ESTIMATING PRINCIPAL COMPONENTS OF LARGE COVARIANCE MATRICES USING THE NYSTRÖM METHOD

Nicholas Arcolano and Patrick J. Wolfe

Statistics and Information Sciences Laboratory, Harvard University
33 Oxford Street, Cambridge, MA 02138, USA

ABSTRACT

Covariance matrix estimates are an essential part of many signal processing algorithms, and are often used to determine a low-dimensional principal subspace via their spectral decomposition. However, for sufficiently high-dimensional matrices exact eigenanalysis is computationally intractable, and in the case of limited data, sample eigenvalues and eigenvectors are known to be poor estimators of their true counterparts. To address these issues, we propose a covariance estimator that is computationally efficient while also performing shrinkage on the sample eigenvalues. Our approach is based on the Nyström method, which uses a data-dependent orthogonal projection to obtain a fast low-rank approximation of a large positive semidefinite matrix. We provide a theoretical analysis of the error properties of our estimator as well as empirical results.

***Index Terms*—** approximate spectral analysis, high-dimensional data, low-rank approximation, Nyström extension, shrinkage estimators

1. INTRODUCTION

The need to estimate the covariance matrix of a random vector given observed data is one that arises in many areas within signal processing; examples include beamforming [1], speech processing [2], and source separation [3]. For many of these applications, the main purpose of covariance matrix estimation is to determine the principal subspace containing some signal of interest. Accomplishing this goal requires spectral analysis of the estimated covariance.

Two challenges commonly arise in these types of problems. The first is that as computational power and data storage capacity continue to grow, practitioners often wish to consider signals (and thus covariances) of increasing size and dimensionality. Since the fundamental computational complexity of the eigenvalue problem scales as $O(n^3)$, determining a principal subspace quickly becomes prohibitively expensive. An alternative is to perform the spectral analysis using some computationally efficient approximate method.

The second challenge is that the eigenvalues and eigenvectors the sample covariance matrix are known to be be poor estimates of the true eigenvalues and eigenvectors, especially when operating in high dimensions with limited observations [4, 5]. In particular, the sample eigenvalues have been shown to be over-dispersed, and much effort has been focused on developing shrinkage estimators that provide better estimates of the true spectrum [6, 7].

Despite prevalence of these two issues and the wealth of theoretical results concerning each of them, there exist few approaches that address them concurrently. Consequently, we propose a covariance estimator that is not only computationally efficient, but also shrinks the eigenvalues of the sample covariance. Our approach is based on the Nyström method, a technique for obtaining low-rank approximations of high-dimensional positive semidefinite matrices that has proven effective in a variety of areas, including machine learning [8], image segmentation [9], and computer vision [10].

The rest of the discussion proceeds as follows. First, we review the Nyström method and develop its use as an estimator of a low-rank covariance matrix (and its corresponding low-dimensional subspace). We then analyze its error characteristics and shrinkage properties, and conclude with some empirical results.

2. THE NYSTRÖM METHOD

The Nyström method is a classical technique for obtaining numerical solutions to eigenfunction problems. When applied to matrices, it can be used to construct a low-rank approximation of a positive semidefinite matrix as follows. Let $\mathbf{S}$ be a $p \times p$ positive semidefinite matrix, represented in block form as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22} \end{bmatrix},$$

where $\mathbf{S}_{11}$ is $k \times k$. The Nyström approximation of $\mathbf{S}$ preserves $\mathbf{S}_{11}$ and $\mathbf{S}_{12}$ while approximating $\mathbf{S}_{22}$ by its Nyström extension:

$$\widehat{\mathbf{S}} \equiv \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{12}^T & \mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} \mathbf{S}_{12} \end{bmatrix}. \tag{1}$$

It is straightforward to show that the rank of $\widehat{\mathbf{S}}$ is equal to the rank of $\mathbf{S}_{11}$, which is at most k. (Note that when $\mathbf{S}_{11}$ is singular, $\mathbf{S}_{11}^{-1}$ in the above expression can be replaced by the Moore-Penrose pseudoinverse $\mathbf{S}_{11}^{+}$.) Since the approximation reconstructs $\mathbf{S}_{11}$ and $\mathbf{S}_{12}$ perfectly, the approximation error $\mathbf{S} - \widehat{\mathbf{S}}$ is characterized entirely by the Schur complement of $\mathbf{S}_{11}$ in $\mathbf{S}$, defined as

$$\overline{\mathbf{S}}_{11} \equiv \mathbf{S}_{22} - \mathbf{S}_{12}^T \mathbf{S}_{11}^{-1} \mathbf{S}_{12}.$$

If we view $\mathbf{S}$ as a Gram matrix (inner product) or covariance (outer product) of some underlying data, an alternative way to characterize the Nyström approximation—one which will prove very enlightening for our purposes—is as a function of an orthogonal projection. Let $\mathbf{X}$ be a $p \times n$ matrix, which we partition as

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Z} \end{bmatrix} \tag{2}$$

where $\mathbf{Y}$ is $k \times n$, and let

$$\mathbf{S} = \mathbf{X}\mathbf{X}^T = \begin{bmatrix} \mathbf{Y}\mathbf{Y}^T & \mathbf{Y}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{Y}^T & \mathbf{Z}\mathbf{Z}^T \end{bmatrix}.$$

This work is sponsored by the United States Air Force under contract FA8721-05-C-0002. Opinions, interpretations, recommendations, and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

Define the $n \times n$ symmetric idempotent matrix

$$\mathbf{P} \equiv \mathbf{Y}^T \left(\mathbf{Y}\mathbf{Y}^T\right)^{-1} \mathbf{Y},$$

which represents an orthogonal projection onto the k-dimensional subspace of $\mathbb{R}^n$ spanned by the rows of $\mathbf{Y}$. In this context, we obtain the same expression as in (1) by approximating $\mathbf{X}$ with its projection $\mathbf{XP}$, yielding

$$\begin{aligned}\widehat{\mathbf{S}} = \mathbf{XP}(\mathbf{XP})^T &= \mathbf{XPX}^T \\ &= \begin{bmatrix} \mathbf{YY}^T & \mathbf{YZ}^T \\ \mathbf{ZY}^T & \mathbf{ZY}^T(\mathbf{YY}^T)^{-1}\mathbf{YZ}^T \end{bmatrix}.\end{aligned}$$

This interpretation highlights the fact that the Nyström method is not inherently specialized to the submatrix $\mathbf{Y}$; we may choose instead to base the approximation on any k of the p rows of $\mathbf{X}$. Since different subsets typically yield different approximations, we can view $\widehat{\mathbf{S}}$ as a function of a set of k indices $I \subseteq \{1, \ldots, p\}$.[1]

Accordingly, we can view $\widehat{\mathbf{S}}$ in some sense as transforming the problem of low-rank approximation into one of subset selection. The most common solution (used, for example, in [8, 9]) is to sample I randomly with uniform probability from the set of all k-subsets of $\{1, \ldots, p\}$. However, this approach ignores any particular structure inherent in $\mathbf{S}$, and as an alternative, many recent results have focused on robust data-dependent sampling methods. For example, in [10] the authors show that by sampling I according to $P(I) \propto \det(\mathbf{S}_I)$, the expected approximation error can be bounded within within a factor of $k+1$ times the error of the optimal approximation.

Once we have selected a suitable subset, the primary advantage of the Nyström method is its computational efficiency. Given a set of k indices, the approximation $\widehat{\mathbf{S}}$ can be reconstructed for a cost of $O(p^2k)$, as opposed to $O(p^3)$ for the optimal rank-k approximation obtained by performing the full spectral decomposition of $\mathbf{S}$ and retaining its k largest principal components. Moreover, using approaches such as those described in [9], one can obtain the eigenvalues and eigenvectors of $\widehat{\mathbf{S}}$ at computational cost of only $O(pk^2)$.

3. NYSTRÖM COVARIANCE ESTIMATOR

As discussed earlier, one of the most common motivations for estimating an unknown covariance is determining its principal components. Unfortunately, the combined costs of the estimation step (particularly when using sophisticated shrinkage-style estimators) and spectral analysis can be overwhelming for very large data sets. One solution is to reduce computation by using a fast low-rank covariance matrix estimator whose principal components are inexpensive to compute. To this end we propose the Nyström covariance estimator, defined as follows.

Definition 1 (Nyström covariance estimator)**.** *Let $\mathbf{X}$ be a $p \times n$ matrix whose columns $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are i.i.d. random samples such that $\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$ and $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T) = \mathbf{\Sigma}$ for $i = 1, \ldots, n$. Given a set of k indices $I \subseteq \{1, \ldots, p\}$, we define the Nyström covariance estimator of $\mathbf{\Sigma}$ as*

$$\widehat{\mathbf{S}}(I) \equiv \tfrac{1}{n}\mathbf{XPX}^T,$$

where $\mathbf{P}$ denotes the $n \times n$ orthogonal projection onto the subspace

$$R(\mathbf{P}) = \text{span}\{\mathbf{x}_i : i \in I\}.$$

[1] Throughout the paper, we we will use $\mathbf{A}_{IJ}$ to denote the submatrix of a matrix $\mathbf{A}$ whose rows and columns are specified by respective index sets I and J, and define $\mathbf{A}_I \equiv \mathbf{A}_{II}$.

Note that we have restricted ourselves to the zero-mean case to clarify subsequent exposition and analysis. Also, note that in light of the derivation of the Nyström method given in Section 2, $\widehat{\mathbf{S}}(I)$ could equally be viewed as a low-rank approximation of the sample covariance $\mathbf{S} = \frac{1}{n}\mathbf{XX}^T$.

Assume now that the data are drawn independently from a p-variate normal distribution with zero mean and covariance $\mathbf{\Sigma}$, denoted $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. In this case, we can compute the bias and expected square error of the Nyström covariance estimator directly, as illustrated in the following theorems. Note that while we arguably could define these quantities with respect to optimal low-rank estimate of $\mathbf{\Sigma}$ (denoted $\mathbf{\Sigma}_k$) instead, this would make theoretical analysis all but impossible. Furthermore, we can always make an equitable comparison between different estimators with respect to $\mathbf{\Sigma}_k$, since the difference $\mathbf{\Sigma} - \mathbf{\Sigma}_k$ is fixed.

Theorem 1 (Bias of Nyström covariance estimator)**.** *Let $\mathbf{X}$ be a $p \times n$ matrix whose columns $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are i.i.d. random samples from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Let $\widehat{\mathbf{S}}(I)$ be the Nyström covariance estimator of $\mathbf{\Sigma}$ given a set of k indices $I \subseteq \{1, \ldots, p\}$, and define $J = \{1, \ldots, p\} \setminus I$. Then the bias matrix*

$$\mathbf{B} \equiv \mathbf{\Sigma} - \mathbb{E}\big(\widehat{\mathbf{S}}(I)\big)$$

is zero except for elements within the submatrix $\mathbf{B}_J$, whose values are given by

$$\mathbf{B}_J = \frac{n-k}{n}\overline{\mathbf{\Sigma}}_I = \frac{n-k}{n}\left[\mathbf{\Sigma}_J - \mathbf{\Sigma}_{IJ}^T\mathbf{\Sigma}_J^{-1}\mathbf{\Sigma}_{IJ}\right].$$

Proof. Without loss of generality we may let $I = \{1, \ldots, k\}$ and $J = \{k+1, \ldots, p\}$. The true covariance can be partitioned as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_I & \mathbf{\Sigma}_{IJ} \\ \mathbf{\Sigma}_{IJ}^T & \mathbf{\Sigma}_J \end{bmatrix}.$$

Partitioning $\mathbf{X}$ as in (2), the Nyström covariance estimate is given by

$$\widehat{\mathbf{S}} = \frac{1}{n}\mathbf{XPX}^T = \frac{1}{n}\begin{bmatrix} \mathbf{YY}^T & \mathbf{YZ}^T \\ \mathbf{ZY}^T & \mathbf{ZPZ}^T \end{bmatrix},$$

where $\mathbf{P} = \mathbf{Y}^T\left(\mathbf{YY}^T\right)^{-1}\mathbf{Y}$. It is then straightforward to prove that the first three submatrices have zero bias by showing that $\mathbb{E}(\frac{1}{n}\mathbf{YY}^T) = \mathbf{\Sigma}_I$, $\mathbb{E}(\frac{1}{n}\mathbf{YZ}^T) = \mathbf{\Sigma}_{IJ}$, and $\mathbb{E}(\frac{1}{n}\mathbf{ZY}^T) = \mathbf{\Sigma}_{IJ}^T$.

Next, we must compute $\mathbb{E}(\frac{1}{n}\mathbf{ZPZ}^T)$. This calculation can be performed directly using the conditional multivariate normal distribution along with the iterated expectation

$$\mathbb{E}(\mathbf{ZPZ}^T) = \mathbb{E}_Y\Big[\mathbb{E}(\mathbf{ZPZ}^T \mid \mathbf{Y})\Big].$$

However, a simpler approach is to exploit the properties of Schur complements of Wishart distributions. Recall that the sample covariance

$$\mathbf{S} = \tfrac{1}{n}\mathbf{XX}^T \sim \mathcal{W}_p\left(n, \tfrac{1}{n}\mathbf{\Sigma}\right).$$

From [11], we have that the Schur complement of $\mathbf{S}_I$ in $\mathbf{S}$ is also Wishart, with a distribution given by

$$\begin{aligned}\overline{\mathbf{S}}_I &= \tfrac{1}{n}\mathbf{ZZ}^T - \tfrac{1}{n}\mathbf{ZY}^T\left(\mathbf{YY}^T\right)^{-1}\mathbf{YZ}^T \\ &\sim \mathcal{W}_{p-k}\big(n-k, \tfrac{1}{n}\mathbf{\Sigma}_J - \tfrac{1}{n}\mathbf{\Sigma}_{IJ}^T\mathbf{\Sigma}_I^{-1}\mathbf{\Sigma}_{IJ}\big),\end{aligned}$$

and thus taking the expected value of $\overline{\mathbf{S}}_I$ yields the result. □

Theorem 2 (MSE of Nyström covariance estimator)**.** *Let $\mathbf{X}$ be a $p \times n$ matrix whose columns $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are i.i.d. random samples from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Let $\widehat{\mathbf{S}}(I)$ be the Nyström covariance estimator of $\boldsymbol{\Sigma}$ given a set of k indices $I \subseteq \{1, \ldots, p\}$, and define $J = \{1, \ldots, p\} \setminus I$. Then mean square error in Frobenius norm is*

$$\mathbb{E}\left\|\boldsymbol{\Sigma} - \widehat{\mathbf{S}}(I)\right\|_F^2 = \delta_{\text{ML}}(\boldsymbol{\Sigma}) + \rho\left[\left\|\overline{\boldsymbol{\Sigma}}_I\right\|_F^2 - \delta_{\text{ML}}\left(\overline{\boldsymbol{\Sigma}}_I\right)\right],$$

where $\rho = (n-k)^2/n^2$, $\delta_{\text{ML}}(\boldsymbol{\Sigma})$ is the mean square error of the sample covariance, given by

$$\delta_{\text{ML}}\left(\boldsymbol{\Sigma}\right) \equiv \frac{1}{n}\left[\text{tr}\left(\boldsymbol{\Sigma}^2\right) + \text{tr}^2\left(\boldsymbol{\Sigma}\right)\right],$$

and $\delta_{\text{ML}}\left(\overline{\boldsymbol{\Sigma}}_I\right)$ is the mean square error of the Schur complement, given by

$$\delta_{\text{ML}}\left(\overline{\boldsymbol{\Sigma}}_I\right) \equiv \frac{1}{n-k}\left[\text{tr}\left(\overline{\boldsymbol{\Sigma}}_I^2\right) + \text{tr}^2\left(\overline{\boldsymbol{\Sigma}}_I\right)\right].$$

Proof. As in the proof of Theorem 1, we can let $I = \{1, \ldots, k\}$ and $J = \{k+1, \ldots, p\}$ without loss of generality. The error is given by

$$\begin{aligned}
\mathbb{E}\left\|\boldsymbol{\Sigma} - \widehat{\mathbf{S}}(I)\right\|_F^2 &= \mathbb{E}\left\|\boldsymbol{\Sigma} - \tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right\|_F^2 \\
&= \mathbb{E}\,\text{tr}\left[\left(\boldsymbol{\Sigma} - \tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right)^T\left(\boldsymbol{\Sigma} - \tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right)\right] \\
&= \text{tr}\left(\boldsymbol{\Sigma}^2\right) - 2\,\text{tr}\left[\boldsymbol{\Sigma}\,\mathbb{E}\left(\tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right)\right] \\
&\quad + \text{tr}\left[\mathbb{E}\left(\tfrac{1}{n^2}\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{X}^T\right)\right]. \qquad (3)
\end{aligned}$$

From Theorem 1, we have that

$$\mathbb{E}\left(\tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right) = \begin{bmatrix} \boldsymbol{\Sigma}_I & \boldsymbol{\Sigma}_{IJ} \\ \boldsymbol{\Sigma}_{IJ}^T & \frac{k}{n}\boldsymbol{\Sigma}_J + \frac{(n-k)}{n}\boldsymbol{\Sigma}_{IJ}^T\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{IJ} \end{bmatrix},$$

and thus

$$\begin{aligned}
\text{tr}\left[\boldsymbol{\Sigma}\,\mathbb{E}\left(\tfrac{1}{n}\mathbf{X}\mathbf{P}\mathbf{X}^T\right)\right] &= \text{tr}\left(\boldsymbol{\Sigma}_I^2\right) + 2\,\text{tr}\left(\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{IJ}^T\right) + \frac{k}{n}\text{tr}\left(\boldsymbol{\Sigma}_J^2\right) \\
&\quad + \frac{(n-k)}{n}\text{tr}\left(\boldsymbol{\Sigma}_J\boldsymbol{\Sigma}_{IJ}^T\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{IJ}\right). \qquad (4)
\end{aligned}$$

Next, we must compute $\mathbb{E}\left(\frac{1}{n^2}\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{X}^T\right)$. Partitioning $\mathbf{X}$ as in (2) and expanding in block form yields

$$\begin{aligned}
\text{tr}\left(\mathbf{X}\mathbf{P}\mathbf{X}^T\mathbf{X}\mathbf{P}\mathbf{X}^T\right) &= \text{tr}\left(\mathbf{Y}\mathbf{Y}^T\mathbf{Y}\mathbf{Y}^T\right) + 2\,\text{tr}\left(\mathbf{Y}\mathbf{Z}^T\mathbf{Z}\mathbf{Y}^T\right) \\
&\quad + \text{tr}\left(\mathbf{Z}\mathbf{P}\mathbf{Z}^T\mathbf{Z}\mathbf{P}\mathbf{Z}^T\right).
\end{aligned}$$

The expected value of the first term can be computed directly using standard properties of normal random variables; for the second and third terms, we must also use iterated expectation to condition on $\mathbf{Y}$ and $\mathbf{P}$. Performing these calculations yields

$$\mathbb{E}\,\text{tr}\left(\mathbf{Y}\mathbf{Y}^T\mathbf{Y}\mathbf{Y}^T\right) = \left(n^2+n\right)\text{tr}\left(\boldsymbol{\Sigma}_I^2\right) + n\,\text{tr}^2\left(\boldsymbol{\Sigma}_I\right), \qquad (5)$$

$$\begin{aligned}
\mathbb{E}\,\text{tr}\left(\mathbf{Y}\mathbf{Z}^T\mathbf{Z}\mathbf{Y}^T\right) &= \left(n^2+n\right)\text{tr}\left(\boldsymbol{\Sigma}_{IJ}\boldsymbol{\Sigma}_{IJ}^T\right) \qquad (6) \\
&\quad + n\,\text{tr}\left(\boldsymbol{\Sigma}_I\right)\text{tr}\left(\boldsymbol{\Sigma}_J\right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\,\text{tr}\left(\mathbf{Z}\mathbf{P}\mathbf{Z}^T\mathbf{Z}\mathbf{P}\mathbf{Z}^T\right) &= \left(k^2+k\right)\text{tr}\left(\overline{\boldsymbol{\Sigma}}_I^2\right) + k\,\text{tr}^2\left(\overline{\boldsymbol{\Sigma}}_I\right) \qquad (7) \\
&\quad + 2n\left(k+1\right)\text{tr}\left(\overline{\boldsymbol{\Sigma}}_I\mathbf{R}\right) \\
&\quad + 2n\left(k+1\right)\text{tr}\left(\overline{\boldsymbol{\Sigma}}_I\right)\text{tr}\left(\mathbf{R}\right) \\
&\quad + \left(n^2+n\right)\text{tr}\left(\mathbf{R}^2\right) + k\,\text{tr}^2\left(\mathbf{R}\right),
\end{aligned}$$

where $\mathbf{R} \equiv \boldsymbol{\Sigma}_{IJ}^T\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{IJ}$. Substituting (4)–(7) into (3) and rearranging and simplifying terms yields the desired result. □

Finally, we show that the Nyström covariance estimator is guaranteed to shrink the sample eigenvalues.

Theorem 3 (Shrinkage property of Nyström covariance estimator)**.** *Let $\mathbf{S} \succeq 0$ be a $p \times p$ sample covariance matrix, and let $\widehat{\mathbf{S}}(I)$ be the corresponding Nyström covariance estimator given a set of k indices $I \subseteq \{1, \ldots, p\}$. Let $\{\lambda_i(\mathbf{A})\}$ denote the eigenvalues a $p \times p$ real symmetric matrix $\mathbf{A}$, ordered such that*

$$\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_p(\mathbf{A}).$$

Then

$$\lambda_i\left(\mathbf{S}\right) \geq \lambda_i\left(\widehat{\mathbf{S}}(I)\right)$$

for all $i = 1, \ldots, p$.

Proof. Again let $I = \{1, \ldots, k\}$ and $J = \{k+1, \ldots, p\}$. Recall that the Nyström covariance estimator can be interpreted as a low rank approximation of the sample covariance $\mathbf{S}$, and that the error in this approximation is given by

$$\mathbf{E} \equiv \mathbf{S} - \widehat{\mathbf{S}}(I) = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \overline{\mathbf{S}}_I \end{bmatrix}.$$

Since the Schur complement is positive semidefinite, we have

$$\mathbf{S} = \widehat{\mathbf{S}}(I) + \mathbf{E},$$

where $\mathbf{E} \succeq 0$. The result then follows from Weyl's Monotonicity Theorem [12], which states that for any $p \times p$ symmetric matrix $\mathbf{A}$, $\mathbf{E} \succeq 0$ implies $\lambda_i(\mathbf{A} + \mathbf{E}) \geq \lambda_i(\mathbf{A})$ for $i = 1, \ldots, p$. □

4. EXPERIMENTAL RESULTS

We continue with an empirical study the performance of the Nyström covariance estimator. To investigate its properties in the context of low-rank approximation, we consider the "spiked covariance model" of [5], wherein the true covariance $\boldsymbol{\Sigma}$ has k large eigenvalues α and $p-k$ small eigenvalues β.

Since the error performance of the Nyström approximation depends on chosen subset I, we examine its performance with respect to two sampling methods. The first is uniform sampling, where all index sets are equally likely. The second is the data-dependent determinant sampling distribution of [10], where index sets are drawn according to the distribution $P(I) \propto \det(\mathbf{S}_I)$. Note that in practice, sampling from this distribution directly requires evaluating the determinants of a combinatoric number principal submatrices. For large problems, however, [10] offers a Metropolis algorithm for approximate sampling that performs quite well in practice.

We also examine two "oversampled" versions of these approaches. For these cases, we compute the Nyström covariance estimate for some $k' > k$. As the spectral decomposition of the Nyström approximation can be obtained for $O(pk^2)$, it is still relatively inexpensive to compute for a larger $k' \ll n$. Thus, one can choose to keep only the k largest principal components of the resulting Nyström estimate and discard the rest. While this approach results in a matrix that is no longer technically a Nyström approximation, it does retain the shrinkage properties of the original estimator. Such adaptations of the Nyström method have been shown to work well in practice [9].

Finally, we include two other methods as points of reference. The first is the optimal low-rank approximation of the sample covariance, denoted $\mathbf{S}_k$. This matrix is obtained by performing the full spectral decomposition of the sample covariance at a cost of $O(p^3)$,

Table 1. Empirical MSE and percent deviation with respect to $\mathbf{\Sigma}_k$.

	U	D	U-OS	D-OS	LW	LRS
MSE	1851.70	1849.05	1538.36	1533.73	1332.76	1232.50
Δ_{MSE}	87.0%	86.8%	55.4%	54.9%	34.6%	24.4%

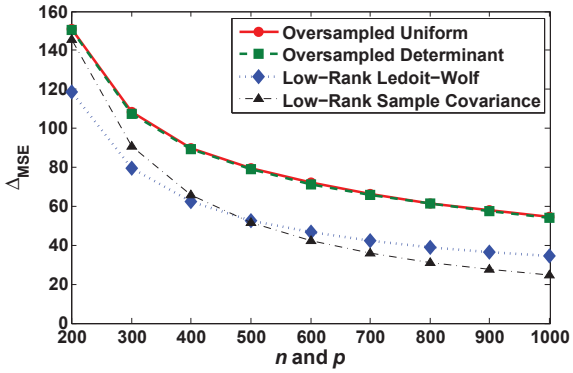

Fig. 1. Δ_{MSE} versus n, p for various low-rank covariance estimators.

and then retaining the k principal components. We also include the optimal low-rank approximation of the Ledoit-Wolf covariance estimator [7], which attempts to shrink the sample eigenvalues by computing an optimally-weighted combination of the sample covariance and the identity matrix.

For the first experiment, we let $p = 1000$. This scenario is large enough to illustrate high-dimensional behavior, while still remaining manageable enough to allow for exact eigenanalysis. We considered a spiked covariance model with $k = 10$ large eigenvalues of size $\alpha = 10$, and $p - k = 990$ small eigenvalues of size $\beta = 1$. We then generated 1000 covariance matrices by subjecting a diagonal matrix with these eigenvalues to uniformly random rotations. Using each matrix, we generated $n = 1000$ normal random vectors and computed the various low-rank covariance estimates. (In the case of oversampled estimators, we used $k' = 5k$.)

Table 1 lists the results for the uniform sampling (U), determinant sampling (D), oversampled uniform (U-OS), oversampled determinant (D-OS) Nyström estimators, as well as the low-rank versions of the Ledoit-Wolf estimator (LW) and the optimal low-rank sample covariance (LRS). We list the empirical MSE for each estimator $\widehat{\mathbf{\Sigma}}$, as well as

$$\Delta_{\text{MSE}} \equiv \frac{\mathbb{E}\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\|_F^2 - \left\|\mathbf{\Sigma} - \mathbf{\Sigma}_k\right\|_F^2}{\left\|\mathbf{\Sigma} - \mathbf{\Sigma}_k\right\|_F^2} \times 100,$$

which is the percent increase in the MSE over that of $\mathbf{\Sigma}_k$, the optimal low-rank estimate of the true sample covariance.

Next, we repeated the previous experiment for n and p varying from 200 to 1000. Because of their improved error performance, we consider only the oversampled versions of the uniform and determinant sampling Nyström estimators. Figure 1 shows a plot of Δ_{MSE} versus n and p for the Ledoit-Wolf, sample covariance, and Nyström estimators.

For lower values of n and p, all four of the methods are fairly poor estimators of the true principal eigenvalues. Consistent with the results of [7], the shrinkage provided by the Ledoit-Wolf estimator allows it to outperform the sample covariance for n and p less than around 500. However, this advantage does not persist as n and p grow large. While the Nyström estimators have the largest error overall, they remain competitive with both the sample covariance and Ledoit-Wolf estimators, especially for high dimension. This performance is especially impressive when recalling that the computational complexity of the Nyström estimators scales as $O(p^2)$ (or $O(p)$, if we only want to compute principal components without actually constructing the approximation), versus the $O(p^3)$ cost required for the other two estimators. The Nyström estimators are able to operate with such low computational demands because their inherent shrinkage properties help to balance the error induced by approximate spectral analysis.

5. SUMMARY

In conclusion, we have proposed a covariance estimator based on the Nyström method that is computationally efficient while also performing shrinkage on the sample eigenvalues. In addition to providing analytical results for its error performance, we have illustrated via simulation that it is a viable option for low-rank covariance estimation, requiring only a fraction of the computational cost of other estimators.

6. REFERENCES

[1] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proceedings F: Radar and Signal Processing*, vol. 140, no. 6, pp. 362–370, 1993.

[2] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[3] A. Belouchrani, A.-M. Karim, J. F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[4] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR: Sbornik*, vol. 1, pp. 457–483, 1967.

[5] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[6] C. Stein, "Estimation of a covariance matrix," Reitz Lecture, 39th Annual Meeting IMS, Atlanta, GA, 1975.

[7] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[8] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Neural Information Processing Systems*, 2000, vol. 13, pp. 682–688.

[9] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 1–12, 2004.

[10] M.-A. Belabbas and P. J. Wolfe, "On landmark selection and sampling in high-dimensional data analysis," *Philosophical Transactions of the Royal Society, Series A*, vol. 367, no. 1906, pp. 4295–4312, 2009.

[11] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*, Chapman & Hall / CRC, Boca Raton, 2000.

[12] R. Bhatia, *Matrix analysis*, Springer, New York, 1997.