NYSTRÖM APPROXIMATION OF WISHART MATRICES

Nicholas Arcolano and Patrick J. Wolfe

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

ABSTRACT

Spectral methods requiring the computation of eigenvalues and eigenvectors of a positive definite matrix are an essential part of signal processing. However, for sufficiently high-dimensional data sets, the eigenvalue problem cannot be solved without approximate methods. We examine a technique for approximate spectral analysis and low-rank matrix reconstruction known as the Nyström method, which recasts the eigendecomposition of large matrices as a subset selection problem. In particular, we focus on the performance of the Nyström method when used to approximate random matrices from the Wishart ensemble. We provide statistical results for the approximation error, as well as an experimental analysis of various subset sampling techniques.

***Index Terms*—** high-dimensional data analysis, kernel methods, Nyström extension, Wishart distribution

1. INTRODUCTION

In many areas of signal processing, the growth of computational power and data storage capacity has allowed researchers and practitioners to consider data sets of ever-increasing size and dimensionality. This trend has led to renewed interest in dimensionality-reduction techniques, which enable analysis of massive data sets by extracting low-dimensional characterizations of high-dimensional data. Most approaches to dimensionality reduction fall under the heading of **spectral methods**, in that they depend upon eigenanalysis of the positive definite kernel matrices that encode relationships between data points. Examples of spectral methods range from traditional approaches such as principal components analysis (PCA), Fisher discriminant analysis, and multidimensional scaling, to more recent manifold learning techniques such as isomap [1], spectral clustering [2], Laplacian [3] and Hessian [4] eigenmaps, and diffusion maps [5].

One advantage of spectral methods is that the eigendecomposition of positive definite matrices has been well studied, resulting in a wealth of theoretical results and efficient algorithms. Nevertheless, the fundamental computational complexity of the eigenvalue problem scales as $O(n^3)$, and thus as the size and dimensionality of data sets increase, the use of spectral methods becomes prohibitively expensive. A common alternative in these cases is to select a set of "landmark" coordinates or samples to represent the data as a whole; one can then perform spectral analysis directly on the landmark examples and extrapolate the results at a reduced computational cost. In effect, this process is equivalent to generating an approximation to the desired kernel matrix, based on a subset of coordinates or samples of the original data.

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.

One approximation technique of this type is the **Nyström method**, which generates a low-rank approximation to a high-dimensional positive definite matrix, given a subset of rows and columns (i.e., a principal submatrix). It has been shown to be effective in a variety of applications, including machine learning [6], image segmentation [7], and computer vision [8], and is particularly appropriate for use in spectral analysis because the approximation is characterized by a set of approximate eigenvectors derived from the eigenstructure of the specified submatrix. In this manner, the Nyström method transforms the matrix approximation problem into one of finding a principal submatrix whose eigenvalues and eigenvectors best predict the behavior of the data over the omitted coordinates—in essence, an optimal subset selection problem.

Unfortunately, despite the fundamental importance of proper coordinate selection to the Nyström method, it has yet to be studied thoroughly in the existing literature. Most of the current results focus on the case of uniform sampling [6, 7], though more recent work considers sampling as a function of the diagonal elements [9, 10] or determinant [8] of the submatrices. Moreover, there exists no analysis of the Nyström method in the context of random matrix ensembles, which are a natural setting for statistical approaches to the subset selection problem.

For these reasons, we focus in this article on the performance of the Nyström method (and its dependence on various sampling methods) when the matrix to be approximated is randomly drawn from a probability distribution. Since no previous results in this vein exist, we choose as an obvious starting point the Wishart distribution, which acts a suitable model for a wide variety of positive definite matrices, most notably when the matrix represents the sample covariance of normally-distributed data.

2. THE NYSTRÖM METHOD

The Nyström method is a classical technique for obtaining numerical solutions to eigenfunction problems. In the context of matrix approximation, it entails solving a large eigenvalue problem on a subset of coordinates, and then using this solution to generate approximate eigenvectors to reconstruct a low-rank approximation to the original matrix.

Let $\mathbf{S}$ be a $p \times p$ strictly positive definite matrix, represented in block form as

$$\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix},$$

where $\mathbf{A}$ is a $k \times k$ matrix. (We refer to a $p \times p$ symmetric matrix $\mathbf{S}$ as positive definite, denoted $\mathbf{S} \succeq 0$, if for all $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}^T\mathbf{S}\mathbf{x} \geq 0$, and as strictly positive definite, denoted $\mathbf{S} \succ 0$, if for all $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^T\mathbf{S}\mathbf{x} > 0$.) Assume we want a rank-k approximation to $\mathbf{S}$. Let $\mathbf{\Lambda}$ and $\mathbf{U}$ be $k \times k$ matrices containing the eigenvalues and corresponding eigenvectors of $\mathbf{A}$. We generate the approximation by

mapping the k eigenvectors of $\mathbf{A}$ to the remaining $p-k$ dimensions using the **Nyström extension**:

$$\widehat{\mathbf{U}} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B}^T \end{bmatrix} \mathbf{U}\mathbf{\Lambda}^{-1} = \begin{bmatrix} \mathbf{U} \\ \mathbf{B}^T\mathbf{U}\mathbf{\Lambda}^{-1} \end{bmatrix},$$

where $\widehat{\mathbf{U}}$ is a $p \times k$ matrix of approximate eigenvectors of $\mathbf{S}$. Note that these k vectors are not orthogonal, though we can still use them to reconstruct a rank-k approximation, given by

$$\widehat{\mathbf{S}}_k \equiv \widehat{\mathbf{U}}\mathbf{\Lambda}\widehat{\mathbf{U}}^T = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{bmatrix}.$$

Of course, this approach is not inherently specialized to the submatrix $\mathbf{A}$; we may choose instead to base the approximation on any $k \times k$ principal submatrix of $\mathbf{S}$, as indicated in the following definition. (Throughout the paper, we use $\mathbf{S}_{IJ}$ to denote the submatrix of $\mathbf{S}$ whose rows and columns are specified by respective index sets I and J, and define $\mathbf{S}_I \equiv \mathbf{S}_{II}$.)

Definition 1 (Nyström method)**.** *Let $\mathbf{S}$ be a $p \times p$ strictly positive definite matrix, and specify a set of k indices $I \subseteq \{1, \ldots, p\}$. Define the permuted matrix $\mathbf{S}^*$, given in block form as*

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{S}_I & \mathbf{S}_{IJ} \\ \mathbf{S}_{IJ}^T & \mathbf{S}_J \end{bmatrix},$$

where $J = \{1, \ldots, p\} \setminus I$. We define the rank-$k$ approximation of $\mathbf{S}^*$ given I as the $p \times p$ matrix

$$\widehat{\mathbf{S}}_k^*(I) \equiv \begin{bmatrix} \mathbf{S}_I & \mathbf{S}_{IJ} \\ \mathbf{S}_{IJ}^T & \mathbf{S}_{IJ}^T\mathbf{S}_I^{-1}\mathbf{S}_{IJ} \end{bmatrix}.$$

Since the approximation reconstructs the submatrices $\mathbf{S}_I$ and $\mathbf{S}_{IJ}$ perfectly, the approximation error is characterized entirely by

$$\overline{\mathbf{S}}_I \equiv \mathbf{S}_J - \mathbf{S}_{IJ}^T\mathbf{S}_I^{-1}\mathbf{S}_{IJ},$$

which is the **Schur complement** of $\mathbf{S}_I$ in $\mathbf{S}$. Also, note that the Nyström approximation $\widehat{\mathbf{S}}_k^*$ can be related to the original matrix through a permutation, i.e. if $\mathbf{\Pi}$ is the matrix that maps $\{1, \ldots, p\}$ to $\{I, J\}$, then the Nyström approximation to $\mathbf{S}$ is

$$\widehat{\mathbf{S}}_k(I) = \mathbf{\Pi}^T\widehat{\mathbf{S}}_k^*(I)\,\mathbf{\Pi}.$$

3. WISHART MATRICES

As described earlier, analysis of the Nyström method in the literature to date has been limited to deterministic matrices. Here we adopt as a natural starting point the family of Wishart matrices, defined as follows.

Definition 2 (Wishart distribution)**.** *For $n \geq p$, let $\mathbf{X}$ be a $p \times n$ matrix whose columns are i.i.d. multivariate normal random vectors with zero mean and covariance $\mathbf{\Sigma} \succ 0$. Then, the $p \times p$ random matrix $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ is strictly positive definite with probability one, and follows a **Wishart distribution** with n degrees of freedom and parameter $\mathbf{\Sigma}$, denoted $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$.*

Because of their fundamental connection to the multivariate normal distribution, Wishart matrices possess many useful properties [11], the following of which we require here.

Linear transformations. Let $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$, and let $\mathbf{A}$ be a $q \times p$ matrix of rank $q \leq p$. Then,

$$\mathbf{A}\mathbf{S}\mathbf{A}^T \sim \mathcal{W}_q\big(n, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\big). \tag{1}$$

Diagonal elements. Let $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$ for integer $n \geq p$ and $\mathbf{\Sigma} = \{\sigma_{ij}\} \succ 0$, and let χ_n^2 be a chi-square random variable with n degrees of freedom. Then,

$$s_{ii} \overset{\mathcal{D}}{=} \sigma_{ii}\chi_n^2, \quad i \in \{1, \ldots, p\}. \tag{2}$$

Schur complement. Let $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$ for integer $n \geq p$ and $\mathbf{\Sigma} \succ 0$. Then,

$$\overline{\mathbf{S}}_I \sim \mathcal{W}_{p-k}\left(n-k, \overline{\mathbf{\Sigma}}_I\right), \tag{3}$$

where $\overline{\mathbf{S}}_I$ is the Schur complement of $\mathbf{S}_I$ in $\mathbf{S}$ and $\overline{\mathbf{\Sigma}}_I$ is the Schur complement of $\mathbf{\Sigma}_I$ in $\mathbf{\Sigma}$.

4. NYSTRÖM APPROXIMATION OF WISHART MATRICES

Let $\mathbf{S} = \mathbf{X}\mathbf{X}^T \sim \mathcal{W}_p(n, \mathbf{\Sigma})$ for integer $n \geq p$ and covariance $\mathbf{\Sigma} \succ 0$. Using the Nyström method, we wish to construct $\widehat{\mathbf{S}}_k(I)$, a rank-$k$ approximation to $\mathbf{S}$. We compute the approximation error in terms of the matrix trace norm, defined for a matrix $\mathbf{A}$ as

$$\|\mathbf{A}\|_{\text{tr}} = \text{tr}\Big(\sqrt{\mathbf{A}^T\mathbf{A}}\Big) = \text{tr}(\mathbf{A}),\ \mathbf{A} \succeq 0.$$

This norm is a suitable choice for several reasons. First, it is **unitarily invariant**, meaning it is invariant to orthogonal transformations and thus is characterized entirely by the eigenvalues of its argument. Second, the trace norm is dominant among all unitarily invariant norms [8], and so a small approximation error with respect to the trace norm will guarantee a small error with respect to many other norms of interest. Finally, in the case where $\mathbf{A} \succeq 0$, we have $\mathbf{A} = \mathbf{B}\mathbf{B}^T$ for some matrix $\mathbf{B}$, and the trace norm is related to the Frobenius norm by

$$\|\mathbf{A}\|_{\text{tr}} = \text{tr}(\mathbf{A}) = \text{tr}(\mathbf{B}\mathbf{B}^T) = \|\mathbf{B}\|_F^2.$$

This last property has important implications in the Wishart case, where the trace norm characterizes the Frobenius norm of the underlying multivariate-normal random variables.

4.1. Conditional error statistics

As we continue, we focus on two questions. First, given a particular index set I, what are the *conditional* statistics of the approximation error? Second, given a particular method for choosing index sets—such as a sampling from a prescribed mass function $P(I \mid \mathbf{\Sigma})$—what are the *marginal* statistics of the approximation error? The following result answers the first question.

Theorem 1 (Conditional error distribution)**.** *Let $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$ for integer $n \geq p$ and $\mathbf{\Sigma} \succ 0$, and let $\widehat{\mathbf{S}}_k(I)$ be the rank-k Nyström approximation to $\mathbf{S}$ given an index set $I \subseteq \{1, \ldots, p\}$. Then, conditioned on I and $\mathbf{\Sigma}$, the approximation error*

$$\delta(\mathbf{S}, I) \equiv \frac{\|\mathbf{S} - \widehat{\mathbf{S}}_k(I)\|_{\text{tr}}}{n} \overset{\mathcal{D}}{=} \frac{1}{n}\sum_{m=1}^{p-k} \lambda_m(\overline{\mathbf{\Sigma}}_I)\,\gamma_m,$$

where $\{\gamma_m\}$ are i.i.d. chi-square random variables with $n-k$ degrees of freedom, and $\lambda_m(\overline{\mathbf{\Sigma}}_I)$ is the m-th eigenvalue of $\overline{\mathbf{\Sigma}}_I$ for $m = 1, \ldots, p-k$.

Proof. Let $\mathbf{S}$ and $\widehat{\mathbf{S}}_k(I)$ be defined as above. The approximation error is

$$\delta(\mathbf{S}, I) = \frac{\operatorname{tr}\left[\mathbf{S} - \widehat{\mathbf{S}}_k(I)\right]}{n} = \frac{\operatorname{tr}\left(\overline{\mathbf{S}}_I\right)}{n},$$

where $\overline{\mathbf{S}}_I$ is the Schur complement of $\mathbf{S}_I$ in $\mathbf{S}$. By (3),

$$\overline{\mathbf{S}}_I \sim \mathcal{W}_{p-k}\left(n-k, \overline{\mathbf{\Sigma}}_I\right).$$

Let $\overline{\mathbf{\Sigma}}_I = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda}$ and $\mathbf{U}$ are $(p-k)\times(p-k)$ matrices containing the eigenvalues and corresponding eigenvectors of $\overline{\mathbf{\Sigma}}_I$. By (1),

$$\mathbf{U}^T\overline{\mathbf{S}}_I\mathbf{U} \sim \mathcal{W}_{p-k}\left(n-k, \mathbf{\Lambda}\right),$$

and by the cyclic property of the trace,

$$\operatorname{tr}\left(\mathbf{U}^T\overline{\mathbf{S}}_I\mathbf{U}\right) = \operatorname{tr}\left(\overline{\mathbf{S}}_I\mathbf{U}\mathbf{U}^T\right) = \operatorname{tr}\left(\overline{\mathbf{S}}_I\right).$$

Since $\mathbf{\Lambda}$ is diagonal, the diagonal elements of $\mathbf{U}^T\overline{\mathbf{S}}_I\mathbf{U}$ are independent, and thus by (2),

$$\operatorname{tr}\left(\overline{\mathbf{S}}_I\right) \overset{\mathcal{D}}{=} \sum_{m=1}^{p-k} \lambda_m(\overline{\mathbf{\Sigma}}_I)\,\gamma_m. \qquad \square$$

Due to the independence of the chi-square random variables in Theorem 1, it is straightforward to obtain the following corollary.

Corollary 2. *The conditional mean and variance of $\delta(\mathbf{S}, I)$ are given by*

$$\mathbb{E}[\delta(\mathbf{S}, I) \mid I, \mathbf{\Sigma}] = \frac{n-k}{n}\sum_{m=1}^{p-k}\lambda_m(\overline{\mathbf{\Sigma}}_I) = \frac{n-k}{n}\|\overline{\mathbf{\Sigma}}_I\|_{\text{tr}},$$

$$\operatorname{Var}[\delta(\mathbf{S}, I) \mid I, \mathbf{\Sigma}] = \frac{2(n-k)}{n^2}\sum_{m=1}^{p-k}\lambda_m^2(\overline{\mathbf{\Sigma}}_I) = \frac{2(n-k)}{n^2}\|\overline{\mathbf{\Sigma}}_I\|_F^2.$$

Although a weighted sum of chi-square random variables does not admit a density function in closed form, we can compute its conditional moment-generating function, given by

$$M_\delta(t \mid I, \mathbf{\Sigma}) \equiv \prod_{m=1}^{p-k} M\big(\lambda_m(\overline{\mathbf{\Sigma}}_I)\,t\big),$$

where $M(t)$ is the moment-generating function of a chi-square random variable with $n-k$ degrees of freedom.

4.2. Marginal error statistics

We now address the second question: given a particular method for choosing index sets, what are the marginal statistics of the approximation error? Assuming we have full knowledge of $\mathbf{\Sigma}$, define a mass function $P(I \mid \mathbf{\Sigma})$ from which to sample an index set $I \subseteq \{1, \ldots, p\}$. Then, we can calculate marginal statistics via iterated expectation:

$$\mathbb{E}[\delta(\mathbf{S}, I) \mid \mathbf{\Sigma}] = \mathbb{E}_I[\mathbb{E}(\delta \mid I, \mathbf{\Sigma})] = \frac{n-k}{n}\mathbb{E}_I\|\overline{\mathbf{\Sigma}}_I\|_{\text{tr}}, \qquad (4)$$

$$\begin{aligned}\operatorname{Var}[\delta(\mathbf{S}, I) \mid \mathbf{\Sigma}] &= \mathbb{E}_I[\operatorname{Var}(\delta \mid I, \mathbf{\Sigma})] + \operatorname{Var}_I[\mathbb{E}(\delta \mid I, \mathbf{\Sigma})] \\ &= \frac{2(n-k)}{n^2}\mathbb{E}_I\|\overline{\mathbf{\Sigma}}_I\|_F^2 + \frac{(n-k)^2}{n^2}\operatorname{Var}_I\|\overline{\mathbf{\Sigma}}_I\|_{\text{tr}}.\end{aligned}$$

Because of the combinatorial number of Schur complements involved, in general it is difficult to evaluate these expressions further. However, when $\mathbf{\Sigma}$ is diagonal and $P(I \mid \mathbf{\Sigma})$ is uniform, we can simplify (4) using an existing bound on the Nyström approximation error that applies to *any* (not necessarily random) positive definite matrix.

Theorem 3 (Uniform sampling bound [8])**.** *Let $\mathbf{\Sigma} \succeq 0$ be a $p \times p$ matrix, and $\widehat{\mathbf{\Sigma}}_k$ be the rank-k Nyström approximation to $\mathbf{\Sigma}$ given an index set $I \subseteq \{1, \ldots, p\}$. If I is chosen randomly with uniform probability, then*

$$\mathbb{E}_I\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_k\|_{\text{tr}} = \mathbb{E}_I\|\overline{\mathbf{\Sigma}}_I\|_{\text{tr}} \leq \frac{p-k}{p}\operatorname{tr}(\mathbf{\Sigma}),$$

where $\overline{\mathbf{\Sigma}}_I$ is the Schur complement of $\mathbf{\Sigma}_I$ in $\mathbf{\Sigma}$.

One can show that this bound is sharp, achieving equality for diagonal $\mathbf{\Sigma}$. Thus,

$$\mathbb{E}(\delta \mid \mathbf{\Sigma}) = \frac{n-k}{n}\mathbb{E}_I\|\overline{\mathbf{\Sigma}}_I\|_{\text{tr}} = \frac{n-k}{n}\frac{p-k}{p}\operatorname{tr}(\mathbf{\Sigma}). \qquad (5)$$

4.3. Data-dependent sampling

For many applications, we may not have knowledge of the covariance parameter $\mathbf{\Sigma}$; in fact, the purpose of $\mathbf{S}$ is often to act as an (unnormalized) estimate of an unknown $\mathbf{\Sigma}$. In these cases, sampling indices directly from a mass function $P(I \mid \mathbf{\Sigma})$ is impossible.

Alternatively, assume that we can sample from a data-dependent mass function $P(I \mid \mathbf{S})$. Then, the marginal distribution of the index sets given $\mathbf{\Sigma}$ is

$$P(I \mid \mathbf{\Sigma}) = \int_{\mathbf{S}} P(I \mid \mathbf{S})\,\mathcal{W}_p(\mathbf{S} \mid n, \mathbf{\Sigma})\,d\mathbf{S} = \mathbb{E}_{\mathbf{S}}[P(I \mid \mathbf{S})]. \qquad (6)$$

In other words, if we cannot sample directly from $P(I \mid \mathbf{\Sigma})$, we can achieve equivalent performance (in distribution) by using a data-dependent mass function $P(I \mid \mathbf{S})$—provided the latter produces the same expected proportion of samples (with respect to the ensemble of $\mathbf{S}$). As discussed in the next section, this property is true of several sampling distributions of interest.

5. DISCUSSION AND EXAMPLE

We return now to our original discussion of high-dimensional data analysis using spectral methods. As stated earlier, spectral methods cannot be applied to data of sufficiently large size or dimension without some means of reducing the computational burden. We consider the Nyström method as one approach to this problem, noting that its effective use relies upon selecting an appropriate subset of coordinates. The importance of the subset selection problem is evident in the Wishart case, where we have shown that given an index set of coordinates I, both the mean and variance of the approximation error scale as functions of norms of $\overline{\mathbf{\Sigma}}_I$.

To support this point, let us consider a small-scale example. Assume we want to approximate matrices drawn from $\mathcal{W}_p(n, \mathbf{\Sigma})$ with $p = 10$, $n = 20$, and $\mathbf{\Sigma} = \operatorname{diag}\{10, 10, 10, 10, 10, 1, 1, 1, 1, 1\}$. This choice of $\mathbf{\Sigma}$ represents a standard covariance type known in multivariate statistics as the "spiked covariance" model [12], wherein a $p \times p$ covariance is dominated by $q < p$ large eigenvalues. Under this model, we should expect that for $\mathbf{S} \sim \mathcal{W}_p(n, \mathbf{\Sigma})$, a highly effective sampling approach will yield a small Nyström approximation error when the the rank $k \geq 5$, because the missing coordinates will be characterized only by small eigenvalues.

We compare the performance of the Nyström method given three different coordinate sampling approaches. The first is **uniform sampling**, where all index sets are equally likely. Intuitively, one should

expect this method to perform least well of the three, since it takes no information from the model or data into account. The second approach is **trace sampling**, where

$$P(I \mid \mathbf{\Sigma}) \propto \mathrm{tr}(\mathbf{\Sigma}_I). \tag{7}$$

Because this mass function places more weight on submatrices with large diagonal elements, one should expect it to perform better than uniform sampling. The third approach is **determinant sampling**, where

$$P(I \mid \mathbf{\Sigma}) \propto \det(\mathbf{\Sigma}_I). \tag{8}$$

This mass function corresponds to the "annealed determinantal distributions" of [8], with the annealing exponent set to unity. Note that since p is small in our case, we can compute the normalizing coefficients of (7) and (8) in order to sample from them directly. However, for larger matrices these values are typically too costly to compute, and thus samples must be drawn using an approximate method such as the Metropolis algorithm of [10].

Also note that while we consider sampling approaches that depend on $\mathbf{\Sigma}$, we could easily condition instead on the observed data, sampling according to the trace or determinant of $\mathbf{S}_I$. It can be shown, though, that both approaches satisfy (6), and thus we are guaranteed to observe the same marginal distribution of approximation errors, regardless of whether we condition on $\mathbf{\Sigma}$ or $\mathbf{S}$.

In analyzing this example, we sampled 10^4 matrices from $\mathcal{W}_p(n, \mathbf{\Sigma})$. For each matrix, we sampled index sets according to each of the three methods and constructed the corresponding rank-k Nyström approximations, for $k = 1, \ldots, p$. Because p is small, we also computed for each matrix the optimal Nyström approximation given the minimum-error index set (obtained via exhaustive search). We compare these results to the optimal rank-k approximation obtained by performing PCA on each matrix and retaining only the first k principal components.

Figure 1 shows the mean of the normalized error of the approximations as a function of k for the three sampling approaches, the optimal Nyström approximation, and the optimal approximation given by PCA. As predicted under the spiked covariance model, the optimal error decreases rapidly, becoming small for $k \geq 5$. Of the three methods, this behavior is best approximated by the determinant sampling approach. Also as expected, the uniform sampling approach performs the least well, though the trace sampling does not perform much better. Finally, note that the uniform sampling error matches the performance predicted by (5). From these results, it is clear that the choice of sampling approach can have a significant effect on the performance of the Nyström approximation, most notably when using determinant sampling.

6. SUMMARY

In this paper, we discussed the role of approximation techniques in large-scale data analysis, in particular when the use of spectral methods is restricted by the size and dimensionality of the data. We focused on the Nyström method, which generates a low-rank reconstruction of a positive definite matrix given a subset of coordinates. Through studying the Nyström approximation of random matrices from the Wishart ensemble, we investigated the effect of the choice of coordinate set on the approximation error. This approach allowed us to provide a complete statistical characterization of the approximation error given a choice of index set, as well as compare specific methods for choosing this set. Our results indicate that of the three methods discussed (uniform, trace, and determinant sampling), choosing a submatrix with probability proportional to its determinant

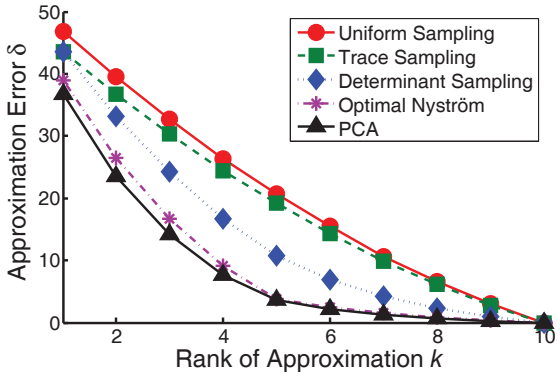

Fig. 1. Average approximation error for various sampling methods.

results in expected error performance that most closely approaches that of the optimal Nyström approximation and of PCA.

7. REFERENCES

[1] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[2] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[3] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[4] D. L. Donoho and C. Grimes, "Hessian eigenmaps: locally linear embedding techniques for high-dimensional data," in *Proc. Natl. Acad. Sci. USA*, 2003, vol. 100, pp. 5591–5596.

[5] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps," in *Proc. Natl. Acad. Sci. USA*, 2005, vol. 102, pp. 7426–7431.

[6] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Neural Inform. Process. Syst.*, 2000, vol. 13, pp. 682–688.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 1–12, 2004.

[8] M.-A. Belabbas and P. J. Wolfe, "On landmark selection and sampling in high-dimensional data analysis," *Phil. Trans. R. Soc. A*, vol. 367, no. 1906, pp. 4295–4312, 2009.

[9] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *J. Mach. Learn. Res.*, vol. 6, pp. 2153–2175, 2005.

[10] M.-A. Belabbas and P. J. Wolfe, "Spectral methods in machine learning: new strategies for very large data sets," in *Proc. Natl. Acad. Sci. USA*, 2009, vol. 106, pp. 369–374.

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

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