185 & 0.8825 & 0.9753 & 0.7827 & 0.9139 & 1.5469 & 0.7334 & 0.9465 \\
0.6538 & 0.8781 & 0.9094 & 0.9560 & 0.6538 & 0.3447 & 0.8607 & 0.7334 & 1.9575 & 0.8564 \\
0.8958 & 0.9851 & 0.9512 & 0.9512 & 0.8694 & 0.6005 & 0.8914 & 0.9465 & 0.8564 & 1.9649
\end{bmatrix}
\]

The eigenvalues of \( M \) are 9.2521, 1.6413, 1.0326, 0.9460, 0.9386, 0.7530, 0.5925, 0.4736, 0.4142 and 0.1946. Thus, the eigenvalues of \( M^{-1} \) are 5.1387, 2.4143, 2.1115, 1.6878, 1.3280, 1.0654, 1.0571, 0.9684, 0.6093, 0.1081. Our current purpose is to approximate \( M^{-1} \) by using the ridge approximation of \( M \). That is, we first implemented the ML estimates of \( A \) and \( \delta \) and then calculated \( (\mathbf{AA}^T + \delta \mathbf{I}_m)^{-1} \)—an approximation of \( M^{-1} \)—in terms of (3). In this example, \( b = 0 \) which implies that there is no constraint for \( M \).

In the EM iteration we randomly generated 10 \( q \)-dimensional vectors from \( N_q(0, I_q) \) as the initial value \( A^{(0)} \) of \( A \) and set the initial value of \( \delta \) as \( \delta^{(0)} = 0.0001 \). We implemented our analysis for \( q = 1, \ldots, 9 \). After taking about 20 step, the EM iterations converge to the conventional ML solution based on the spectral decomposition method. Table 2 reports the SD-based ML estimates and the EM iteration estimates of \( \delta \) for \( q = 1, \ldots, 9 \). The corresponding values are almost identical.

We evaluate the performance of \( (\mathbf{AA}^T + \delta \mathbf{I}_m)^{-1} \), as an approximation of \( M^{-1} \), by employing the following two criteria:

\[
e_F = \frac{1}{\sqrt{m}} \| \mathbf{I}_m - M(\mathbf{AA}^T + \delta \mathbf{I}_m)^{-1} \|_F \quad \text{and} \quad e_2 = \| \mathbf{I}_m - M(\mathbf{AA}^T + \delta \mathbf{I}_m)^{-1} \|_2.
\]

The \( e_F \) and \( e_2 \) are given in Figure 1. We see that \( e_F \) and \( e_2 \) become small as \( q \) increases. Especially, when \( q = 9 \), their values are 0.0024 and 0.0076 respectively. Moreover, in this case, the eigenvalues of \( (\mathbf{AA}^T + \delta \mathbf{I}_m)^{-1} \) are 5.1381, 2.4143, 2.1115, 1.6877, 1.3281, 1.0654, 1.0571, 0.9684, 0.6093 and 0.1081, which are almost equal to those of \( M^{-1} \).

For comparison, we also performed the two-step method based on the QR orthogonal iteration (see Section 3). We define the initial column-orthonormal matrix \( Q^{(0)} = A^{(0)}(A^{(0)T}A^{(0)})^{-\frac{1}{2}} \), where \( A^{(0)} \) is the same to that for the EM iteration. As we see from Figure 1, for \( q = 1 \) and \( q = 2 \), the two-step method has approximation errors similar to the EM iteration method. However, the two-step method fails to obtain a good approximation in other cases. When \( q = 9 \), the errors of the method are \( e_F = 2.8462 \) and \( e_2 = 9.00 \).
Additionally, the eigenvalues of \((\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m)^{-1}\) with the QR iteration are 51.2821, 2.3057, 2.0281, 1.6339, 1.2945, 1.0437, 1.0357, 0.9505, 0.6021, and 0.1079.

Table 2: The estimated values of \(\delta\) with the ML based on the SD and EM iteration methods.

| \(q=1\) | \(q=2\) | \(q=3\) | \(q=4\) | \(q=5\) | \(q=6\) | \(q=7\) | \(q=8\) | \(q=9\) |
|---|---|---|---|---|---|---|---|---|
| SD | 0.7763 | 0.6681 | 0.6161 | 0.5611 | 0.4856 | 0.4187 | 0.3608 | 0.3044 | 0.1946 |
| EM | 0.7763 | 0.6681 | 0.6161 | 0.5614 | 0.4856 | 0.4187 | 0.3608 | 0.3044 | 0.1945 |

Figure 1: (a) The errors between \(\mathbf{M}^{-1}\) and its approximate \(\mathbf{A} \mathbf{A}^T + \delta \mathbf{I}_m\) where \(\mathbf{A}\) and \(\delta\) were estimated by the EM method and the QR method respectively, for \(q = 1, \ldots, 9\).

Let us see the estimates of \(\mathbf{A}\) in the cases that \(q = 1\) and \(q = 3\). First, when \(q = 1\), the EM estimate of \(\mathbf{A}\) is

\[
\mathbf{A} = (-0.9563, -0.9790, -0.9126, -0.9774, -0.9308, -0.6513, -0.9108, -0.9579, -0.8809, -1.0007)^T.
\]

It is further seen that \(\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-\frac{1}{2}} = (-0.3285, -0.3363, -0.3135, -0.3357, -0.3197, -0.2237, -0.3128, -0.3290, -0.3026, -0.3437)^T\) is the principal eigenvector of \(\mathbf{M}\).

When \(q = 3\), the matrix \(\mathbf{U}_3\) of the first three eigenvectors of \(\mathbf{M}\) and the EM estimate of \(\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-\frac{1}{2}}\) are respectively given by

\[
\mathbf{U}_3 = \begin{bmatrix}
-0.3285 & 0.4057 & 0.1792 \\
-0.3363 & -0.1540 & -0.4530 \\
-0.3135 & -0.0746 & 0.0302 \\
-0.3357 & -0.3073 & 0.1697 \\
-0.3197 & 0.3362 & 0.1897 \\
-0.2237 & 0.4044 & 0.1239 \\
-0.3128 & -0.0221 & 0.1241 \\
-0.3290 & 0.2035 & 0.0465 \\
-0.3026 & -0.6230 & 0.4150 \\
-0.3437 & -0.0711 & -0.7013
\end{bmatrix}
\]

and

\[
\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-\frac{1}{2}} = \begin{bmatrix}
0.1658 & -0.3889 & -0.3549 \\
0.3884 & -0.1290 & 0.4159 \\
0.0359 & -0.2973 & 0.1226 \\
-0.1863 & -0.3615 & 0.2647 \\
0.1194 & -0.3819 & -0.3019 \\
0.1769 & -0.2709 & -0.3525 \\
-0.0128 & -0.3352 & 0.0344 \\
0.1695 & -0.3296 & -0.1203 \\
-0.5560 & -0.4134 & 0.4152 \\
0.6341 & -0.0423 & 0.4610
\end{bmatrix}
\]
Assume that \( \text{range}(\mathbf{L})^{\frac{1}{2}} = \mathbf{U}_3 \mathbf{V} \) where
\[
\mathbf{V} = \begin{bmatrix}
-0.3130 & 0.9214 & -0.2304 \\
0.5098 & -0.0417 & -0.8593 \\
-0.8013 & -0.3864 & -0.4566
\end{bmatrix}
\]
is a 3\times3 orthogonal matrix. This is in line with the theoretical justification given in Section 4.

In the case that \( q = 3 \), we also implement the incomplete Cholesky decomposition of \( \mathbf{M} \) as \( \mathbf{M} \approx \mathbf{L} \mathbf{L}^T \) where \( \mathbf{L} \) and \( \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} \) are given as \(^1\)
\[
\mathbf{L} = \begin{bmatrix}
0.6391 & 0.2092 & 1.16727 \\
0.7028 & 0.3565 & 0.2924 \\
0.6786 & 0.3931 & 0.3103 \\
0.6786 & 0.4302 & 0.2458 \\
0.6202 & 0.2184 & 0.4694 \\
0.4284 & 0.0669 & 0.4621 \\
0.6359 & 0.3751 & 0.3331 \\
0.6752 & 0.2549 & 0.4286 \\
0.6110 & 1.2587 & 0 \\
1.4017 & 0 & 0
\end{bmatrix}
\]
and \( \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} = \begin{bmatrix}
0.0633 & 0.0016 & 0.8150 \\
0.2621 & 0.1315 & 0.0757 \\
0.2365 & 0.1659 & 0.0948 \\
0.2428 & 0.1977 & 0.0407 \\
0.2051 & 0.0321 & 0.2437 \\
0.1292 & -0.0505 & 0.2843 \\
0.2108 & 0.1598 & 0.1236 \\
0.2379 & 0.0512 & 0.1968 \\
0.0816 & 0.8915 & -0.1696 \\
0.8037 & -0.2992 & -0.3104
\end{bmatrix} \).

Assume that \( \text{range}(\mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}}) = \text{range}(\mathbf{U}_3) \). Then we have \( \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} = \mathbf{U}_3 \mathbf{R} \) where
\[
\mathbf{R} = \mathbf{U}_3^T \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} = \begin{bmatrix}
0.6248 & -0.5466 & 0.4554 \\
-0.6634 & -0.5633 & 0.3665 \\
0.16903 & -0.5628 & -0.6421
\end{bmatrix} .
\]

We further have \( \mathbf{I}_3 = (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} \mathbf{L}^T \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} = \mathbf{R}^T \mathbf{U}_3^T \mathbf{U}_3 \mathbf{R} = \mathbf{R}^T \mathbf{R} \). However, it is directly computed that \( \mathbf{R}^T \mathbf{R} \neq \mathbf{I}_3 \), yielding a conflict. This implies that the assumption \( \text{range}((\mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}}) = \text{range}((\mathbf{U}_3) \) is not true. Thus, this example shows that the incomplete Cholesky decomposition can not be used to find the top eigenvectors of an arbitrary positive definite matrix.

Additionally, we defined a new positive definite matrix \( \mathbf{K} \) as
\[
\mathbf{K} = \mathbf{M} + \alpha^2 \mathbf{I}_{10} ,
\]
which has an explicit form as in (4). As mentioned in Section 4, we employed the incomplete Cholesky decomposition to approximate \( \mathbf{K}^{-1} \). In particular, we first obtained the incomplete Cholesky decomposition of \( \mathbf{M} \) as \( \mathbf{M} \approx \mathbf{L} \mathbf{L}^T \) and then computed \( \alpha^2 \mathbf{I}_{10} + \mathbf{L} (\mathbf{L}^T \mathbf{L})^{-\frac{1}{2}} \mathbf{L}^T \) as the approximation to \( \mathbf{K}^{-1} \). Let \( q = 3 \). We took \( \alpha^2 = 0.1 \) and \( \alpha^2 = 0.0001 \) to implement the empirical analysis. When \( \alpha^2 = 0.1 \), \( e_F \) and \( e_2 \) with the incomplete Cholesky decomposition are respectively 7.0688 and 14.8785; \( e_F \) and \( e_2 \)

---

1. Our implementation is based on the code from http://theoval.cmp.uea.ac.uk/~gcc/matlab/default.html, which was written for the incomplete Cholesky decomposition algorithm described by Fine et al. (2001).
with our method are 0.3030 and 0.5886. When \(\alpha^2 = 0.0001\), \(e_F\) and \(e_2\) with the incomplete Cholesky decomposition are respectively \(7.0680 \times 10^3\) and \(1.4877 \times 10^3\); \(e_F\) and \(e_2\) with our method are 0.3522 and 0.6840. This shows that the incomplete Cholesky decomposition fails when \(\alpha^2\) takes a very small value. However, our method is numerically stable in every case. The reason is in that our method makes \(AA^T + \delta I_{10}\) better-conditioned than \(K\). But we see that \(LL^T + \alpha^2 I_{10}\) is more ill-conditioned than \(K\).

Finally, we performed a simulation on a cluster to further validate efficiency of our approach in inverting large-size matrices. We randomly generated a 50000 \(\times\) 50000 positive definite matrix \(M\) from Wishart distribution \(W_{50000}(50020, \Sigma)\) where \(\Sigma = 0.5I_{50000}I_{T_{50000}} + 0.5I_{50000}\). The running time of the direct computation for \(M^{-1}\) is 5.4416 \(\times\) \(10^4\) (s), while our approximate approach with \(q = 224\) \((\approx \sqrt{50000})\) took 2.5452 \(\times\) \(10^3\) seconds. Moreover, the errors are \(e_f = 0.9744\) and \(e_2 = 2.4798\), respectively.

### 6.2 The Matrix Ridge Approximation for Spectral Clustering

The matrix ridge approximation (RA) with the EM iteration has potentially wide applications in those methods who involve the inversion or SD of a large-scale positive semidefinite matrix. In this section we apply RA to spectral clustering.

Spectral clustering (Shi and Malik, 2000, Ng et al., 2001) is a method for partitioning data into classes by relaxing an intractable partitioning problem into a tractable eigenvector problem, specifically a problem that can be reduced to the eigenvector problem in (5) for a particular matrix \(M\) (Zhang and Jordan, 2008). The solution of the relaxation is then “rounded” to yield a partition, where standard rounding methods include \(K\)-means and Procrustes analysis (Zhang and Jordan, 2008).

In the following experiments, we used the EM-based RA methods to solve the eigenvector relaxation associated with spectral clustering, and compared the results with the conventional direct spectral decomposition (SD) method. We also implemented the \(K\)-means and Procrustean transformation (PT) rounding algorithms given in Zhang and Jordan (2008) to obtain complete spectral clustering algorithms. This yields four spectral clustering algorithms, which we refer to as RA-KM, SD-KM, RA-PT and SD-PT.

Assume we are given a dataset \(\{x_1, \ldots, x_m\}\). We defined \(M\) as a kernel matrix \(K\) via the RBF kernel with single parameter \(\beta\), i.e., \([K]_{ij} = K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\beta)\). Let \(P = I_m - \frac{1}{m}1_m1_m^T\) (i.e., \(b = \frac{1}{\sqrt{m}}1_m\)). We then formed the \(m \times m\) matrix \(T = PKP\), whose top \(q\) eigenvectors are the solution of the eigenvector problem in (5). That is, the top \(q\) eigenvectors of \(T\) are just the eigenvector relaxation associated with spectral clustering. Recall that \(\hat{A}(\hat{A}^T\hat{A})^{-\frac{1}{2}} = U_qV = \hat{X}\), which implies that RA-KM and RA-PT employ the EM-based RA to find such \(q\) eigenvectors. However, SD-KM and SD-PT employ the standard direct SD to obtain the \(q\) eigenvectors.

We conducted the experiments on eight publicly available datasets from the UCI Machine Learning Repository: the dermatology data, the soybean data, the “A-J” letter data, the image segmentation data, the NIST optical handwritten digit data, the CTG (Cardiotocograms) data, the pen-based recognition of handwritten digits data, and the Statlog (Landsat Satellite) data. Table 6 gives a summary of these datasets.

In the clustering setup, \(q+1\) is the number of classes. We initialized \(K\)-means by the orthogonal initialization method in Ng et al. (2001) and the Procrustean transformation by
The values of $\beta$ that we used are given in the last row of Table 3; they were set to be empirically optimal for these algorithms.

Table 3: Summary of the benchmark datasets: $m$—# of samples; $p$—# of features; $q+1$—# of classes; $\beta$—parameter in the kernel function $K(\cdot, \cdot)$.

|       | Derma | Soybean | Letter | CTG | Segmen | NIST | Landsat | Pen |
|-------|-------|---------|--------|-----|--------|------|---------|-----|
| $m$   | 358   | 630     | 1978   | 2126| 2310   | 3823 | 4435    | 7494|
| $p$   | 34    | 35      | 16     | 23  | 18     | 59   | 36      | 16  |
| $q+1$ | 6     | 19      | 10     | 10  | 7      | 7    | 7       | 10  |
| $\beta$ | 100   | 100     | 100    | 100 | 100    | 100  | 100     | 100 |

To evaluate the performance of the various clustering algorithms, we employed the Rand index (RI) (Rand, 1971). Given a set of $m$ samples $X = \{x_1, \ldots, x_m\}$, suppose that $U = \{U_1, \ldots, U_r\}$ and $V = \{V_1, \ldots, V_s\}$ are two different partitions of the samples in $X$ such that $\bigcup_{i=1}^r U_i = X = \bigcup_{j=1}^s V_j$ and $U_i \cap U_{i'} = \emptyset = V_j \cap V_{j'}$ for $i \neq i'$ and $j \neq j'$. Let $a$ be the number of pairs of samples that are in the same set in $U$ and in the same set in $V$, and $b$ the number of pairs of samples that are in different sets in $U$ and in different sets in $V$. The RI is given by $RI = \frac{a + b}{m(m-1)/2}$. If $RI = 1$, the two partitions are identical. Since the true partitions are available for our datasets, we calculated the RI between the true partition and the partition obtained from each clustering algorithm.

We conducted 50 replicates of each of those algorithms with K-means rounding because of the random initialization required by K-means (this is not necessary for the Procrustean transformation, because it is initialized to the identity matrix). The results shown in Table 4 are based on the average of these 50 realizations.

From Table 4 we see that the clustering methods based on RA and SD have the almost same clustering performance. In Table 5 we reported the CPU times of the direct SD method and the EM-based RA method for computing the top $q$ eigenvectors. We see that RA method can be significantly more efficient than the SD method for large $m$, and this is borne out by our results. For example, on pen-based recognition of handwritten digits data ($m = 7494$), the direct SD method takes twenty two minutes, the EM-based RA method only needs about four minutes.
Table 5: CPU times (s) of running the spectral relaxation with the direct SD and EM-based ridge approximation (RA) which are performed in Matlab on a Core 2 Duo computer with a 2.27 GHz CPU and 8 GB of RAM.

|         | SD         | RA         |
|---------|------------|------------|
| Derma   | 0.2202     | 0.1862     |
| Soybean | 0.9148     | 0.7485     |
| Letter  | 28.3251    | 7.8297     |
| CTG     | 32.9942    | 9.4303     |
| Segment | 35.2645    | 11.1811    |
| NIST    | 212.1458   | 42.7599    |
| Landsat | 289.9903   | 64.054     |
| Pen     | 1361.8     | 266.9177   |

6.3 The Matrix Ridge Approximation for GPR

In this section we applied the matrix ridge approximation with the EM iteration to Gaussian process regression (GPR), and compared with the Nyström method (Williams and Seeger, 2001) and the incomplete Cholesky decomposition method (Fine et al., 2001).

Assume we are given a training dataset \( D = \{ (x_1, y_1), \ldots, (x_m, y_m) \} \), where the \( x_i \in \mathbb{R}^p \) are the input vectors and \( y_i \in \mathbb{R} \) are the corresponding outputs. In the GPR model \( y \) is defined as

\[
y = u + f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2),
\]

where \( f(x) \) follows a Gaussian process with mean function 0 and covariance function \( K(\cdot, \cdot) \). This implies that \( f = (f(x_1), f(x_2), \ldots, f(x_m))^T \), corresponding outputs of the input vectors in the training dataset \( D \), has multivariate Gaussian distribution \( N(0, K) \), where \( K \) is the \( m \times m \) covariance matrix with \( [K]_{ij} = K(x_i, x_j) \).

We employed the Gaussian RBF kernel function \( K(\cdot, \cdot) \) with a separate length-scale parameter for each variate of the input vector, plus the signal and noise variance parameters \( \sigma_f^2 \) and \( \sigma^2 \). These parameters are trained by optimizing the marginal likelihood on a subset of the training data. Here we ignored the learning details and directly used the code provided by Rasmussen and Williams (2006) to implement the training. We concentrated our attention on the test procere.

For a test input vector \( x_* \), the prediction of the corresponding output \( y_* \) is based on the conditional posterior distribution \( p(y_*|y) \), which is also Gaussian. In particular, the predicted mean at \( x_* \) is given by

\[
\hat{y}_* = k^T(x_*)(K + \sigma^2 I_m)^{-1}y,
\]

where \( y = (y_1, \ldots, y_m)^T \) and \( k(x_*) = (K(x_*, x_1), \ldots, K(x_*, x_m))^T \) (see, Rasmussen and Williams, 2006). As we can see, GPR requires us to compute the inverse of \( K + \sigma^2 I_m \), which is an \( m \times m \) positive definite matrix. When the size \( m \) of the training dataset is very large, this limits the efficient application of GPR.

Since \( K + \sigma^2 I_m \) has the explicit form mentioned in Section 4, Williams and Seeger (2001) considered the Nyström approximation for its inverse when \( m \) is large. The Nyström method randomly chooses \( q \) columns of \( K \) without replacement. Let \( K_{m,q} \) denote the \( m \times q \) matrix consisting of such \( q \) columns. Then the Nyström approximation of \( K \) is \( K_{m,q}K_{q,q}^{-1}K_{m,q}^T \). Here we also compared the approximate method based on the incomplete Cholesky decomposition; that is, we first implemented the incomplete Cholesky decomposition of \( K \) as \( K \approx LL^T \) where \( L \) is an \( m \times q \) lower triangular matrix. After having obtained the
Table 6: Summary of the datasets: $m$—# of training samples; $n$—# of test samples; $p$—# of features,

|       | Housing | CCS | CMC | Abalone | Sat | SARCOS | YPMSD1 | YPMSD2 |
|-------|---------|-----|-----|---------|-----|--------|--------|--------|
| $m$   | 455     | 700 | 1,000 | 3,133   | 4,435 | 5,000  | 60,000 | 100,000 |
| $n$   | 51      | 330 | 473  | 1,044   | 2,000 | 4,449  | 455,345 | 415,345 |
| $p$   | 13      | 8   | 9    | 8       | 36   | 21     | 90     | 90     |

Nyström approximation or the incomplete Cholesky decomposition, we then computed $(K_{m,q}K_{q,q}^{-1}K_{m,q}^T+\sigma^2I_m)^{-1}$ or $(LL^T+\sigma^2)^{-1}I_m$ via the Sherman-Morrison-Woodbury formula. Recall that we applied the ridge approximation directly on $K+\sigma^2I_m$, rather than on $K$.

We conducted the experiments on seven publicly available datasets from the UCI Machine Learning Repository: the Boston Housing data, the Concrete Compressive Strength (CCS) data, the Contraceptive Method Choice (CMC) data, the Abalone data, the Landsat Satellite (Sat) data, the SARCOS data, and the YearPredictionMSD (YPMSD) data. We employed the setting given in the UCI Machine Learning Repository for training and testing for the first six datasets. For the YPMSD data, we employed two settings for training and testing. In the first setting (YPMSD1) we used the first 60,000 samples for training and the rest of the samples for testing, while in the second setting (YPMSD2) we used the first 100,000 samples for training and the rest of the samples for testing. Table 6 gives a summary of these datasets.

We evaluated the performance of predictions using the standardized mean squared error (SMSE) (Rasmussen and Williams, 2006). We set the rank of the matrix $A$ in the ridge approximation, the columns of the matrix $L(K \approx LL^T)$ in the incomplete Cholesky decomposition, and the columns uniformly sampled from the original kernel matrix in the Nyström method to the same number $q$. We then compared the performance of the three methods.

For the Nyström method, for each given $q$, we repeated the experiment 50 times. We found that the results are very sensitive to the columns randomly selected. The method works well in a few instances, but in most cases its performance is extremely poor. Thus, given $q$, we reported the smallest SMSE for the Nyström method.

Figure 2 shows SMSE values over the first six datasets. It should be worth pointing out that the performance of the Nyström method is very poor on the Sat and CCS datasets. Thus, we omitted the SMSE values on the two datasets for the Nyström method. Also, when $q$ is less than 4396 for the Sat dataset, the performance of the incomplete Cholesky decomposition method is poor. For this reason, we also omitted the SMSE values for the incomplete Cholesky decomposition method on the Sat dataset.

From Figure 2, we see that the performance of the ridge approximation method is nearly the same as that of standard GPR. Moreover, the ridge approximation is not sensitive to the value of $q$. For a wide range of $q$, the GPR prediction varies very little. When $q$ takes a small value, the ridge approximation still works well. Contrarily, when $q$ is small, the incomplete Cholesky decomposition is not very effective, because it results in an underfitting problem in which $LL^T+\sigma^2I_m$ is ill-conditioned. However, the ridge approximation can avoid this
Figure 2: Comparisons of the ridge approximation method, the Nyström method and the incomplete Cholesky decomposition method.
problem, because it makes $\mathbf{AA}^T+\delta\mathbf{I}_m$ better-conditioned than $\mathbf{K}$ itself (see the discussion in Section 6.1).

For the YPMSD data, we did not include the results with the Nyström method and the incomplete Cholesky decomposition method because the performance of these two methods is very poor when $q \ll m$ (e.g., $q \leq \sqrt{m}$). We only took $q = 245$ for YPMSD1 and $q = 316$ for YPMSD2 ($\approx \sqrt{m}$) to implement the ridge approximation method. Since the size of the matrix $\mathbf{K}+\sigma^2\mathbf{I}_m$ is too large, we partitioned it into the $2\times2$ block submatrices in the direct computation of $(\mathbf{K}+\sigma^2\mathbf{I}_m)^{-1}$. Although this does not reduce the computational cost, it can make the computation more numerically stable. To reduce the storage space of data, all computations were carried out with single precision in Matlab. However, we still could not complete the experiment with the direct computation method on the YPMSD2 dataset due to limited storage space.

The SMSE values with the direct method and the ridge approximation method for computing $(\mathbf{K}+\sigma^2\mathbf{I}_m)^{-1}$ on YPMSD1 are $3.7713\times10^{-5}$ and $2.9232\times10^{-5}$, respectively. We see that the ridge approximation slightly outperforms the direct computation. We hypothesize that this phenomenon is a result of roundoff error in the floating point computations. The SMSE value with the ridge approximation on YPMSD2 is $2.8959\times10^{-5}$. Therefore, the ridge approximation method is effective.

Finally, in Table 7 we report the running times with our matrix ridge approximation and the direct calculation for $(\mathbf{K}+\sigma^2\mathbf{I}_m)^{-1}$ on the datasets. The reported results with our method are based on that $q$ is taken as the integer closest to $\sqrt{m}$. We see that our method is able to reduce computation when $m$ is vary large. For example, on the YPMSD1 dataset the direct computation took $1.2436\times10^5$ seconds, while the ridge approximation took $5.814\times10^3$ seconds. In summary, our proposed approach is efficient and effective.

|                  | YPMSD1 | YPMSD2 |
|------------------|--------|--------|
| Direct           | 1.2436×10^5 | NA     |
| EM-RA            | 5.814×10^3  | 1.0664×10^6 |

Table 7: CPU times (s) of running the test procedure of GPR with the direct computation and EM-based ridge approximation (RA) which are performed in Matlab on a Workstation with a 3.07 GHz CPU and 24 GB of RAM.

7. Conclusion

In this paper we have proposed the matrix ridge approximation method, which tries to find an approximation for a symmetric positive semidefinite matrix. We have also developed probabilistic formulations for this method. The probabilistic formulation not only provides a statistical interpretation but also leads us to an efficient EM iterative procedure for the matrix ridge approximation. The matrix ridge approximation with the EM iteration has potentially broad applicability in machine learning problems that involve the inversion or spectral decomposition of a large-scale positive semidefinite matrix. In particular, we have empirically illustrated the effectiveness and efficiency of the matrix ridge approximation in the case of spectral clustering and Gaussian process regression.
The support vector machine (SVM) and Gaussian process classification (GPC) are two classical kernel classification methods. When applying them to large-scale data sets, we also meet a computational challenge. The matrix ridge approximation technique is a potentially useful approach for handling this challenge. We will study this issue in future work. Recall that each EM iteration for the matrix ridge approximation takes time \(O(m^2 q)\) and it mainly involves matrix multiplications. To make the method more efficient, we can consider the parallel implementation of the matrix multiplications.

Appendix A. Several Lemmas

In order to prove the theorems, we first present several lemmas that will be used.

**Lemma 4** Suppose \(B \in \mathbb{R}^{m \times m}\). Let \(c_i + id_i\) for \(i = 1, \ldots, m\) be the eigenvalues of \(B\) where \(i^2 = -1\) and the \(c_i, d_i \in \mathbb{R}\). Then,

(i) \(\text{tr}(B) = \sum_{i=1}^{m} c_i\) and \(\text{tr}(B^2) = \sum_{i=1}^{m} (c_i^2 - d_i^2)\).

(ii) \(\sum_{i=1}^{m} c_i^2 + d_i^2 \leq \text{tr}(BB^T)\), \(\sum_{i=1}^{m} c_i^2 \leq \frac{1}{2}\text{tr}(BB + BB^T)\), and \(\sum_{i=1}^{m} d_i^2 \leq \frac{1}{2}\text{tr}(BB^T - BB)\).

**Proof** It is obvious that \(c_i + id_i\) is the eigenvalue of \(B\) iff \(c_i - id_i\) is the eigenvalue of \(B\). Accordingly, we have Part (i).

In addition, let the Schur factorization of \(B\) be \(B = QTQ^*\) where \(Q\) is unitary and \(T\) is upper-triangular with the eigenvalues of \(B\) at the diagonals. Thus,

\[
\text{tr}(BB^T) = \text{tr}(TT^*) \geq \sum_{i=1}^{m} (c_i + id_i)(c_i - id_i) = \sum_{i=1}^{m} (c_i^2 + d_i^2).
\]

In addition, we also have

\[
\frac{1}{2}\text{tr}(BB^T + BB) = \frac{1}{4}\text{tr}((T + T^*)(T + T^*)) \geq \sum_{i=1}^{m} c_i^2
\]

and

\[
\frac{1}{2}\text{tr}(BB^T - BB) = \frac{1}{4}\text{tr}((T - T^*)(T - T^*)) \geq \sum_{i=1}^{m} d_i^2.
\]

The proof completes.

We now turn to our proposed approach and follow the notations in Table 1. Without loss of generality, we only consider the case that \(b \neq 0\) and \(b^Tb = 1\). In this case, \(P = I_m - bb^T\) is idempotent, symmetric and of rank \(m-1\). Thus we can express it as \(P = \Psi[I_{m-1} 0 0] \Psi^T\) where \(\Psi^T \Psi = \Psi \Psi^T = I_m\). Let \(\Psi_1\) be an \(m \times (m-1)\) matrix containing the first \(m-1\) columns of \(\Psi\). Then \(\Psi = [\Psi_1, b]\) so that \(\Psi_1^T \Psi_1 = I_{m-1}, \Psi_1^T b = 0\) and \(P = \Psi_1^T \Psi_1\).

In order to prove the theorems given in Section 3, we use the same notation as in Section 3. Moreover, we here and later denote \(Z = \Psi_1^T A ((m-1) \times q)\), \(G = \Psi_1^T T \Psi_1 =\)
\[ \Psi_1^T PMP \Psi_1 = \Psi_1^T M \Psi_1 \quad ((m-1) \times (m-1)) \] and \[ \Theta = \Psi_1^T \Omega \Psi_1 = \Psi_1^T (A A^T + \delta I_m) \Psi_1 = Z Z^T + \delta I_{m-1} \quad ((m-1) \times (m-1)). \] With these notations, we present the following several lemmas.

**Lemma 5** Let \( \lambda(C) \) be the set of the all eigenvalues of \( C \). Then \( \lambda(T) = \lambda(G) \cup \{0\} \). Furthermore, if \( \phi \) is the eigenvector of \( G \) associated with its eigenvalue \( \gamma \), then \( \Psi_1 \phi \) is the eigenvector of \( T \) associated with its eigenvalue \( \gamma \). Conversely, if \( u \) satisfying \( u^T b = 0 \) is the eigenvector of \( T \) associated with its eigenvalue \( \gamma \), then \( \Psi_1^T u \) is the eigenvector of \( G \) associated with its eigenvalue \( \gamma \).

**Proof** Recall that

\[ T = PTP = \Psi \begin{bmatrix} \Psi_1^T T \Psi_1 & 0 \\ 0 & 0 \end{bmatrix} \Psi^T = \Psi \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \Psi^T. \]

Thus, \( \lambda(T) = \lambda(G) \cup \{0\} \). Letting \( G \phi = \gamma \phi \), we have

\[ S \Psi_1 \phi = \Psi \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \Psi^T \Psi_1 \phi = \Psi \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{m-1} \\ 0 \end{bmatrix} \phi = \Psi \begin{bmatrix} I_{m-1} \\ 0 \end{bmatrix} G \phi = \Psi \begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \Psi^T \phi = \gamma \Psi_1 \phi, \]

which shows that \( \Psi_1 \phi \) is the eigenvectors of \( T \). Also, since

\[ G \Psi_1^T u = \Psi_1^T \Psi_1 \Psi_1^T u = \Psi_1^T T u = \gamma \Psi_1^T u \]

\( \Psi_1^T u \) is the eigenvector of \( G \) associated with its eigenvalue \( \gamma \).

**Lemma 6** Assume that \( k \) is an arbitrary integer. Then,

(i) \( \Psi_1^T \Omega^{-k} \Psi_1 = \Theta^{-k} \), \( b^T \Omega^{-k} \Psi_1 = 0 \) and \( b^T \Omega^{-k} b = \delta^{-k} \);

(ii) \( \text{tr}(\Omega^{-k}) = \text{tr}(\Theta^{-k}) + \delta^{-k} \).

**Proof** As for (i), we first have

\[ \Psi_1^T \Omega^{-1} \Psi_1 = \Psi_1^T (A A^T + \delta I_m)^{-1} \Psi_1 = \Psi_1^T (\delta^{-1} I_{m-1} - \delta^{-1} Z (\delta I_q + Z^T Z)^{-1} Z) \Psi_1 = \Theta^{-1}, \]

due to \( A^T A = A^T P A = A^T \Psi_1 \Psi_1^T A = Z^T Z \). Assume that \( \Psi_1^T \Omega^{-l} \Psi_1 = \Theta^{-l} \) for some positive integer \( l \). Then

\[ \Psi_1^T \Omega^{-l} \Psi_1 = \Psi_1^T \Omega^{-l} (\Psi_1 \Psi_1^T + b b^T) \Omega^{-1} \Psi_1 = \Theta^{-l} + \Psi_1^T \Omega^{-l} b b^T \Omega^{-1} \Psi_1 = \Theta^{-l} \]

due to \( b^T \Omega^{-k} \Psi_1 = 0 \). Thus, we obtain \( \Psi_1^T \Omega^{-k} \Psi_1 = \Theta^{-k} \) by the induction. Similarly, we

\[ b^T \Omega^{-k} \Psi_1 = 0 \] and \( b^T \Omega^{-k} b = \delta^{-k} \). Finally, it follows from (i) that

\[ \text{tr}(\Omega^{-k}) = \text{tr} \left( \left[ \begin{array}{c} \Psi_1^T \\ b^T \end{array} \right] \Omega^{-k} [\Psi_1, b] \right) = \text{tr}(\Theta^{-k}) + \delta^{-k}. \]
Appendix B. Proof for Theorem 1

In order to prove Theorem 1, we present a more general alternative which is based on two variants of \(F\) and \(G\). In particular, the first variant is

\[
F_1(A, \delta) = \|M - AA^T - \delta I_m\|_F^2 = \text{tr}((M - AA^T - \delta I_m)^2),
\]

while the second variant is

\[
G_1(A, \delta) = \log |AA^T + \delta I_m| + \text{tr}((AA^T + \delta I_m)^{-1}M).
\]

Obviously, \(F_1\) and \(F\) (or \(G_1\) and \(G\)) become identical when \(b = 0\). The minimizers of \(F_1\) as well as \(G_1\) are given in the following theorem.

**Theorem 7** Let \(\gamma_1 \geq \cdots \geq \gamma_q \geq \cdots \geq \gamma_m \geq 0\) be the eigenvalues of \(T = \text{PMP}\), \(V\) be an arbitrary \(q \times q\) orthogonal matrix, \(\Gamma_q\) be a \(q \times q\) diagonal matrix containing the first \(q\) principal (largest) eigenvalues \(\gamma_i\), and \(U_q\) be an \(n \times q\) column-orthonormal matrix in which the \(q\) column vectors are the principal eigenvectors corresponding to \(\Gamma_q\). Assume that \(\delta > 0\) and that \(A \in \mathbb{R}^{m \times q}\) (\(q < \min(m, p)\)) is of full column rank and satisfies \(A^Tb = 0\). If the following conditions are satisfied

\[
\gamma_i > \frac{1}{m-q} \left(b^T Mb + \sum_{j=q+1}^m \gamma_j\right), \quad \text{for } i = 1, \ldots, q, \quad (8)
\]

then the strict local minimum of \(F_1(A, \delta)\) and \(G_1(A, \delta)\) w.r.t. \((A, \delta)\) are respectively obtained when

\[
\hat{A} = U_q(\Gamma_q - \delta I_q)^{1/2}V \quad \text{and} \quad \hat{\delta} = \frac{1}{m-q} \left(b^T Mb + \sum_{j=q+1}^m \gamma_j\right).
\]

Note that \(\text{PSP} = S\) and \(b^T S b = 0\). Thus, when viewing \(S\) as \(M\) in Theorem 7, we immediately obtain Theorem 1 from Theorem 7.

Theorem 7 shows the connection between the estimates of \(A\) and \(\delta\) based on the minimizations of \(F_1\) and \(G_1\). In particular, the estimates of \(A\) and \(\delta\) via minimizing \(F_1\) are equivalent to those of \(A\) and \(\delta\) via minimizing \(G_1\). We note that the minimizer \((\hat{A}, \hat{\delta})\) of \(G_1\) under \(b = 0\) was given in Magnus and Neudecker (1999). The conditions in (8) aim to ensure that \((\Gamma_q - \delta I_q)^{1/2}\) exists and \(\hat{A}\) is of full column rank. In the case that \(b = 0\), \(\gamma_q > \gamma_{q+1}\) suffices for the conditions. In fact, they are always satisfied whenever there is at least one \(\gamma_j\) where \(j \in \{q+1, \ldots, m\}\) such that \(\gamma_q > \gamma_j > 0\). Thus, the conditions in (8) are trivial when \(b = 0\).

However, the conditions are not always satisfied when \(b^T b = 1\). For example, let

\[
M = \begin{bmatrix} b & \Psi_1 \end{bmatrix} \begin{bmatrix} 1 + \alpha^2 & 0 \\ 0 & \alpha^2 \end{bmatrix}^{-1} \begin{bmatrix} b^T \\ \Psi_1^T \end{bmatrix}
\]

for \(\alpha \neq 0\) such that \(\Psi_1^T b = 0\) and \(\Psi_1^T \Psi_1 = I_{m-1}\). It is clear that \(b^T Mb = 1 + \alpha^2\) and \(T = \text{PMP} = \alpha^2 \text{P}\). This implies that the eigenvalues \(\gamma_i\) of \(T\) are \(\alpha^2\) with multiplicity \(m-1\) and 0 with multiplicity 1. As a result, for any \(i \leq q < m\), we always have

\[
\gamma_i = \alpha^2 < \alpha^2 + 1/(m-q) = \frac{1}{m-q} \left(b^T Mb + \sum_{j=q+1}^m \gamma_j\right).
\]
Thus, the condition in (8) is not satisfied. Consequently, this condition would limit the use of $F_1$ and $G_1$ in the matrix ridge approximation. This is the reason why we employ $F$ and $G$ instead of $F_1$ and $G_1$ respectively.

B.1 Proof for the Minimizer of $F_1(A, \delta)$ w.r.t. $(A, \delta)$

Consider the Lagrangian function of

$$L = \text{tr}(M - AA^T - \delta I_m)^2 + 4b^T A a$$

where $a$ is a $q \times 1$ vector of Lagrangian multipliers. We now compute

$$dL = -2\text{tr}[(M - AA^T - \delta I_m)((dA)A^T + A(dA^T)) + 4b^T (dA)a$$

$$= -4\text{tr}(A^T(M - AA^T - \delta I_m)(dA)) + 4b^T (dA)a,$$

$$dL = -2\text{tr}(M - AA^T - \delta I_m)d\delta.$$ 

Using the first-order condition, we obtain

$$-A^T(M - AA^T - \delta I_m) + ab^T = 0,$$

$$\text{tr}(M - AA^T - \delta I_m) = 0.$$ 

Postmultiplying the above first equation by $b$, we obtain $a = A^T Mb$ because of $A^T b = 0$. As a result, we have

$$TA = A(A^T A + \delta I_q),$$

where $T = PMP$. Assume the spectral decomposition of $A^T A$ as $A^T A = V \Lambda V^T$. Hence,

$$TAVA^{-\frac{1}{2}} = AVA^{-\frac{1}{2}}(\Lambda + \delta I_q).$$

This implies that the diagonal elements of $\Lambda + \delta I_q$ are the $q$ eigenvalues of $T$, and $AVA^{-\frac{1}{2}}$ is a corresponding matrix of orthonormal eigenvectors. This motivates us to define $\Gamma_q = \Lambda + \delta I_q$ and $U_q = AVA^{-\frac{1}{2}}$. That is, $\Lambda = \Gamma_q - \delta I_q$ and $\hat{A} = U_q A^{\frac{1}{2}} V^T$.

On the other hand, we have

$$\text{tr}(M) = \text{tr}(\Psi M \Psi) = \text{tr}(\Psi_1^T M \Psi_1) + b^T M b$$

and $\text{tr}(T) = \text{tr}(PMP) = \text{tr}(MP) = \text{tr}(M \Psi_1 \Psi_1^T) = \text{tr}(\Psi_1^T M \Psi_1)$. It then follows from $\text{tr}(M - AA^T - \delta I_m) = 0$ that

$$m\delta + \sum_{i=1}^{q} \gamma_i - q\delta = \sum_{i=1}^{m} \gamma_i + b^T M b.$$ 

Thus we let $\delta = \frac{1}{m-q}(\sum_{i=q+1}^{m} \gamma_i + b^T M b)$. Condition 8 shows that $A^{\frac{1}{2}} = (\Gamma_q - \delta I_q)^{\frac{1}{2}}$ exists and $\hat{A} = U_q A^{\frac{1}{2}} V^T$ is of full column rank.
To verify that \((\hat{A}, \delta)\) is a minimizer of \(F_1(A, \delta)\), we compute the Hessian matrix of \(L\) w.r.t. to \((A, W)\). Let \(\text{vec}(A) = (y_{11}, \ldots, y_{m1}, y_{12}, \ldots, y_{mq})^T\). The Hessian matrix is then given by

\[
H(A, \delta) \triangleq \begin{bmatrix}
\frac{\partial^2 L}{\partial \text{vec}(A) \partial \text{vec}(A)^T} & \frac{\partial^2 L}{\partial \text{vec}(A) \partial \delta^T} \\
\frac{\partial^2 L}{\partial \delta \partial \text{vec}(A)^T} & \frac{\partial^2 L}{\partial \delta^T}(\delta)
\end{bmatrix} = 4 \begin{bmatrix}
I_q \otimes (\delta I_m + AA^T - M) + A^T A \otimes I_m + (A^T \otimes A) C_{mq} \text{vec}(A) & \text{vec}(A)
\end{bmatrix},
\]

where \(C_{mq}\) is the \(mq \times mq\) commutation such that \(C_{mq} \text{vec}(B) = \text{vec}(B^T)\) for any \(m \times q\) matrix \(B\).

Let \(X\) be an arbitrary nonzero \(m \times q\) matrix such that \(X^T b = 0\), and \(a\) be a nonzero real number. Hence,

\[
\zeta \triangleq \frac{1}{4} [\text{vec}(X)^T, a] H(\hat{A}, \hat{\delta}) [\text{vec}(X)^T, a]^T
\]

\[
= \text{tr}(X^T (\hat{\delta} I_m + \hat{A} \hat{A}^T - M) X) + \text{tr}(X^T X \hat{A} \hat{A}^T) + 2 \text{atr}(X \hat{A}^T) + \frac{m}{2} a^2
\]

\[
= \text{tr}(X^T (\hat{\delta} I_m + \hat{A} \hat{A}^T - T) X) + \text{tr}(X^T X \hat{A} \hat{A}^T) + 2 \text{atr}(X \hat{A}^T) + \frac{m}{2} a^2
\]

due to \(X^T M X = X^T T X\).

Let \(T = UTU^T\) where \(U = [U_q, U_2]\) and \(T = \text{diag}(\Gamma_q, \Gamma_2)\) such that \(U_2^T U_q = 0\), \(U_2^T U_2 = I_{m-q}\) and \(\Gamma_2 = \text{diag}(\gamma_{q+1}, \ldots, \gamma_m)\). Thus,

\[
\hat{\delta} I_m + \hat{A} \hat{A}^T - T = [U_q, U_2] \begin{bmatrix} 0 & 0 \\ 0 & \hat{\delta} I_{m-q} - \Gamma_2 \end{bmatrix} \begin{bmatrix} U_q^T \\ U_2^T \end{bmatrix} = U_2 (\hat{\delta} I_{m-q} - \Gamma_2) U_2^T.
\]

Furthermore, we have \(\text{tr}(X \hat{A}^T) = \text{tr}(B_q A^{\frac{1}{2}})\),

\[
\text{tr}(X^T (\hat{\delta} I_m + \hat{A} \hat{A}^T - T) X) = \text{tr}(B_q^T (\hat{\delta} I_{m-q} - \Gamma_2) B_2).
\]

\[
\text{tr}(X^T X \hat{A} \hat{A}^T) = \text{tr}(U_q^T X V A^{\frac{1}{2}} U_q^T X V A^{\frac{1}{2}}) = \text{tr}(B_q A^{\frac{1}{2}} B_q A^{\frac{1}{2}})\]

\[
\text{tr}(X^T X \hat{A} \hat{A}^T) = \text{tr}(V^T X^T U U^T X V A) = \text{tr}(B^T B) = \text{tr}(B_q^T B_q A) + \text{tr}(B_2^T B_2 A)
\]

where \(B_q = U_q^T X V (q \times q), B_2 = U_2^T X V ((m-q) \times q), \) and \(B = U^T X V = [B_q^T, B_2^T]^T = [b_1, \ldots, b_m]^T (m \times q)\). Accordingly, we obtain

\[
\zeta = \text{tr}(B_q^T (\hat{\delta} I_{m-q} - \Gamma_2) B_2) + \text{tr}(B_q^T B_q A) + \text{tr}(B_q A^{\frac{1}{2}} B_q A^{\frac{1}{2}}) + 2 \text{atr}(B_q A^{\frac{1}{2}}) + \frac{m}{2} a^2.
\]

Recall that \(A = \Gamma_q \hat{\delta} I_q\). It is easily verified that \(\text{tr}(B_q^T (\hat{\delta} I_{m-q} - \Gamma_2) B_2) + \text{tr}(B_q^T B_q A) \geq 0\).

In addition, let the real parts of the eigenvalues of \(B_q A^{\frac{1}{2}}\) be \(\eta_i\) for \(i = 1, \ldots, q\). It follows
from Lemma 4 that
\[
\begin{align*}
\text{tr}(B_q^T B_q A) + \text{tr}(B_q A^2 B_q A^T) + 2\text{tr}(B_q A^T) + \frac{m}{2} a^2 \\
\geq \sum_{i=1}^{q} [2\eta_i^2 + 2a\eta_i] = \sum_{i=1}^{q} \frac{1}{2} (2\eta_i + a)^2 + \frac{m - q}{2} a^2 \\
> 0.
\end{align*}
\]

In summary, we obtain \([\text{vec}(X^T)^T, a]H(\hat{A}, \hat{\delta})[\text{vec}(X^T)^T, a]^T > 0\). Thus, this implies that \((\hat{A}, \hat{\delta})\) is the strict local minimizer of \(F_1(A, \delta)\).

Replacing \(S\) for \(M\) in \(F_1(A, \delta)\) and considering \(\text{PSP} = S\), we immediately obtain the strict local minimizer of \(F(A, \delta)\). In this case, we have \(\hat{\delta} = \frac{1}{m-q} \sum_{i=q+1}^{m} \gamma_i\) due to \(S^T b = 0\).

**B.2 Proof for the Minimizer of \(G_1(A, \delta)\) w.r.t. \((A, \delta)\)**

To prove that the \((\hat{A}, \hat{\delta})\) is also the minimizer of \(G_1(A, \delta)\), we consider the following the Lagrangian function:

\[
L(A, \delta) = \log |\Omega| + \text{tr}(\Omega^{-1} M) + 2b^T A a
\]

where \(a\) is the \(q \times 1\) vector of Lagrangian multipliers. We have

\[
dL = \text{tr}(\Omega^{-1}(d\Omega)) - \text{tr}(M \Omega^{-1}(d\Omega) \Omega^{-1}) + 2b^T (dA)a \\
= \text{tr}(\Omega^{-1}((dA)A^T + A(dA^T))) - \text{tr}(\Omega^{-1} M \Omega^{-1}((dA)A^T + A(dA^T))) + 2b^T (dA)a \\
= 2\text{tr}(A^T \Omega^{-1}(dA)) - 2\text{tr}(A^T \Omega^{-1} M \Omega^{-1}(dA)) + 2b^T (dA)a
\]

Then, using the first-order condition, we have \(\text{tr}(\Omega^{-1}) - \text{tr}(\Omega^{-1} M \Omega^{-1}) = 0\) and

\[
A^T \Omega^{-1} - A^T \Omega^{-1} M \Omega^{-1} + ab^T = 0.
\]

Postmultiplying the above equation by \(b\), we obtain \(a = (A^T \Omega^{-1} M \Omega^{-1} - A^T \Omega^{-1}) b\). As a result, we have the first-order condition for \(A\) as

\[
A^T \Omega^{-1} P = A^T \Omega^{-1} M \Omega^{-1} P,
\]

which is equivalent to that

\[
A^T \Psi_1 \Psi_1^T \Omega^{-1} \Psi_1 \Psi_1^T = A^T \Psi_1 \Psi_1^T \Omega^{-1} (\Psi_1 \Psi_1^T + bb^T) M (\Psi_1 \Psi_1^T + bb^T) \Omega^{-1} \Psi_1 \Psi_1^T = A^T \Psi_1 \Psi_1^T \Omega^{-1} \Psi_1 \Psi_1^T T \Psi_1 \Psi_1^T \Omega^{-1} \Psi_1 \Psi_1^T
\]

due to \(P = \Psi_1 \Psi_1^T\), \(T = \text{PMP}\), \(A^T P = A^T\) and \(b^T \Omega^{-1} \Psi_1 = 0\). We thus obtain

\[
\Psi_1^T T \Psi_1 \Psi_1^T \Omega^{-1} \Psi_1 Z = Z,
\]

where \(Z = \Psi_1^T A\). According to Lemma 6, the first-order condition for \(A\) becomes

\[
G \Theta^{-1} Z = Z.
\]
where $G = \Psi_1^T T \Psi = \Psi_1^T P M \Psi = \Psi_1^T M \Psi$. In addition, from Lemma 6, we have

\[
\text{tr}(\Omega^{-1} M \Omega^{-1}) = \text{tr}
\begin{pmatrix}
\Psi_1^T \\
\Psi_1 T M \Psi_1 \\
\Psi_1 T M \Psi_1 + \delta - I
\end{pmatrix}
\begin{pmatrix}
\Theta^{-1} \\
\Psi_1 T M \Psi_1 \\
0 & \delta^{-1}
\end{pmatrix}
\begin{pmatrix}
\Theta^{-1} \\
\Psi_1 T M \Psi_1 \\
0 & \delta^{-1}
\end{pmatrix}
\]

\[
= \text{tr}(\Theta^{-1} \Psi_1^T M \Psi \Theta^{-1}) + \delta^2 b^T M b.
\]

The first-order condition for $\delta$ thus becomes

\[
\delta^2 \left[ \text{tr}(\Theta^{-1}) - \text{tr}(\Theta^{-1} G \Theta^{-1}) \right] + \delta - b^T M b = 0.
\]  

(10)

It follows from $ZZ^T = \Theta - \delta I_{m-1}$ that

\[
ZZ^T Z = G \Theta^{-1} ZZ^T Z = G Z - \delta Z,
\]

which yields

\[
G Z = Z (\delta I_q + Z^T Z).
\]  

(11)

Assume that the rank of $Z$ is $q$ ($\leq m-1$). There exists a semi-orthogonal $q \times q$ matrix $V$ ($VV^T = I_q$) and a $q \times q$ diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q)$ such that

\[
Z^T Z = V \Lambda V^T.
\]

It is clear that $V$ and $\Lambda$ are the eigenvector matrix and eigenvalue matrix of $Z^T Z$, respectively. Then we can rewrite (11) as

\[
G Z V = Z V (\delta I_q + \Lambda)
\]

which gives

\[
G Z V \Lambda^{-1/2} = Z V \Lambda^{-1/2} (\delta I_q + \Lambda).
\]

Denote $\Phi_q = Z V \Lambda^{-1/2}$ ($(m-1) \times q$). It is easy to see $\Phi_q^T \Phi_q = I_q$. Thus, $\Phi_q$ and $\delta I_q + \Lambda$ are the eigenvector and eigenvalue matrices of $G$, respectively. This motivates us to equalize $\delta I_q + \Lambda = \Gamma_q$ and $Z V \Lambda^{-1/2} = \Psi_1^T U_q$. That is, we let $\hat{\Lambda} = U_q (\Gamma_q - \delta I_q)^{1/2} V^T$.

On the other hand, since

\[
\Theta^{-1} = \delta^{-1} I_{m-1} - \delta^{-1} Z (\delta I_q + Z^T Z)^{-1} Z^T
\]

and from (11), we have

\[
G \Theta^{-1} = \delta^{-1} G - \delta^{-1} G Z (\delta I_q + Z^T Z)^{-1} Z^T = \delta^{-1} (G - ZZ^T).
\]

Hence

\[
\delta^2 (\Theta^{-1} G \Theta^{-1} - \Theta^{-1}) = G - ZZ^T - \delta I_{m-1}.
\]

Combining this equation with (10) yields

\[
m \delta = \text{tr}(G) - \text{tr}(Z^T Z) + b^T M b.
\]
We thus set \( \hat{\delta} = \frac{1}{m-q}(b^T M b + \sum_{j=q+1}^{m} \gamma_j) \).

It is clearly seen that \( (\hat{\delta}, \hat{A}) \) satisfy the first-order conditions of \( L \) w.r.t. \( (\delta, A) \). To verify that \( (\hat{\delta}, \hat{A}) \) are the minimizer of \( g(A, \delta) \), we compute

\[
\frac{1}{2}d^2 L = \text{tr}[(dA^T)\Omega^{-1}(dA)] - \text{tr}[A^T \Omega^{-1}(dA)A^T \Omega^{-1}(dA)] - \text{tr}[A^T \Omega^{-1} A(dA^T)\Omega^{-1}(dA)]
\]

\[
- \text{tr}[(dA^T)\Omega^{-1}M\Omega^{-1}(dA)] + \text{tr}[A^T \Omega^{-1}(dA)A^T \Omega^{-1}M\Omega^{-1}(dA)]
\]

\[
+ \text{tr}[A^T \Omega^{-1} A(dA^T)\Omega^{-1}\Omega^{-1}(dA)] + \text{tr}[A^T \Omega^{-1}M\Omega^{-1}(dA)A^T \Omega^{-1}(dA)]
\]

\[
+ \text{tr}[A^T \Omega^{-1}M\Omega^{-1} A(dA^T)\Omega^{-1}(dA)]
\]

\[
\frac{1}{2}d^2 L = -\text{tr}[A^T \Omega^{-2}(dA)](d\delta) + \text{tr}[A^T \Omega^{-1}M\Omega^{-2}(dA)](d\delta) + \text{tr}[A^T \Omega^{-2}M\Omega^{-1}(dA)](d\delta),
\]

\[
\frac{1}{2}d^2 L = -\frac{1}{2}\text{tr}[\Omega^{-2}](d\delta)(d\delta) + \text{tr}[\Omega^{-3}M](d\delta)(d\delta).
\]

We thus have the Hessian matrix:

\[
H(A, \delta) \triangleq \begin{bmatrix}
\frac{\partial^2 L}{\partial \text{vec}(A)^T \partial \text{vec}(A)^T} & \frac{\partial^2 L}{\partial \text{vec}(A)^T \partial \delta} \\
\frac{\partial^2 L}{\partial \delta \partial \text{vec}(A)^T} & \frac{\partial^2 L}{\partial \delta^2}
\end{bmatrix}
\]

where \( \frac{1}{2} \frac{\partial^2 L}{\partial \delta^2} = \text{tr}[\Omega^{-3}M] - \frac{1}{2} \text{tr}[\Omega^{-2}], \)

\[
\frac{1}{2} \frac{\partial^2 L}{\partial \text{vec}(A)^T \partial \delta} = [I_q \otimes (\Omega^{-1}M\Omega^{-2} + \Omega^{-2}M\Omega^{-1} - \Omega^{-2})]\text{vec}(A),
\]

\[
\frac{1}{2} \frac{\partial^2 L}{\partial \text{vec}(A)^T \partial \text{vec}(A)} = [I_q - A^T \Omega^{-1} A] \otimes [\Omega^{-1} - \Omega^{-1}M\Omega^{-1}] + A^T \Omega^{-1}M\Omega^{-1} A \otimes \Omega^{-1}
\]

\[
+ C_{qm} [\Omega^{-1} A \otimes A^T \Omega^{-1}M\Omega^{-1} + \Omega^{-1}M\Omega^{-1} A \otimes A^T \Omega^{-1} - \Omega^{-1} A \otimes A^T \Omega^{-1}].
\]

Given an arbitrary nonzero matrix \( X \in \mathbb{R}^{m \times q} \) such that \( X^T b = 0 \), and a nonzero number \( a \in \mathbb{R} \), we have

\[
B \triangleq \frac{1}{2} [\text{vec}(X)^T, a]H(A, \delta)[\text{vec}(X)^T, a]^T
\]

\[
= \text{tr} [X(I_q - A^T \Omega^{-1} A)X^T (\Omega^{-1} - \Omega^{-1}M\Omega^{-1})] + \text{tr} [XA^T \Omega^{-1}M\Omega^{-1} AX^T \Omega^{-1}]
\]

\[
+ 2 \text{tr} (XA^T \Omega^{-1}XAX^T \Omega^{-1}M\Omega^{-1}) - \text{tr} [XA^T \Omega^{-1}XAX^T \Omega^{-1}]
\]

\[
+ 2 \text{tr} [X^T (\Omega^{-1}M\Omega^{-2} + \Omega^{-2}M\Omega^{-1} - \Omega^{-2}) A] + [\text{tr}(\Omega^{-3}M) - \frac{1}{2} \text{tr}(\Omega^{-2})] a^2
\]

\[
= \text{tr} [X_1(I_q - Z^T \Theta^{-1} Z)X_1^T (\Theta^{-1} - \Theta^{-1}G\Theta^{-1})] + \text{tr} [X_1 Z^T \Theta^{-1}G\Theta^{-1} ZX_1^T \Theta^{-1}]
\]

\[
+ 2 \text{tr} (X_1 Z^T \Theta^{-1}X_1 Z^T \Theta^{-1}G\Theta^{-1}) - \text{tr} [X_1 Z^T \Theta^{-1}X_1 Z^T \Theta^{-1}]
\]

\[
+ 2 \text{tr} [X_1^T (\Theta^{-1}G\Theta^{-2} + \Theta^{-2}G\Theta^{-1} - \Theta^{-2}) Z] + \frac{1}{2} [2 \text{tr}(\Omega^{-3}M) - \text{tr}(\Theta^{-2}) - \delta^{-2}] a^2
\]

\[
= \text{tr} [X_0(I_q - Z^T \Theta^{-1} Z)X_0^T (\Theta - G)] + \text{tr} [X_0 Z^T \Theta^{-1}G\Theta^{-1} ZX_0^T \Theta]
\]

\[
+ 2 \text{tr} (X_0 Z^T X_0 Z^T \Theta^{-1} G) - \text{tr} [X_0 Z^T X_0 Z^T] + 2 \text{tr} [X_0^T (G\Theta^{-2} + \Theta^{-1}G\Theta^{-1} - \Theta^{-1}) Z]
\]

\[
+ \frac{1}{2} [2 \text{tr}(\Theta^{-3} G) - \text{tr}(\Theta^{-2}) - \delta^{-2}] a^2 + \frac{\alpha^2}{\delta^3} b^T M b
\]
where \( X_1 = \Psi_1^T X \) and \( X_0 = \Theta^{-1} X_1 \). Here we use the fact that \( X = PX = \Psi_1 \Psi_1^T X \), \( b^T \Omega^{-1} \Psi_1 = 0 \), \( G = \Psi_1^T M \Psi_1 \) and \( \text{tr}(\Omega^{-3} M) = \text{tr}(\Theta^{-3} G) + \delta^3 b^T M b \).

Recall that the eigenvalues of \( G \) are also the eigenvalues of \( T \). Let \( T_2 = \text{diag}(\lambda_{q+1}, \ldots, \lambda_{m-1}) \).

We can express the SVD of \( G \) as \( G = [\Phi_q \Phi_2] \left[ \begin{array}{c|c} \Gamma_q & 0 \\ \hline 0 & \Gamma_2 \end{array} \right] \left[ \begin{array}{c} \Phi_q^T \\ \Phi_2^T \end{array} \right] = \Phi_q \Gamma_q \Phi_q^T + \Phi_2 \Gamma_2 \Phi_2^T \).

Then \( \hat{Z} = \Phi_h (\Gamma_q - \delta)^2 V^T \). Substituting \( (\hat{Z}, \hat{\delta}) \) for \( (Z, \delta) \) yields \( \hat{Z} \hat{Z}^T = \Phi_q (\Gamma_q - \delta I_q) \Phi_q^T \) and

\[
\hat{\Theta}^{-1} = (\delta I_{m-1} + \hat{Z} \hat{Z}^T)^{-1} = \delta^{-1}[\Gamma_{m-1} - \Phi_q (\Gamma_q - \delta I_q) \Gamma_q^{-1} \Phi_q^T] \delta^{-1} \Phi_2 \Phi_2^T + \Phi_q \Gamma_q^{-1} \Phi_q^T,
\]

which in turn lead to \( \hat{\Theta} = \delta \Psi_q \Phi_2^T - \Psi_2 \Gamma_2 \Phi_2^T, \hat{\Theta}^{-1} G = \delta^{-1} \Phi_2 \Gamma_2 \Phi_2^T + \Phi_q \Phi_q^T, \hat{\Theta}^{G} = \delta^{-2} \Phi_2 \Phi_2^T + \Phi_q \Gamma_q^{-2} \Phi_q^T, \hat{\Theta}^{-3} G = \delta^{-3} \Phi_2 \Gamma_2 \Phi_2^T + \Phi_q \Gamma_q^{-3} \Phi_q^T \)

and

\[
G \hat{\Theta}^{-2} + G \hat{\Theta}^{-1} \hat{\Theta}^{-1} = \frac{1}{\delta^2} \Phi_2(2 \Gamma_2 - \delta I_2) \Phi_2^T + \Phi_q \Gamma_q^{-2} \Phi_q^T.
\]

Let \( E_1 = \Phi_h X_0 V \) and \( E_2 = \Phi_2 X_0 V \). It is then obtained that

\[
B_1 \triangleq \text{tr}\left[ X_0 (I_q - \hat{Z} \hat{Z}^T \hat{\Theta}^{-1} \hat{Z}) X_0^T (\hat{\Theta} - G) \right] = \delta \text{tr}\left[ E_2 \Gamma_q^{-1} E_2^T (\delta I_{m-q-1} - \Gamma_2) \right],
\]

\[
B_2 \triangleq \text{tr}\left[ X_0 \hat{Z} \hat{Z}^T \hat{\Theta}^{-1} G \hat{Z} X_0^T \hat{\Theta} \right] = \text{tr}\left[ E_1 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_1^T \Gamma_q \right] + \delta \text{tr}\left[ E_2 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_2^T \right],
\]

\[
B_3 \triangleq 2 \text{tr}\left[ X_0 \hat{Z} \hat{Z}^T X_0 \hat{Z} \hat{Z}^T \hat{\Theta}^{-1} \Gamma_q \right] - \text{tr}\left[ X_0 \hat{Z} \hat{Z}^T X_0 \hat{Z} \hat{Z}^T \hat{\Theta}^{-1} G \right] = \text{tr}\left[ E_1 (\Gamma_q - \delta I_q) \hat{\Theta} E_1^T (\Gamma_q - \delta I_q) \right],
\]

\[
B_4 \triangleq 2 \text{tr}\left[ X_0 (G \hat{\Theta}^{-2} + \hat{\Theta}^{-1} G \hat{\Theta}^{-1} \hat{\Theta}^{-1} \hat{Z}) \right] = 2 \text{tr}\left[ E_1^T \Gamma_q^{-1} (\Gamma_q - \delta I_q) \right],
\]

\[
B_5 \triangleq a^2 \text{tr}\left( \hat{\Theta}^{-3} G \right) - \frac{a^2}{2} \left[ \text{tr}(\hat{\Theta}^{-2}) + \delta^{-2} \right] + \frac{a^2}{\delta^3} b^T M b = \frac{a^2}{2} \left[ \text{tr}(\hat{\Theta}^{-2}) + \delta^{-2} \right]
\]

Thus,

\[
B = B_1 + B_2 + B_3 + B_4 + B_5
\]

\[
= \delta \left\{ \text{tr}\left[ E_2 \Gamma_q^{-1} E_2^T (\delta I_{m-q-1} - \Gamma_2) \right] + \text{tr}\left[ E_2 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_2^T \right] \right\} + \frac{m-q}{2} \delta^{-2} a^2
\]

\[
+ \frac{1}{2} \text{tr}\left[ a \Gamma_q^{-1} + 2 E_1 (\Gamma_q - \delta I_q) \right] \frac{a^2}{2} \text{tr}\left[ E_1 (\Gamma_q - \delta I_q) \right] + \text{tr}\left[ E_1 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_1^T \Gamma_q \right].
\]

It is easily verified that \( \text{tr}\left[ E_2 \Gamma_q^{-1} E_2^T (\delta I_{m-q-1} - \Gamma_2) \right] + \text{tr}\left[ E_2 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_2^T \right] \geq 0 \). On the other hand, let the \( c_i + d_i \) for \( i = 1, \ldots, q \) be the eigenvalues of \( a \Gamma_q^{-1} + 2 E_1 (\Gamma_q - \delta I_q) \). It then follows from Lemma 4 that

\[
\frac{1}{2} \text{tr}\left[ a \Gamma_q^{-1} + 2 E_1 (\Gamma_q - \delta I_q) \right] \frac{a^2}{2} \text{tr}\left[ E_1 (\Gamma_q - \delta I_q) \right] = \frac{1}{2} \sum_{i=1}^q (c_i^2 - d_i^2).
\]

Furthermore, Lemma 4 (ii) shows that

\[
\frac{1}{2} \sum_{i=1}^q d_i^2 \leq \text{tr}\left[ E_1 \Gamma_q^{-1} (\Gamma_q - \delta I_q) E_1^T \Gamma_q \right] - \text{tr}\left[ E_1 (\Gamma_q - \delta I_q) \frac{a^2}{2} E_1 (\Gamma_q - \delta I_q) \right].
\]
In summary, we prove that $B > 0$. This thus implies that $(\hat{A}, \hat{\delta})$ is the strict local minimizer of $G_1(A, \delta)$ under the constraint $A^Tb = 0$.

Also, replacing $S$ for $M$ in $G_1(A, \delta)$, we immediately obtain the strict local minimizer of $G(A, \delta)$. In this case, since $S^Tb = 0$, we have

$$\hat{\delta} = \frac{1}{m-q} \sum_{i=q+1}^{m} \gamma_i.$$ 

### Appendix C. The Proof of Lemma 2

We prove the lemma by induction on $t$. Let the rank of $S$ be $k \ge q$. Then we can write the condensed SVD of $S$ as $S = BdB^T$ where $B$ is an $m \times k$ matrix with orthonormal columns and $D$ is a $k \times k$ diagonal matrix with positive diagonal entries. Since $\text{range}(A_{(0)}) \subseteq \text{range}(S)$, we are able to express $A_{(0)}$ as $A_{(0)} = BC$ where $C$ is a $k \times q$ matrix of full-column rank. Subsequently, we have

$$Z_{(1)} = SA_{(0)} = BDC,$$

which implies the rank of $Z_{(1)}$ is $q$. We now assume that $A_{(t)}$ is of full-column rank. In this case, the columns of $Z_{(t+1)} = SA_{(t)}$ are mutually independent. By induction, we can derive $A_{(t+1)}$ is a matrix of full-column rank.

### Appendix D. The Proof of Theorem 3

We now prove that the $\delta$ computed by (2) is positive. Assume that we set the initial value of $\delta$ to a positive number, i.e., $\delta_{(0)} > 0$. Now supposing $\delta_{(t)} > 0$, we want to prove that $\delta_{(t+1)} > 0$. Substituting (1) into (2), we have

$$\delta_{(t+1)} = \frac{1}{m} \left[ \text{tr}(S) - \text{tr} \left( SA_{(t)} (\delta_{(t)} I_q + \Sigma_{(t)}^{-1} A^T A_{(t)}^{-1} \Sigma_{(t)}^{-1} A^T S) \right) \right].$$

Denote $B = S - SA(\delta \Sigma + A^T SA)^{-1} A^T S$. Eq. (1) shows that $SS^+ A_{(t+1)} = S^+ SA_{(t+1)} = A_{(t+1)}$ due to $S^+ SS = S$ and $SS^+ S = S$. It is then easily proven that $B_{(t)}$ is the Moore-Penrose inverse of $S^+ + \delta_t^{-1} A_{(t)} \Sigma_t^{-1} A^T_{(t)}$ (Harville, 1977). As a result, $B_{(t)}$ is p.s.d. due to positive semidefiniteness of $S$ and $A \Sigma^{-1} A^T$. Thus, $\text{tr}(B_{(t)})$ is positive.

It is well known that the standard EM algorithm converges to a local minimum or a saddle point. In any case, assume $A_{(t)} \to \hat{A}$ and $\delta_{(t)} \to \hat{\delta}$. It follows from (1) and (2) that

$$\hat{A} = SA \left( \delta I_q + \Sigma^{-1} \hat{A}^T S \hat{A} \right)^{-1}$$

$$\hat{\delta} = \frac{1}{m} \left[ \text{tr}(S) - \text{tr}(\hat{A} \Sigma^{-1} \hat{A}^T S) \right].$$

We thus have $\hat{A}(\delta I_q + \Sigma^{-1} \hat{A}^T S) = S \hat{A}$. Since $A \Sigma^{-1} = A(\delta I_q + \hat{A}^T \hat{A})^{-1} = (\delta I_m + \hat{A} \hat{A}^T)^{-1} \hat{A}$, we obtain $S \hat{A} = \hat{A}(\delta I_q + \hat{A}^T \hat{A})$. Let $V \Lambda V^T = \hat{A}^T \hat{A}$ be SVD of $\hat{A}^T \hat{A}$. Then $S \hat{A} V \Lambda^{-\frac{1}{2}} = \hat{A} \Lambda \Lambda^{-\frac{1}{2}} (\delta I_q + \Lambda)$. This implies that $\delta I_q + \Lambda$ and $\hat{A} \Lambda \Lambda^{-\frac{1}{2}}$ are the eigenvalue matrix and corresponding eigenvector matrix of $S$. According to Appendix B, we have $\hat{A} = U_q (\Gamma_q - \delta I_q)^{\frac{3}{2}} V^T$. In this case, because of $\text{tr}(\hat{A} \Sigma^{-1} \hat{A}^T S) = \text{tr}(\Gamma_q) - q \hat{\delta}$, we have

$$\hat{\delta} = \frac{1}{m-q} \sum_{j=q+1}^{m} \gamma_j.$$
Appendix E. Derivation of the EM Algorithm

In the case that \( \mathbf{u} = \frac{1}{\text{Im}} \mathbf{F}^T \mathbf{b} \), we have \( \mathbf{F} - \text{Im} \mathbf{u}^T = \mathbf{H}_b \mathbf{F} \). It is readily seen that

\[
\mathbf{H}_b \mathbf{F} | \mathbf{W} \sim N_{m,r} (\mathbf{A} \mathbf{W}, \delta (\text{Im} \otimes \mathbf{I}_r)/r).
\]

Using Bayes’ rule, we can compute the conditional distribution of \( \mathbf{W} \) given \( \mathbf{H}_b \mathbf{F} \) as

\[
\mathbf{W} | \mathbf{H}_b \mathbf{F} \sim N_{q,r} (\Sigma^{-1} \mathbf{A}^T \mathbf{H}_b \mathbf{F}, \delta (\Sigma^{-1} \otimes \mathbf{I}_r)/r),
\]

where \( \Sigma = \delta \mathbf{I}_q + \mathbf{A}^T \mathbf{A} \).

Considering \( \mathbf{W} \) as the missing data, \{\( \mathbf{W}, \mathbf{H}_b \mathbf{F} \)\} as the complete data, and \( \mathbf{A} \) and \( \delta \) as the model parameters, we now devise an EM algorithm for the ridge approximation. First, the complete-data log-likelihood is

\[
L_c = \log p(\mathbf{W}, \mathbf{H}_b \mathbf{F}) = \log p(\mathbf{H}_b \mathbf{F} | \mathbf{W}) + \log p(\mathbf{W})
\]

\[
\propto -\frac{mr}{2} \log \delta - \frac{r}{2} \text{tr} (\mathbf{WW}^T) - \frac{r}{2\delta} \text{tr} (\mathbf{H}_b \mathbf{F} - \mathbf{A} \mathbf{W}) (\mathbf{H}_b \mathbf{F} - \mathbf{A} \mathbf{W})^T,
\]

where we have omitted the terms independent of \( \mathbf{A} \) and \( \delta \). It is easy to find that \( \mathbf{W} \) and \( \mathbf{WW}^T \) are the complete-data sufficient statistics for \( \mathbf{A} \) and \( \delta \).

Using some properties of matrix-variate normal distributions (Gupta and Nagar, 2000, Page 60), we have

\[
\mathbb{E}(\mathbf{W}|\mathbf{H}_b \mathbf{F}) = \Sigma^{-1} \mathbf{A}^T \mathbf{H}_b \mathbf{F},
\]

(13)

\[
\mathbb{E}(\mathbf{WW}^T|\mathbf{H}_b \mathbf{F}) = \delta \Sigma^{-1} + \Sigma^{-1} \mathbf{A}^T \mathbf{S} \mathbf{A} \Sigma^{-1}.
\]

(14)

Given the \( t \)th estimates, \( \mathbf{A}_{(t)} \) and \( \delta_{(t)} \), of \( \mathbf{A} \) and \( \delta \), the E-step computes the expectation of \( L_c \) w.r.t. \( p(\mathbf{W}|\mathbf{H}_b \mathbf{F}, \mathbf{A}_{(t)}, \delta_{(t)}) \), namely,

\[
Q(\mathbf{A}, \delta|\mathbf{H}_b \mathbf{A}_{(t)}, \delta_{(t)}) = -\frac{mr}{2} \log \delta - \frac{r}{2} \text{tr} (\mathbf{WW}^T) - \frac{r}{2\delta} \text{tr} (\mathbf{S})
\]

\[
-\frac{r}{2\delta} \text{tr} (\mathbf{A} \mathbf{WW}^T \mathbf{A}^T) + \frac{r}{\delta} \text{tr} (\mathbf{A} \mathbf{W} \mathbf{F}^T \mathbf{H}_b^T),
\]

where \( \langle \mathbf{W} \rangle = \mathbb{E}(\mathbf{W}|\mathbf{H}_b \mathbf{F}, \mathbf{A}_{(t)}, \delta_{(t)}) \) and \( \langle \mathbf{WW}^T \rangle = \mathbb{E}(\mathbf{WW}^T|\mathbf{H}_b \mathbf{F}, \mathbf{A}_{(t)}, \delta_{(t)}) \). It follows from (13) and (14) that

\[
\langle \mathbf{W} \rangle = \Sigma_{(t)}^{-1} \mathbf{A}_{(t)}^T \mathbf{H}_b \mathbf{F},
\]

(15)

\[
\langle \mathbf{WW}^T \rangle = \delta_{(t)} \Sigma_{(t)}^{-1} + \Sigma_{(t)}^{-1} \mathbf{A}_{(t)}^T \mathbf{S} \mathbf{A}_{(t)} \Sigma_{(t)}^{-1}.
\]

(16)

The M-step maximizes \( Q(\mathbf{A}, \delta|\mathbf{A}_{(t)}, \delta_{(t)}) \) w.r.t. \( \mathbf{A} \) and \( \delta \), giving their new estimates:

\[
\mathbf{A}_{(t+1)} = \mathbf{H}_b \mathbf{F} \langle \mathbf{WW}^T \rangle (\langle \mathbf{WW}^T \rangle)^{-1},
\]

(17)

\[
\delta_{(t+1)} = \frac{1}{m} \left[ \text{tr}(\mathbf{S}) + \text{tr} \left( \mathbf{A}_{(t+1)}^T \langle \mathbf{WW}^T \rangle \mathbf{A}_{(t+1)} - 2 \mathbf{A}_{(t+1)}^T \mathbf{H}_b \mathbf{F} \langle \mathbf{WW}^T \rangle \right) \right].
\]

(18)

It then follows from (17) that

\[
\mathbf{A}_{(t+1)} \langle \mathbf{WW}^T \rangle = \mathbf{H}_b \mathbf{F} \langle \mathbf{WW}^T \rangle.
\]
Thus, we can rewrite (18) as

$$\delta_{(t+1)} = \frac{1}{m} \left[ \text{tr}(S) - \text{tr}(A^T_{(t+1)} H_\delta F\langle W^T \rangle) \right].$$

(19)

Now substituting $\langle W \rangle$ and $\langle WW^T \rangle$ from (15) and (16) into (17) and (19), we can combine the E-step and M-step into (1) and (2).

References

D. Achlioptas, F. McSherry, and B. Schölkopf. Sampling techniques for kernel methods. *Advances in Neural Information Processing Systems 13*, 2001.

J. H. Ahn and J. H. Oh. A constrained EM algorithm for principal component analysis. *Neural Computation*, 15:57–65, 2003.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, second edition, 1984.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.

S. Fine, K. Scheinberg, N. Cristianini, J. Shawe-Taylor, and B. Williamson. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

G. H. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third edition, 1996.

J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarities coefficients. *Journal of Classification*, 3:5–48, 1986.

P. J. F. Groenen, R. Mathar, and W. J. Heiser. The majorization approach to multidimensional scaling for Minkowski distance. *Journal of Classification*, 12:3–19, 1995.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, 2000.

D. A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80C86, 1970.

I. T. Jolliffe. *Principal component analysis*. Springer, New York, second edition edition, 2002.

N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems 16*, 2004.
M. Lázaro-Gredilla, J. Quiñonero Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

Q. Le, T. Sarlós, and A. J. Smola. Fastfood—approximating kernel expansions in loglinear time. In *The 30th International Conference on Machine Learning*, 2013.

J. R. Magnus and H. Neudecker. *Matrix Calculus with Applications in Statistics and Econometric*. John Wiley & Sons, New York, revised edition edition, 1999.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, volume 14, 2001.

M.-H. Oh and A. E. Raftery. Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, 96(455):1031–1044, 2001.

J. Quiñonero-Candela, C. E. Rasmussen, and C. K. I. Williams. Approximation methods for Gaussian process regression. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Machine*, pages 203–223. MIT Press, 2007.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2008.

J. O. Ramsay. Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society Series A*, 145:285–312, 1982.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006.

R. Rosipal and M. Girolami. An expectation-maximization approach to nonlinear component analysis. *Neural Computation*, 13:505–510, 2001.

S. Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems 10*, 1998.

B. Schölkopf and A. Smola. *Learning with Kernels*. The MIT Press, 2002.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. V. N. Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10:2615–2637, 2009.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *The 17th International Conference on Machine Learning*, 2000.
M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3):611–622, 1999.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, 2001.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems 25*, 2012.

Z. Zhang and M. I. Jordan. Multiway spectral clustering: A margin-based perspective. *Statistical Science*, 23(2):383–403, 2008.

Z. Zhang, J. T. Kwok, and D.-Y. Yeung. Model-based transductive learning of the kernel matrix. *Machine Learning*, 63(1):69–101, 2006.