Find the dimension that counts: Fast dimension estimation and Krylov PCA

Shashanka Ubaru* Abd-Krim Seghouane† Yousef Saad‡

Abstract

High dimensional data and systems with many degrees of freedom are often characterized by covariance matrices. In this paper, we consider the problem of simultaneously estimating the dimension of the principal (dominant) subspace of these covariance matrices and obtaining an approximation to the subspace. This problem arises in the popular principal component analysis (PCA), and in many applications of machine learning, data analysis, signal and image processing, and others. We first present a novel method for estimating the dimension of the principal subspace. We then show how this method can be coupled with a Krylov subspace method to simultaneously estimate the dimension and obtain an approximation to the subspace. The dimension estimation is achieved with no additional cost. The proposed method operates on a model selection framework, where the novel selection criterion is derived based on random matrix perturbation theory ideas. We present theoretical analyses which (a) show that the proposed method achieves strong consistency (i.e., yields optimal solution as the number of data-points $n \to \infty$), and (b) analyze conditions for exact dimension estimation in the finite n case. Using recent results, we show that our algorithm also yields near optimal PCA. The proposed method avoids forming the sample covariance matrix (associated with the data) explicitly and computing the complete eigen-decomposition. Therefore, the method is inexpensive, which is particularly advantageous in modern data applications where the covariance matrices can be very large. Numerical experiments illustrate the performance of the proposed method in various applications.

1 Introduction

In many applications, for a given set of data observations, covariance matrices are used to capture the interactions in high dimensions, among the many degrees of freedom. A popular approach to analyze such high dimensional data is to look for the principal (components) subspace of the covariance matrix, which is of much lower dimension. For this, it is often required to first estimate the dimension of this principal (dominant) subspace of the covariance matrix associated with the observations [33, 18, 5, 19, 31]. These observations can be treated as high dimensional random quantities embedded in noise.

Low rank approximation is a popular tool used in applications to reduce high dimensional data [16, 10, 17, 30]. Determining the lower dimension (rank k) remains a principal problem in these applications, see [31, 32] for discussions. In statistical signal and array processing, detecting the number of signals in the observations of an array of passive sensors is a fundamental problem [33, 19, 23], which can be posed as the above dimension estimation problem. Similar estimation problems occur in many other fields such as chemo-metrics [20, 18], econometrics and statistics [5], population genetics [24], and reduced rank regression models [4]. Moreover, in most of these applications, once the dimension of the principal subspace (approximate rank) is estimated, it is also desired to obtain an approximation for this principal subspace, e.g., in principal component analysis (PCA) [16, 17], subspace tracking [7] and others. Krylov subspace based methods [27] are the most popular and effective methods used in the literature to compute an approximation for the principal subspace, see [34, 28, 13, 22, 25] for examples.

Prior Work: The problem of estimating the rank or the dimension of the principal subspace has been studied in various fields, and a few different methods have been proposed in the literature. In signal processing, information theory criteria based methods have been proposed for the detection of number of signals [33, 23]. A few hypothesis testing based methods have also been proposed for dimension estimation, see [34, 24, 18, 19]. In econometrics and statistics, various tests and methods have been proposed to estimate the rank and the rank statistic of a matrix, see, e.g., [26, 9, 5].

However, most of these methods require computing the complete eigen-decomposition of the sample covariance matrix, which becomes impractical for large dimensional matrices, e.g., in modern data applications and for large aperture arrays in array signal processing. Even forming the covariance matrix is inviable in many cases. The information criteria based methods are not

*IBM T. J. Watson Research Center, Yorktown Heights, NY, USA. Shashanka.Ubaru@ibm.com.

†The University of Melbourne, Melbourne, Victoria, Australia. abd-krim.seghouane@unimelb.edu.au

‡University of Minnesota, Twin Cities, MN, USA. saad@umn.edu

applicable when the data dimension p is larger than the number of observations n, i.e., when $p > n$. Recently, a set of inexpensive methods were proposed for numerical rank estimation of data matrices [31, 32]. These methods combine ideas such as stochastic trace estimator, eigen-projectors and spectral densities to compute the rank inexpensively without any matrix decomposition. However, methods that simultaneously estimate the dimension and obtain an approximation to the principal subspace are lacking.

Contributions: In this work, we present a novel method for estimating the dimension of the principal subspace of covariance matrices. The method can be combined with the Krylov subspace methods (Krylov PCA) to compute an approximation to the principal subspace, simultaneously. The method operates on a model selection framework, and the proposed selection criterion requires computing only the top k eigenvalues of the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is the matrix containing n observed data of dimension p, for a given integer $k \ll \{n, p\}$. In order to compute these eigenvalues, we can use the popular Lanczos algorithm [27] which requires only matrix-vector products with $\mathbf{S}_n$. Hence, we do not have to form the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, explicitly. Our approach can be viewed as a stopping criterion for the Krylov subspace methods, and we can simultaneously estimate the dimension and compute the principal subspace at no additional cost.

The proposed selection criterion is derived using random matrix perturbation theory results [21], see section 3. The criterion also includes a penalty (function) term which under mild assumptions yields us a strongly consistent estimator, i.e., the method estimates the exact dimension as the number of data observations $n \to \infty$. We establish this strong consistency for the proposed method and also present performance analysis in section 4. We derive conditions on the signal strength and the noise level for avoiding incorrect dimension estimation in the finite n case, using random matrix theory results [14]. Using the recent results in [22], we also show that the method yields near optimal PCA, and the consistency results and the performance analysis hold for eigenvalues computed by the Krylov subspace methods. Numerical experiments illustrate the performance of the proposed method in the number of signals detection application, numerical rank estimation of general data matrices, and in video foreground detection, an application of PCA.

2 Preliminaries

We begin by presenting the problem formulation for dimension estimation of the principal subspace.

Notation: We use lowercase and uppercase bold letters, $\mathbf{x}$ and $\mathbf{A}$ for vectors and matrices, respectively. The Gaussian distribution with mean μ and covariance $\mathbf{\Sigma}$ is denoted by $\mathcal{N}(\mu, \mathbf{\Sigma})$. Identity matrix is depicted as $\mathbf{I}_p$, where p is the order. Convergence in distribution is denoted by $\to_d$.

Problem Formulation: The data observations which form the matrix $\mathbf{X}$ are typically modeled as high dimensional random quantities embedded in noise. We assume the standard Gaussian random model for the set of n data observations each of dimension p. We denote the p-dimensional data as $\{\mathbf{x}_i\}_{i=1}^n$ described as

$$\mathbf{x}_i = \mathbf{M}\mathbf{s}_i + \sqrt{\sigma}\mathbf{n}_i,\ i = 1, \ldots, n \tag{2.1}$$

where $\mathbf{M}$ is a $p \times q$ mixing matrix with q independent columns, $\mathbf{s}_i$ are $q \times 1$ vectors containing the zero mean relevant data and $\mathbf{n}_i$ are p-dimensional Gaussian (white) noise vectors with parameter σ as the unknown noise variance. This is a standard assumption made in PCA [16], probabilistic PCA [29], signal detection and subspace tracking [33, 34], and in modern data analysis [3] and neural networks [11] methods. The true covariance matrix $\mathbf{\Sigma}$ associated with the underlying data is then assumed to be a low rank matrix of rank q, perturbed by noise of variance σ. That is,

$$\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T + \sigma\mathbf{I}_p,$$

where $\mathbf{B} \in \mathbb{R}^{p \times q}, q \ll p$ and $span(\mathbf{B})$ is the principal subspace. The top q eigenvalues λ_i for $i = 1, \ldots, q$ of $\mathbf{\Sigma}$ will correspond to the q dimensional relevant data and the remaining $p - q$ eigenvalues are related to noise and are equal to σ. Hence, the subspace associated with the top q eigenvectors (eigenvalues) forms the principal subspace, which is of interest.

The exact covariance matrix of the underlying data will not be available, and hence we consider the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, using the n (noisy) observations of the data. We wish to estimate q, the dimension of (relevant) data in the observations, using the eigenvalues of the sample covariance matrix $\mathbf{S}_n$ denoted by $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$.

3 Proposed Method

In this section, we first present the proposed method for the principal subspace dimension estimation and derive it. We then discuss the Krylov subspace methods for computing partial eigen-decomposition of matrices, and present the proposed algorithm for simultaneous estimating the dimension and computing an approximation to the principal subspace.

The novel method is based on model selection technique and the proposed criterion is the following:

(3.2)
$$\arg\min_k \left[\frac{n}{2\sigma^2} \sum_{i=k+1}^{p} (\ell_i - \sigma)^2 - C_n \frac{(p-k)(p-k-1)}{2} \right]$$

where ℓ_i, for $i = 1, \ldots, p$ are the eigenvalues of the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, σ is the noise variance, and C_n is a parameter that depends on n (see sec. 4 for details). Note that the first term in the criterion depends on the sum of bottom $p-k$ eigenvalues of $\mathbf{S}_n$, which can be written as

$$\sum_{i=k+1}^{p} (\ell_i - \sigma)^2 = \|\mathbf{S}_n - \sigma \mathbf{I}_p\|_F^2 - \sum_{i=1}^{k} (\ell_i - \sigma)^2.$$

Thus, the method requires computing only the top k eigenvalues of $\mathbf{S}_n$. We can compute the norm as $\|\mathbf{S}_n - \sigma\mathbf{I}_p\|_F^2 = \frac{1}{n^2}\|\mathbf{X}\|_F^4 - \frac{2\sigma}{n}\|\mathbf{X}\|_F^2 + p\sigma^2$. Therefore, if Krylov subspace method such as the Lanczos algorithm [27] is used for computing these eigenvalues, then we do not need to form $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ explicitly.

The Krylov subspace methods will also yield us an approximation to the eigenvectors corresponding to the computed eigenvalues. Therefore, we can use the above method as a stopping criterion for the Krylov subspace methods, and hence, estimate the dimension and approximate the principal subspace of the covariance matrix, simultaneously. We present the resulting algorithm in the latter part of this section. First, we derive the above criterion using concepts from random matrix perturbation theory.

3.1 Derivation We start the derivation of the proposed selection criterion using the following key concept from random matrix theory [21]: The sample covariance matrix $\mathbf{S}_n$ approaches the true covariance matrix $\mathbf{\Sigma}$ only in the expectation, i.e., $\mathbb{E}[\mathbf{S}_n] \to \mathbf{\Sigma}$. More importantly, the sample covariance matrix $\mathbf{S}_n$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Sigma}$.

PROPOSITION 3.1. *$\mathbf{S}_n$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Sigma}$. That is,*

$$\sqrt{n}\mathtt{vec}(\mathbf{S}_n - \mathbf{\Sigma}) \to_d \mathcal{N}(0, \mathbf{\Omega}),$$

where $\mathbf{\Omega} = (I + P_{\mathtt{vec}(\mathbf{S}_n)})(\mathbf{\Sigma} \otimes \mathbf{\Sigma})$ is a $p^2 \times p^2$ covariance matrix with $\otimes$ denoting the Kronecker product and $P_{\mathtt{vec}(\mathbf{S}_n)}$ the transposition-permutation matrix associated to $\mathtt{vec}(\mathbf{S}_n)$.

The proof of this proposition can be found in most standard multivariate statistical theory textbooks, e.g., [2, 21].

Next, we consider the eigen-decomposition of the covariance matrix $\mathbf{\Sigma} = \mathbf{U}\Lambda\mathbf{U}^T$. Let us write $\mathbf{U} = [\mathbf{U}_q, \mathbf{U}_{p-q}]$, where $\mathbf{U}_q$ is a matrix containing the top q eigenvectors (principal subspace) of $\mathbf{\Sigma}$ as columns. Similarly, let us consider the eigen-decomposition of the sample covariance matrix $\mathbf{S}_n = \mathbf{G}\mathbf{L}\mathbf{G}^T$, with $\mathbf{G}_q$ containing the top q eigenvectors of $\mathbf{S}_n$ as columns. We can then prove the consistency of $\mathbf{G}_q$ using the random matrix perturbation approach on $\mathbf{S}_n$.

PROPOSITION 3.2. *Let q be the numerical rank of $\mathbf{\Sigma}$ and we assume that the smallest eigenvalue corresponding to data is well above zero, i.e., $\lambda_q > \varepsilon > 0$ for a small ε. Then as $n \to \infty$,*

$$\mathbf{G}_q \to_d \mathbf{U}_q.$$

A version of proof of this proposition is given in supplementary, which was first derived in [1]. We then have the following result (proof in supplementary).

COROLLARY 3.1. *The orthogonal projector onto the space spanned by the eigenvectors corresponding to the noise related eigenvalues satisfy*

$$\mathbf{Q}_G = \mathbf{G}_{p-q}\mathbf{G}_{p-q}^T = \mathbf{U}_{p-q}\mathbf{U}_{p-q}^T + O_p\left(\frac{1}{\sqrt{n}}\right).$$

We next have the following result that gives the asymptotic behavior of the bottom $p-q$ eigenvalues of $\mathbf{S}_n$.

PROPOSITION 3.3. *The asymptotic distribution of $\sqrt{n}\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma I)\mathbf{Q}_G)$ is given by*

$$\sqrt{n}\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G) \to \mathcal{N}(0, \hat{\mathbf{\Omega}}),$$

where $\hat{\mathbf{\Omega}} = (\mathbf{Q}_U \otimes \mathbf{Q}_U)\mathbf{\Omega}(\mathbf{Q}_U \otimes \mathbf{Q}_U)$, where $\mathbf{Q}_U = \mathbf{U}_{p-q}\mathbf{U}_{p-q}^T$ and $\mathbf{\Omega}$ is as Proposition 3.1.

We defer the proof to supplementary. This leads to the following result.

LEMMA 3.1. *Let $\mathcal{L}$ be defined as*

$$\mathcal{L} = \frac{n}{2\sigma^2} \sum_{i=q+1}^{p} (\ell_i - \sigma)^2,$$

where ℓ_i are the eigenvalues of $\mathbf{S}_n$ and σ is the noise variance. Then $\mathcal{L}$ follows asymptotically a χ^2 chi-square distribution with $\eta = \frac{1}{2}(p-q)(p-q-1)$ degrees of freedom.

Proof. Suppose L_{p-q} is a diagonal matrix with the bottom $p-q$ eigenvalues of $\mathbf{S}_n - \sigma\mathbf{I}_p$ as entries, then we have

$$\begin{aligned} n \sum_{i=q+1}^{p} (\ell_i - \sigma)^2 &= n\mathtt{tr}(L_{p-q}^2) \\ &= n\mathtt{tr}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)^2\mathbf{Q}_G) \\ &= \|\sqrt{n}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)\|_F^2 \\ &= \|\sqrt{n}\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)\|_2^2. \end{aligned}$$

From Proposition 3.3, the above sum follows asymptotically a $\eta = \frac{1}{2}(p-q)(p-q-1)$ weighted χ^2_1 distribution [2], where the η weights correspond to the first η eigenvalues of $\hat{\mathbf{\Omega}} = (\mathbf{Q}_U \otimes \mathbf{Q}_U)\mathbf{\Omega}(\mathbf{Q}_U \otimes \mathbf{Q}_U)$. Note that η is the degree of freedom in $\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G$.

Given the eigenpairs of $\mathbf{\Sigma}$ to be $(\lambda_i, \mathbf{u}_i), i = 1, \ldots, p$, the eigenpairs of $\mathbf{\Omega}$ will be $(\lambda_i * \lambda_j, \mathbf{u}_i \otimes \mathbf{u}_j), i, j = 1, \ldots, p$ from the property of Kronecker products, see [12, Thm. 4.2.12]. $\mathbf{Q}_U$ is a projector onto the span of eigenvectors corresponding to the bottom $p-q$ eigenvalues of $\mathbf{\Sigma}$, which are all equal to σ. Hence, the top η eigenvalues of $\hat{\mathbf{\Omega}}$ will be all equal to σ^2, since $(\mathbf{Q}_U \otimes \mathbf{Q}_U)$ is a projector onto space spanned by the eigenvectors corresponding to the bottom $(p-q)^2$ eigenvalues of $\mathbf{\Omega}$. Hence, the weights of the weighted χ^2 are all equal to σ^2. Thus, by reweighting the above sum, $\mathcal{L}$ will have asymptotically χ^2_η distribution[1].

Therefore, the above $\mathcal{L}(\mathbf{S}_n, q)$ can be used in model selection criterion for estimating q, the dimension of the principal subspace.

THEOREM 3.1. *The following criterion yields an estimation for the dimension q of the principal subspace of the covariance matrix $\mathbf{\Sigma}$:*

(3.3)
$$q = \arg\min_k \left[\frac{n}{2\sigma^2} \sum_{i=k+1}^{p} (\ell_i - \sigma)^2 - C_n \frac{(p-k)(p-k-1)}{2} \right],$$

where ℓ_i, for $i = 1, \ldots, p$ are eigenvalues of the sample covariance matrix $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, σ is the noise variance and C_n is a parameter that depends on n.

Proof of the theorem is given in supplementary. We also give a simulation result which shows that Lemma 3.1 and Theorem 3.1 hold true in practice.

3.2 Krylov subspace methods Krylov subspace methods are popularly used to compute the partial spectrum (top k eigenvalues and eigenvectors) of matrices [27]. Recent results [22] have shown that these methods returns high quality principal components and give nearly optimal PCA for any matrix. The proposed dimension estimation criterion can be used as a stopping criterion for such Krylov subspace approximation of the principal subspace of covariance matrices.

For a symmetric matrix $\mathbf{A}$, the Krylov subspace is defined as $\mathcal{K}^m(\mathbf{A}, \mathbf{v}) = span\{\mathbf{v}, \mathbf{A}\mathbf{v}, \ldots, \mathbf{A}^{m-1}\mathbf{v}\}$, where $\mathbf{v}$ is a random vector of unit norm, $\|\mathbf{v}\| = 1$, $\mathbf{v} \nsubseteq null(\mathbf{A})$ and m is a scalar. The Lanczos algorithm builds an orthonormal basis for this Krylov subspace [27]. We can also define a block Krylov subspace as: $\mathbb{K}^m(\mathbf{A}, \mathbf{V}) = span\{\mathbf{V}, \mathbf{A}\mathbf{V}, \ldots, \mathbf{A}^{m-1}\mathbf{V}\}$, where $\mathbf{V} \in \mathbb{R}^{p \times k}$ is a random matrix such that $\mathbf{V} \nsubseteq null(\mathbf{A})$, see [22] for recent theoretical results for randomized block Krylov subspace methods. We can compute approximate eigenvalues and eigenvectors of $\mathbf{A}$, say $\{\theta_i, \mathbf{y}_i\}_{i=1}^k$ for some k, using the Krylov subspace methods. We have the following result from eqn. 3 and Theorem 1 in [22]:

LEMMA 3.2. *Consider a symmetric PSD matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$ with eigenvalues $\ell_i, i = 1, \ldots, p$. Let $\{\theta_i, \mathbf{y}_i\}_{i=1}^k$ be the k eigenpair computed using m steps of block Krylov subspace method (using the orthonormal basis of $\mathbb{K}^m(\mathbf{A}, \mathbf{V})$ for $\mathbf{V} \in \mathbb{R}^{p \times k}$). If $m = \frac{\log(p)}{\sqrt{\epsilon}}$ for some $0 < \epsilon < 1$, then we have*

$$|\theta_i - \ell_i| \leq \epsilon \ell_{k+1},\ i = 1, \ldots, k.$$

Moreover, suppose $\mathbf{Y}_k$ is a matrix containing the eigenvectors $\{\mathbf{y}_i\}_{i=1}^k$ computed by the Krylov subspace method as columns, then we have for $\xi \in \{2, F\}$

$$\|\mathbf{A} - \mathbf{Y}_k \mathbf{Y}_k^T \mathbf{A}\|_\xi \leq (1+\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_\xi,$$

where $\mathbf{A}_k$ is the best rank k approximation of $\mathbf{A}$ obtained using its exact eigen-decomposition.

Therefore, the Krylov subspace method will return a high quality principal components of $\mathbf{S}_n$ and near

Algorithm 1 Proposed Algorithm

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, noise variance σ, parameter C_n, and a error tolerance ϵ.
Output: Dimension q and an approximation to the principal subspace $\mathbf{Y}_q$.
Set $IC = zeros(p, 1)$, $\mathbf{Q} = [\,], k = 1$, $m = \frac{\log(p)}{\sqrt{\epsilon}}$, $\Phi = \frac{1}{n^2}\|\mathbf{X}\|_F^4 - \frac{2\sigma}{n}\|\mathbf{X}\|_F^2 + p\sigma^2$.
for $k = 1$ to p **do**
1. Generate a random vector $\mathbf{v}_k$ with $\|\mathbf{v}_k\|_2 = 1$.
2. $\mathbf{K} = \frac{1}{n}[\mathbf{X}\mathbf{v}_k, (\mathbf{X}\mathbf{X}^T)\mathbf{X}\mathbf{v}_k, \ldots, (\mathbf{X}\mathbf{X}^T)^{m-1}\mathbf{X}\mathbf{v}_k]$
3. $\mathbf{Q} = orth([\mathbf{Q}, \mathbf{K}])$, $\mathbf{Q} = \mathbf{Q}(:, 1:k)$.
4. $\mathbf{T} = \frac{1}{n}\mathbf{Q}^T\mathbf{X}\mathbf{X}^T\mathbf{Q}$.
5. $[\mathbf{V}, \Theta] = \texttt{eig}(\mathbf{T})$.
6. $IC(k) = n(\Phi - \sum_{i=1}^k (\theta_i - \sigma)^2) - C_n \frac{(p-k)(p-k-1)}{2}$

if $(k > 1$ && $IC(k) > IC(k-1))$ **then**
break;
end if
end for
$q = k - 1$. Output q and $\mathbf{Y} = \mathbf{Q}\mathbf{V}$.

[1] Anderson made a similar observation (of asymptotically χ^2_η distribution) in [1] for a given eigenvalue λ_k of $\mathbf{\Sigma}$ with multiplicity q_k and the sum of eigenvalues of $(\mathbf{S}_n - \lambda_k \mathbf{I})$. In our case, $\lambda_k = \sigma$ with multiplicity $q_k = p - q$.

optimal $(1+\epsilon)$ PCA. In addition, the eigenvalues θ_i's computed are very close to the actual eigenvalues ℓ_is of the sample covariance matrix (within a multiplicative factor). The error ϵ in the above analysis is related to the gap in the spectrum, i.e., we can replace ϵ by $\frac{\ell_k}{\ell_{k+1}} - 1$, see [22, §7]. For $k > q$, the error term $\epsilon\ell_{k+1}$ is related to the noise related eigenvalues and we have $\epsilon\ell_{k+1} = O(\frac{1}{\sqrt{n}})$ from the analysis in section 3.1 and [1]. Asymptotically, this term goes to zero. Thus, θ_i's have the same statistical properties of ℓ_i's, and are good approximation to them. Since ℓ_i's are asymptotically equivalent to λ_i's, θ_i's are good estimates of λ_i's.

Proposed Algorithm: Algorithm 1 presents the proposed algorithm for simultaneously estimating the dimension and computing the principal subspace of the covariance matrix. In step 2, note that only matrix-vector products with the data $\mathbf{X}$ and its transpose are needed to form the Krylov matrix $\mathbf{K}$. In step 3, since $\mathbf{Q}$ is already orthonormal from the previous iteration, the new vectors in $\mathbf{K}$ can be quickly orthonormalized wrt. $\mathbf{Q}$. We can also replace steps 2-5, by a version of the Lanczos algorithm [27], which updates the previous subspace $\mathbf{Q}$ and the tridiagonal matrix $\mathbf{T}$.

Cost: If q is the exact dimension, the computational cost of the algorithm will be $O(\text{nnz}(\mathbf{X})qm + p(qm)^2)$, where $\text{nnz}(\mathbf{X})$ is the number of nonzeros in $\mathbf{X}$. Since both $q \ll p$ and $m = \frac{\log(p)}{\sqrt{\epsilon}}$ are small, the algorithm is quite inexpensive, more so if data $\mathbf{X}$ is sparse.

Choosing σ: In our Algorithm, we need to choose the noise level σ, when it is unknown. In many applications, e.g., in signal processing, typically an estimate of noise level is known. In low rank approximation problems, the maximum approximation error tolerance acceptable might be known. Otherwise, for signal processing applications, σ can be determined using the thresholding method proposed in [19]. For data related applications, article [32] discusses an inexpensive method to estimate σ using the spectral density plot of the data matrix. For further details, see [31, 32].

4 Analysis

In this section, we first show that the proposed method yields a strong consistent estimator for q, the exact dimension. We then analyze the conditions for correct estimation for finite n data observations.

4.1 Strong consistency

THEOREM 4.1. *The criterion defined by* (4.4)

$$IC(k) = \frac{n}{2\sigma^2}\sum_{i=k+1}^{p}(\ell_i - \sigma)^2 - C_n\frac{(p-k)(p-k-1)}{2}$$

can be used to obtain a strong consistent estimator for q, the exact dimension of the principal subspace, i.e., $\lim_{n\to\infty}\hat{k} = q$, where $\hat{k} = \arg\min_k IC(k)$, with value of C_n such that

$$\lim_{n\to\infty}\frac{C_n}{n} = 0 \text{ and } \lim_{n\to\infty}\frac{C_n}{\log\log n} = \infty.$$

Proof of this theorem is given in supplementary. For the right choice of C_n, the proposed estimator is strongly consistent. Next, we consider the eigenvalues computed using the Krylov subspace method in our criterion.

COROLLARY 4.1. *For the choice of C_n in Theorem 4.1, the criterion 3.2 is strongly consistent for the eigenvalues computed using the Krylov subspace method in Algorithm 1 if we set the parameter $\sigma = (1-\epsilon)\sigma_{true}$ in the algorithm, where σ_{true} is the true noise variance of the data.*

Proof can be found in supplementary. Next, we analyze the performance of the proposed method for finite sample size and obtain the conditions for correct detection.

4.2 Performance Analysis The consistency analysis above considered the asymptotic case when $n \to \infty$, and the law of iterated logarithm [21] is used to derive the results. Here, we analyze the performance of the proposed method for finite sample size (general n), and obtain the conditions when the method either *underestimates* or *overestimates* the dimension.

The notorious scenario for wrong detection is when the dimension is off by exactly one ($q \pm 1$), which we analyze here (important in signal detection applications). The analysis trivially generalizes to other cases. First, let us consider underestimation by one, and consider the following difference:

$$\begin{aligned}\Delta_1 &= IC(q-1) - IC(q)\\ &= \frac{n}{2\sigma^2}(\ell_q - \sigma)^2 - C_n(p-q).\end{aligned}$$

Note that we will not have underestimation when $\Delta_1 > 0$, i.e., when

$$\ell_q > \sigma\left(\sqrt{\tfrac{2C_n}{n}(p-q)} + 1\right).$$

So, we need the magnitude of ℓ_q (related to relevant data or the signal strength) to be large enough in order to avoid underestimate the dimension. That is, we need a reasonable gap between relevant eigenvalues and the noise related eigenvalues in the spectrum. For the asymptotic case ($n \to \infty$), we know that the RHS term with C_n goes to zero and, hence we will not have

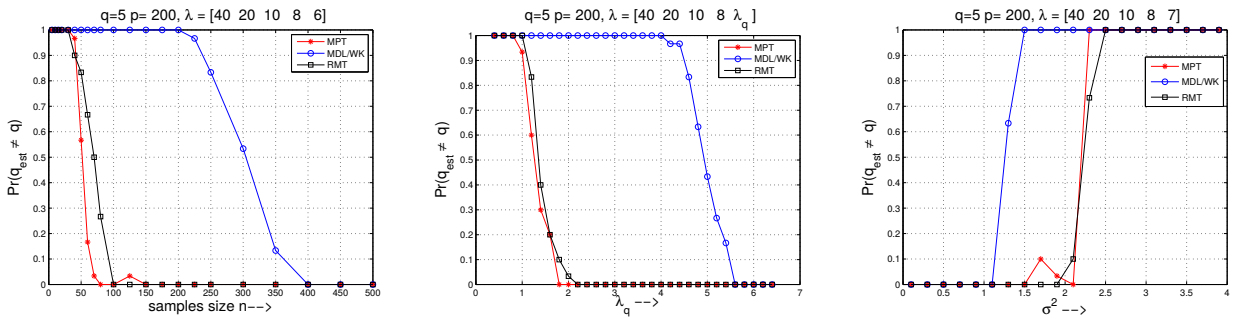


Figure 1: Signal detection: Comparison between the proposed method MPT, RMT and MDL as a function of: (left) the number of samples n, (middle) signal strength (λ_q eigenvalue), and (right) the noise level σ.

any underestimation of dimension as long as the signal strength is more than the noise variance.

Next, let us consider overestimation of the dimension by one, and the following difference:

$$\begin{aligned}\Delta_2 &= IC(q+1) - IC(q) \\ &= C_n(p-q-1) - \frac{n}{2\sigma^2}(\ell_{q+1}-\sigma)^2.\end{aligned}$$

Again, we will not overestimate if $\Delta_2 > 0$, i.e., when

$$\frac{\ell_{q+1}}{\sigma} < \sqrt{\tfrac{C_n}{n}(p-q-1)} + 1.$$

We know that ℓ_{q+1} corresponds to the largest noise related eigenvalue of the covariance matrix. For the asymptotic case ($n \to \infty$), we know $\ell_{q+1} \to \sigma$, hence the equation holds. For finite n, we must choose the noise parameter σ close to the true noise level (reflected in ℓ_{q+1}) in order to avoid overestimation. Assuming the noise variance σ is known, for finite n, when the ratio of p/n or n/p is not too large, we can derive bounds on the parameter C_n in our method to avoid overestimation, using the random matrix theory results in [14, 15].

The largest eigenvalue of the sample covariance matrix (Wishart matrix) of pure noise vectors with Gaussian distribution follows the Tracy-Widom distribution [14, 15]. Then, for finite p, n as long as $\min\{p,n\} \gg 1$ and the ratio of p/n or n/p is not too large, the largest eigenvalue due to noise will be approximately $\sigma(1+\sqrt{p/n})^2$, see [19] for details. Hence, for finite but large values of p, n, we have

$$\ell_{q+1} \approx \sigma\left(1+\sqrt{\frac{p}{n}}\right)^2.$$

Substituting in the condition above for overestimation, we get the following bound for the parameter C_n for exact detection for finite but large values of p, n:

$$C_n > \frac{(p+2\sqrt{np})^2}{n(p-q-1)}.$$

When the ratio of p/n or n/p is not too large, the RHS is fairly small. The above analysis provides us the conditions on the relevant eigenvalue ℓ_q, noise level and the parameter C_n in order to avoid incorrect estimation of the dimension q using the proposed method.

When we consider the eigenvalues obtained by the Krylov subspace method in the criterion, we will have an additional term that depends on ϵ in the denominators of the above conditions. That is, we have approximately the following conditions for exact dimension detection:

$$\ell_q > \frac{\sigma}{(1-\epsilon)}\left(\sqrt{\tfrac{2C_n}{n}(p-q)}+1\right) \text{ and}$$

$$\frac{\ell_{q+1}}{\sigma} < \frac{1}{(1-\epsilon)}\sqrt{\tfrac{C_n}{n}(p-q-1)}+1.$$

For small ϵ, we end up with similar conditions on ℓ_q, noise level and C_n as above.

5 Numerical experiments

In this section, we present some numerical experimental results to illustrate the performance of the proposed method, and compare it to few other popular methods. First, we consider examples for the number of signals detection application in signal and array processing. We then consider few large data matrices and a PCA application to illustrate the method's performance.

5.1 Number of signals detection In the first set of experiments, we consider the signal detection problem to illustrate the accuracy of the proposed method for dimension estimation (exact detection is desired in this application). The results and observations from these experiments are applicable for general data too, see supplementary. We consider p dimensional signals $\mathbf{x}_i$'s that are corrupted by white noise with $\mathcal{N}(0, \sigma\mathbf{I})$, variance σ. There are three parameters in this model, namely the number of samples n, the signal strength or the magnitude of λ_q eigenvalue, and the noise level σ. We compare the performances of the proposed method,

Table 1: Performance of the Krylov Subspace method, Algorithm 1 with $m = 10$.

Dataset	p	Actual q	λ_q	σ	Estimated $\tilde{q}$	$\|\mathbf{A} - \mathbf{Y}_{\tilde{q}}\mathbf{Y}_{\tilde{q}}^T\mathbf{A}\|_F$	Runtime
sprand	5000	50	5	1	50	134.47	6.1 secs
	5000	100	2	0.5	100	159.23	22.8 secs
	10000	100	2	0.5	100	162.52	72.5 secs
	40000	100	2	0.5	100	183.74	101.6 secs
	100000	100	2	0.5	100	210.86	192.1 secs
Harvard	500	63	2.6	1	69	36.14	0.24 secs
lpiceria3d	3576	108	5	1	104	140.52	0.68 secs
EVA	8497	165	5.2	1	172	81.47	2.90 secs
lpstocfor3	16675	981	23.7	3	981	3.05e4	2.29 secs
as-22july	22963	241	54.6	10	237	311.23	137.4 secs
internet	124651	–	–	1	351	7.49e3	797.8 secs

the MDL (Minimum Description Length) method proposed in [33], and the 'state of the art' hypothesis testing method proposed in [19] based on random matrix theory (RMT) for signal detection as a function of these three parameters. In all experiments, we set $C_n = \log n$ to ensure that the asymptotic properties and the finite sample lower bound on C_n above hold.

Figure 1 presents three results for the three methods, the proposed matrix perturbation theory (MPT) based method, the MDL method and the random matrix theory (RMT) based hypothesis testing method. For a chosen signal dimension p (reported in the plot), we generate the signals and the sample covariance matrix based on the considered signal eigenvalues λ (listed in the plot). We then add noise covariance matrix corresponding to the noise level σ considered. We plot the probability of the estimated rank q_{est} being not equal to the actual rank q, i.e., $Pr(q_{est} \neq q)$ over 100 trials. In the first plot of Fig. 1, we plot $Pr(q_{est} \neq q)$ as a function of the number of samples n. We consider small signal dimension $p = 200$ (note that MDL and RMT require complete eigen-decomposition), the actual rank $q = 5$ and the noise level $\sigma = 1.1$. The eigenvalues corresponding to the signals are given in the plot. We note that MDL requires $n \geq p$ to yield exact rank, where as the proposed method MPT yields exact rank for much smaller sample size, and performs even slightly better than the state of the art method RMT which requires all the eigenvalues of the sample covariance matrix.

In the second (middle) plot, we compare the performances wrt. the signal strength, i.e., the magnitude of the qth eigenvalue λ_q of the covariance matrix. Again the signal dimension is $p = 200$, the actual rank $q = 5$ and the noise level $\sigma = 1.1$. The number of samples is $n = 400$. We note that, the proposed method again outperforms MDL and yields more accurate results for much lower signal strength. In the last plot, we compare the performances with respect to the noise level σ. Here too, the signal dimension is $p = 200$, the actual rank $q = 5$ and the number of samples is $n = 400$. The signal eigenvalues are given in the plot and the signal strength $\lambda_q = 6$. The proposed method MPT performs better than MDL wrt. the noise level too and performs was well as RMT. RMT requires parameters, such as confidence level α to be selected. More importantly, both MDL and RMT require computing all the eigenvalues of the sample covariance matrix. Results for our algorithm 1 are reported in the supplementary.

5.2 Data matrices Next, we illustrate the performance of the proposed method for numerical rank estimation of data matrices. We consider general data matrices that have low numerical rank from publicly available database, SuitSparse [8], and a few synthetic sparse random matrices. For these matrices, the Gaussian type distribution assumptions for the data and noise may not hold. We report additional comparative results in the supplementary.

Table 1 presents the performance of the Krylov Subspace method, i.e., Algorithm 1 for dimension estimation and approximation of the principal subspace. The synthetic sparse random matrices are of the form $\mathbf{X} = \mathbf{B}\Lambda\mathbf{B}^T + \mathbf{N}$, where $\mathbf{B}$ is a sparse (relevant) data matrix (unit column norm) of size $p \times q$ (sparsity $\text{nnz}(\mathbf{B})/pq = [0.05, 0.1]$), Λ is a diagonal matrix with the smallest diagonal entry equal to λ_q listed in the 4th column. $\mathbf{N}$ is a Gaussian sparse random matrix with σ listed in fifth column. The number of Lanczos steps per iteration (for each k) is $m = 10$. The exact dimension q and the estimated dimension $\tilde{q}$ are reported (dimension estimation), along with the Frobenius norm error$\|\mathbf{A} - \mathbf{Y}_{\tilde{q}}\mathbf{Y}_{\tilde{q}}^T\mathbf{A}\|_F$, evaluating the quality of approximation to the principal subspace. The runtime of the algorithm is also reported (computed using

Figure 2: *Background subtraction:* for two sample images from two video datasets. Low rank approximation (mean added) and foreground detection with eigenvectors from proposed method and exact eigenvectors.

`cputime` function on an Intel i-5 3.4GHz machine). For the synthetic examples, we vary the parameters: size p, rank q, data strength λ_q and noise level σ, and report the results. We also consider a few sparse data matrices (also see supplementary). We report matrices that have smaller numerical rank ($q \ll \min(n, p)$) and a reasonable gap in the spectrum. The Krylov subspace algorithm works well only when there is a spectral gap. Otherwise, the interior eigenvalues do not converge. For large matrix 'internet', we do not know the exact rank (cannot compute complete decomposition). We observe that the algorithm performs reasonably well for these matrices. The method is also quite inexpensive, particularly for large sparse data matrices.

5.3 Video Foreground Detection In the last experiment, we consider an application of PCA, that of background subtraction in surveillance videos. Here, PCA is used to separate the foreground information from the background noise. We consider two videos datasets: "Lobby in an office building with switching on/off lights" and "Shopping center" available from `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`. Here we illustrate how the proposed Krylov method can be used to obtain an appropriate dimension of the principle subspace (components) to be used for background subtraction, and use the approximate principal components obtained from the algorithm in the application [6].

The Lobby video contains 1546 frames each of size 160×128, and the data matrix size is 1546×20480. Second video is from a shopping mall with 1286 frames each of resolution 320×256. So, the data matrix is of size 1286×81920. This video contains more activities than Lobby video with many people moving in and out of the frames throughout. The performance of the proposed method for background subtraction of these video data is shown in figure 2.

Figure 2(four images on the left) are results on a randomly selected frame from the Lobby video. The four images correspond to the true frame, low rank approximation (after adding back the mean) and the background subtracted image using the eigenvectors obtained from the proposed Krylov method ($m = 10, \sigma = 0.1$), and using the exact eigenvectors, respectively. The images were all mean centered and normalized to have unit norm. The approximate dimension estimated was equal to 1. The matrix has one very large eigenvalue compared to rest, since the video has very little activities (one/two people moving in and out in few frames).

Figure 2(C) and (D) are the background subtracted images for a randomly selected frame from the Shopping Mall video. The approximate dimension estimated by our method was 14. This video has more activities and the dimension estimated here is higher than for the Lobby data. For more details on these datasets and the use of PCA for foreground detection, we refer [6]. We observe that, we can achieve good foreground detection using the proposed method. Also note that, our method does not require forming the covariance matrix for PCA (in the above two video datasets, $p = 20480$ and 81920, respectively), hence requiring less storage (such dense covariance matrices would not fit in the memory). Therefore, this example illustrates how the proposed method can be used to simultaneously estimate the dimension of the principal subspace and use the approximation obtained for the principal subspace in PCA and robust PCA applications.

References

[1] T. W. Anderson, *Asymptotic theory for principal component analysis*, The Annals of Mathematical Statistics, 34 (1963), pp. 122–148.

[2] ———, *An introduction to multivariate statistical analysis*, Wiley-Interscience, 2003.

[3] S. Bradde and W. Bialek, *PCA Meets RG*, Journal of Statistical Physics, 167 (2017), pp. 462–475.

[4] E. Bura and R. D. Cook, *Rank estimation in reduced-rank regression*, Journal of Multivariate Analysis, 87 (2003), pp. 159–176.

[5] G. Camba-Méndez and G. Kapetanios, *Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling*, (2008).

[6] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Journal of the ACM (JACM), 58 (2011), p. 11.

[7] P. Comon and G. H. Golub, *Tracking a few extreme singular values and vectors in signal processing*, Proceedings of the IEEE, 78 (1990), pp. 1327–1343.

[8] T. A. Davis and Y. Hu, *The university of florida sparse matrix collection*, ACM Transactions on Mathematical Software (TOMS), 38 (2011), p. 1.

[9] S. G. Donald, N. Fortuna, and V. Pipiras, *On rank estimation in symmetric matrices: the case of indefinite matrix estimators*, Econometric Theory, 23 (2007), pp. 1217–1232.

[10] N. Halko, P. Martinsson, and J. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review, 53 (2011), pp. 217–288.

[11] G. E. Hinton and R. R. Salakhutdinov, *Reducing the dimensionality of data with neural networks*, science, 313 (2006), pp. 504–507.

[12] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge university press, 1990.

[13] T. Idé and K. Tsuda, *Change-point detection using krylov subspace learning*, in Proceedings of the 2007 SIAM International Conference on Data Mining, SIAM, 2007, pp. 515–520.

[14] K. Johansson, *Shape fluctuations and random matrices*, Communications in mathematical physics, 209 (2000), pp. 437–476.

[15] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Annals of statistics, (2001), pp. 295–327.

[16] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[17] R. Khanna, J. Ghosh, R. Poldrack, and O. Koyejo, *A deflation method for structured probabilistic PCA*, in Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, 2017, pp. 534–542.

[18] S. Kritchman and B. Nadler, *Determining the number of components in a factor model from limited noisy data*, Chemometrics and Intelligent Laboratory Systems, 94 (2008), pp. 19–32.

[19] ———, *Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory*, IEEE Transactions on Signal Processing, 57 (2009), pp. 3930–3941.

[20] M. Meloun, J. Capek, P. Mikik, and R. G. Brereton, *Critical comparison of methods predicting the number of components in spectroscopic data*, Analytica Chimica Acta, 423 (2000), pp. 51–68.

[21] R. J. Muirhead, *Aspects of multivariate statistical theory*, vol. 197, John Wiley & Sons, 2009.

[22] C. Musco and C. Musco, *Randomized block krylov methods for stronger and faster approximate singular value decomposition*, in Advances in Neural Information Processing Systems, 2015, pp. 1396–1404.

[23] B. Nadler, *Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator*, IEEE Transactions on Signal Processing, 58 (2010), pp. 2746–2756.

[24] N. Patterson, A. L. Price, and D. Reich, *Population structure and eigenanalysis*, PLoS genetics, 2 (2006), p. e190.

[25] S. Rachakonda, R. F. Silva, J. Liu, and V. D. Calhoun, *Memory efficient pca methods for large group ica*, Frontiers in neuroscience, 10 (2016), p. 17.

[26] J.-M. Robin and R. J. Smith, *Tests of rank*, Econometric Theory, 16 (2000), pp. 151–175.

[27] Y. Saad, *Numerical Methods for Large Eigenvalue Problems- classics edition*, SIAM, Philadelpha, PA, 2011.

[28] M. K. Schneider and A. S. Willsky, *Krylov subspace estimation*, SIAM Journal on Scientific Computing, 22 (2001), pp. 1840–1864.

[29] M. E. Tipping and C. M. Bishop, *Probabilistic principal component analysis*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61 (1999), pp. 611–622.

[30] S. Ubaru, A. Mazumdar, and Y. Saad, *Low rank approximation and decomposition of large matrices using error correcting codes*, IEEE Transactions on Information Theory, 63 (2017), pp. 5544–5558.

[31] S. Ubaru and Y. Saad, *Fast methods for estimating the numerical rank of large matrices*, in Proceedings of The 33rd International Conference on Machine Learning, 2016, pp. 468–477.

[32] S. Ubaru, Y. Saad, and A.-K. Seghouane, *Fast estimation of approximate matrix ranks using spectral densities*, Neural Computation, 29 (2017), pp. 1317–1351.

[33] M. Wax and T. Kailath, *Detection of signals by information theoretic criteria*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 33 (1985), pp. 387–392.

[34] G. Xu and T. Kailath, *Fast subspace decomposition*, IEEE Transactions on Signal Processing, 42 (1994), pp. 539–551.

A Proofs for the derivation

Here we give the proofs missing in the main paper.

Proof of Proposition 3.2.

Proof. From proposition 3.1, $\mathbf{S}_n$ is a $\sqrt{n}$ consistent estimator of $\mathbf{\Sigma}$, and we can express $\mathbf{S}_n$ as a perturbation

$$\mathbf{S}_n = \mathbf{\Sigma} + \varepsilon \frac{\mathbf{S}_n - \mathbf{\Sigma}}{\varepsilon} = \mathbf{\Sigma} + \varepsilon \mathbf{E},$$

where the perturbation of $\mathbf{\Sigma}$ is of the order $1/\sqrt{n}$. That is, $\varepsilon\mathbf{E} = O_p\left(\frac{1}{\sqrt{n}}\right)$. Then,

$$\begin{aligned} \mathbf{S}_n \mathbf{U}_q \Lambda_q^{-1} &= (\mathbf{\Sigma} + \varepsilon\mathbf{E})\mathbf{U}_q\Lambda_q^{-1} \\ &= \mathbf{U}_q + \varepsilon\mathbf{E}\mathbf{U}_q\Lambda_q^{-1}. \end{aligned}$$

Since $\mathbf{U}_q$ has orthogonal columns and is non-random, and also for Λ_q^{-1} (diagonal matrix with inverse of the top q eigenvalues) is bounded since $\lambda_q > \varepsilon$, the second term in the above equation should be $\varepsilon\mathbf{E}\mathbf{U}_q\Lambda_q^{-1} = O_p(\varepsilon)$. Then, we have $\mathbf{G}_q = \mathbf{U}_q + O_p\left(\frac{1}{\sqrt{n}}\right)$, i.e., $\mathbf{G}_q$ is a $\sqrt{n}$ consistent estimator of $\mathbf{U}_q$. See [1] for further details.

Proof of the corresponding Corollary:

Proof. From proposition 3.2, we have $\mathbf{G}_q = \mathbf{U}_q + O_p\left(\frac{1}{\sqrt{n}}\right)$. Then,

$$\begin{aligned} \mathbf{Q}_G = \mathbf{G}_{p-q}\mathbf{G}_{p-q}^T &= \mathbf{I}_p - \mathbf{G}_q\mathbf{G}_q^T \\ &= \mathbf{I}_p - \left[\mathbf{U}_q + O_p\left(\frac{1}{\sqrt{n}}\right)\right]\left[\mathbf{U}_q + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^T \\ &= \mathbf{I}_p - \mathbf{U}_q\mathbf{U}_q^T + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \mathbf{U}_{p-q}\mathbf{U}_{p-q}^T + O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

Proof of Proposition 3.3.

Proof. Using the Corollary, we have

$$\begin{aligned} \mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G) &= \mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)(\mathbf{Q}_G - \mathbf{Q}_U)) + \mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U) \\ &= \mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \mathtt{vec}((\mathbf{Q}_G - \mathbf{Q}_U)(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U) + \mathtt{vec}(\mathbf{Q}_U(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \mathtt{vec}(\mathbf{Q}_U(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U) + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Thus, $\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)$ has the same asymptotic distribution as $\mathtt{vec}(\mathbf{Q}_U(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_U)$. We know that the bottom $p-q$ eigenvalues of $\mathbf{\Sigma}$ are all σ. Hence we have $\mathbf{Q}_U\mathbf{\Sigma}\mathbf{Q}_U = \mathbf{Q}_U(\sigma\mathbf{I}_p)\mathbf{Q}_U$. So, we have

$$\begin{aligned} \mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G) &= \mathtt{vec}(\mathbf{Q}_U(\mathbf{S}_n - \mathbf{\Sigma})\mathbf{Q}_U) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= (\mathbf{Q}_U \otimes \mathbf{Q}_U)\mathtt{vec}(\mathbf{S}_n - \mathbf{\Sigma}) + O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Thus, in terms of the distribution, we have from above,

$$\sqrt{n}\mathbb{E}\{\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)\} = (\mathbf{Q}_U \otimes \mathbf{Q}_U)\mathbb{E}\{\sqrt{n}\mathtt{vec}(\mathbf{S}_n - \mathbf{\Sigma})\} = 0$$

and

$$\begin{aligned} cov\{\sqrt{n}\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)\} &= (\mathbf{Q}_U \otimes \mathbf{Q}_U)cov\{\sqrt{n}\mathtt{vec}(\mathbf{S}_n - \mathbf{\Sigma})\}(\mathbf{Q}_U \otimes \mathbf{Q}_U) \\ &= (\mathbf{Q}_U \otimes \mathbf{Q}_U)\mathbf{\Omega}(\mathbf{Q}_U \otimes \mathbf{Q}_U). \end{aligned}$$

Proof of Theorem 3.1.

Proof. A model selection criterion takes the form

$$IC(k) = L(n,k) - \mathbb{E}(L(n,k)),$$

as $n \longrightarrow \infty$. In our case, from Lemma 3.1, we have

$$L(n,k) = \sum_{i=1}^{\eta} \mu_i \chi^2_{(1)},$$

where μ_i are the eigenvalues of $\frac{1}{2\sigma^2}(\mathbf{Q}_G \otimes \mathbf{Q}_G)(\mathbf{S}_n \otimes \mathbf{S}_n)(\mathbf{Q}_G \otimes \mathbf{Q}_G)$, an estimate of $\frac{1}{2\sigma^2}\hat{\mathbf{\Omega}}$ from Proposition 3.3, the asymptotic covariance matrix of $\sqrt{\frac{n}{2\sigma^2}}\mathtt{vec}(\mathbf{Q}_G(\mathbf{S}_n - \sigma\mathbf{I}_p)\mathbf{Q}_G)$. To compute an approximation to the mean of the statistic, we use the following Gamma approximation:

$$\begin{aligned}\sum_{i=1}^{\eta} \mu_i \chi^2_{(1)} &= \sum_{i=1}^{\eta} \mu_i \Gamma\left(\frac{1}{2}, 2\right) = \sum_{i=1}^{\eta} \Gamma\left(\frac{1}{2}, 2\mu_i\right) \\ &= \sum_{i=1}^{\eta} \Gamma(\kappa, \theta_i) \simeq \Gamma(K, \Theta)\end{aligned}$$

where

$$\kappa = \frac{1}{2}, \theta_i = 2\mu_i, K = \frac{(\sum_i \kappa\theta_i)^2}{\sum_i \theta_i^2 \kappa} \text{ and } \Theta = \frac{\sum_i \kappa\theta_i}{K}$$

and the mean of the asymptotic approximation of L is given by $\mathbb{E}(L(n,k)) = K\Theta$. Hence, in our case,

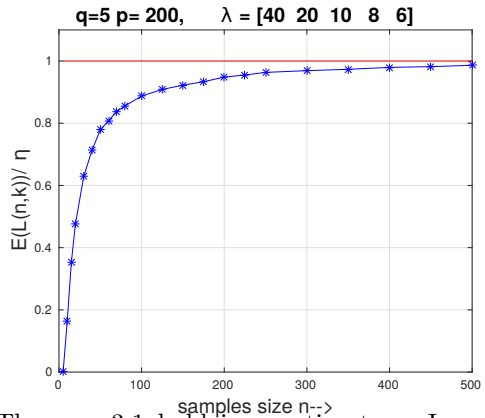
$$\mathbb{E}(L(n,k)) = \sum_{i=1}^{\eta} \kappa\theta_i = \sum_{i=1}^{\eta} \mu_i = \sum_{i,j=k+1; i\neq j}^{p} \frac{\ell_i * \ell_j}{2\sigma^2},$$

where $\{\ell_i\}_{i=1}^p$ are the eigenvalues of the sample covariance matrix $\mathbf{S}_n$ and the last equality is from the property of Kronecker products as seen in the proof of Lemma 3.1.

Note that, asymptotically $\ell_i \to \sigma$,the noise variance, for $i > q$ as $n \to \infty$. Hence, asymptotically

$$\mathbb{E}(L(n,k)) \to \eta = \frac{(p-k)(p-k-1)}{2}.$$

Hence, we use the criterion in (3.2) for model selection, i.e., for the dimension estimation of the principal subspace.



The figure on the left plots the ratio

$$\frac{\mathbb{E}(L(n,q))}{\eta} = \frac{\sum_{i,j=k+1; i\neq j}^{p} \frac{\ell_i * \ell_j}{2\sigma^2}}{\frac{(p-q)(p-q-1)}{2}}$$

as a function of the number of samples n for a small simulation with $p = 200, q = 5$ (similar to the experiment in Figure 1). The true covariance matrix from which the data is sampled has top $q = 5$ eigenvalues of magnitude listed in the figure and the noise level was $\sigma = 1.2$. We plot the average of the ratio over 30 trials. We note that the mean $\mathbb{E}(L(n,q))$ quickly approaches the degree of freedom η, showing that the quantity $L(n,q)$ indeed has χ^2_η distribution for large enough n. Thus, Lemma 3.1 and Theorem 3.1 hold in practice too. In section 5 of the main paper and below, we present several numerical experiments to illustrate the performance of the proposed method.

B Proofs for the analysis

Proof of Theorem 4.1

Proof. In order to prove the strong consistency of

$$\hat{k} = \arg\min_k IC(k),$$

we first consider that $\hat{k} > k_0$, then

$$\begin{aligned}
IC(\hat{k}) - IC(k_0) &= \frac{n}{2\sigma^2}\left(\sum_{i=\hat{k}+1}^{p}(\ell_i-\sigma)^2 - \sum_{i=k_0+1}^{p}(\ell_i-\sigma)^2\right) - C_n\left(\frac{(p-\hat{k})(p-\hat{k}-1)}{2} - \frac{(p-k_0)(p-k_0-1)}{2}\right) \\
&= -\frac{n}{2\sigma^2}\sum_{i=k_0+1}^{\hat{k}}(\ell_i-\sigma)^2 - C_n\left(\frac{(\hat{k}-k_0)(\hat{k}+k_0-2p+1)}{2}\right) \\
\frac{IC(\hat{k}) - IC(k_0)}{n} &= -\frac{1}{2\sigma^2}\sum_{i=k_0+1}^{\hat{k}}(\lambda_i-\sigma)^2 - \frac{C_n}{n}\left(\frac{(\hat{k}-k_0)(\hat{k}+k_0-2p+1)}{2}\right) + O\left(\sqrt{\frac{\log\log n}{n}}\right),
\end{aligned}$$

since $\ell_i = \lambda_i + O\left(\sqrt{\frac{\log\log n}{n}}\right)$ from the law of iterated logarithm [21]. The last two terms in the RHS of the above equation go to zero as n tends to infinity and $\lambda_i > 0$, hence we have

$$IC(\hat{k}) - IC(k_0) < 0 \text{ for all large } n \text{ a.s.}$$

Next, for $\hat{k} < k_1$, we have

$$\begin{aligned}
IC(\hat{k}) - IC(k_1) &= \frac{n}{2\sigma^2}\left(\sum_{i=\hat{k}+1}^{p}(\ell_i-\sigma)^2 - \sum_{i=k_1+1}^{p}(\ell_i-\sigma)^2\right) - C_n\left(\frac{(p-\hat{k})(p-\hat{k}-1)}{2} - \frac{(p-k_1)(p-k_1-1)}{2}\right) \\
&= \frac{n}{2\sigma^2}(k_1-\hat{k})O\left(\frac{\log\log n}{n}\right) - C_n\left(\frac{(\hat{k}-k_1)(\hat{k}+k_1-2p+1)}{2}\right) \\
\frac{IC(\hat{k}) - IC(k_1)}{C_n} &= -\frac{(\hat{k}-k_1)(\hat{k}+k_1-2p+1)}{2} + \frac{(k_1-\hat{k})}{2\sigma^2}.\frac{O(\log\log n)}{C_n}.
\end{aligned}$$

Since, $\ell_i = \lambda_i + O\left(\sqrt{\frac{\log\log n}{n}}\right)$ and for all $i > \hat{k}, \lambda_i = \sigma$. Again, the second term in the RHS of the above equation goes to zero due the the property of C_n. As, $\hat{k} < k_1$ and $\{\hat{k}, k_1\} < p$, the first term is always negative. Hence, we again have

$$IC(\hat{k}) - IC(k_1) < 0 \text{ for all large } n \text{ a.s.}$$

Proof of Corollary 4.1

Proof. For the eigenvalues θ_i computed in Algorithm 1, we have from Lemma 3.2,

$$\ell_i - \epsilon\ell_{k+1} \le \theta_i \le \ell_i,\ i = 1, \ldots, k.$$

Hence, we have

$$IC(k) \le \frac{n}{2\sigma^2}\left(\|\mathbf{S}_n - \sigma\mathbf{I}_p\|_F^2 - \sum_{i=1}^{k}(\ell_i - \epsilon\ell_{k+1} - \sigma)^2\right) - C_n\frac{(p-k)(p-k-1)}{2}.$$

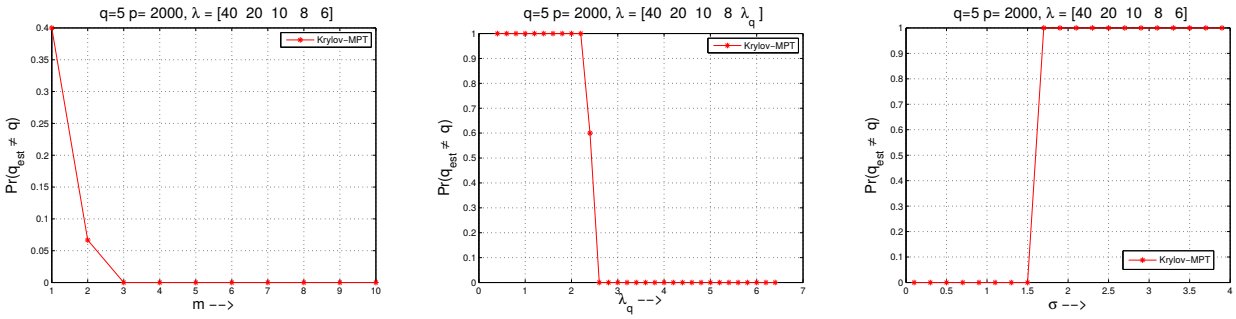


Figure 3: Signal detection using the Krylov method: Detection as a function of: number of Lanczos steps m (left), signal strength (ℓ_q eigenvalue), and (right) the noise level σ.

For the first case when $\hat{k} > k_0$, ignoring the terms that go to zero asymptotically, we will have:

$$\begin{aligned}\frac{IC(\hat{k}) - IC(k_0)}{n} &\leq \frac{1}{2\sigma^2}\left(\sum_{i=1}^{k_0}(\ell_i - \epsilon\ell_{k_0+1} - \sigma)^2 - \sum_{i=1}^{\hat{k}}(\ell_i - \epsilon\ell_{\hat{k}+1} - \sigma)^2\right) \\ &= \frac{1}{2\sigma^2}\left(\sum_{i=1}^{k_0}\left((\ell_i - \epsilon\ell_{k_0+1} - \sigma)^2 - (\ell_i - \epsilon\ell_{\hat{k}+1} - \sigma)^2\right) - \sum_{i=k_0+1}^{\hat{k}}(\ell_i - \epsilon\ell_{\hat{k}+1} - \sigma)^2\right).\end{aligned}$$

For $\epsilon < 1$, note that both terms in RHS is always negative since $\ell_{k_0+1} > \ell_{\hat{k}+1}$. Hence $IC(\hat{k}) - IC(k_0) < 0$ for eigenvalues computed by the Krylov method.

Next, for the case $\hat{k} < k_1$, the term in $\frac{IC(\hat{k}) - IC(k_1)}{C_n}$ which is neither negative nor goes to zero is

$$\begin{aligned}\frac{IC(\hat{k}) - IC(k_1)}{C_n} &\leq \frac{n}{2\sigma^2 C_n}\left(\sum_{i=\hat{k}+1}^{k_1}(\ell_i - \epsilon\ell_{k_1+1} - \sigma)^2\right) \\ &= \frac{n}{2\sigma^2 C_n}(k_1 - \hat{k})\left((1-\epsilon)(\sigma + O\left(\sqrt{\frac{\log\log n}{n}}\right)) - \sigma\right)^2.\end{aligned}$$

Hence, if we replace σ in the algorithm by $(1-\epsilon)\sigma$, this term goes to zero and we will have $IC(\hat{k}) - IC(k_1) < 0$.

C Additional Numerical Results

In section 5 of the main paper, we presented several numerical experiments to illustrate the performance of the proposed method in applications. Here, we present few additional experimental results.

Krylov subspace method: In the the main paper, for the number of signal detection experiments, we used the exact eigenvalues of the covariance matrices (computed using `eig` function in Matlab) for the dimension estimation using the three compared methods (MDL and RMT require all of the eigenvalues). Here, we illustrate how the proposed Krylov subspace based algorithm 1 performs for the dimension estimation. We consider the same signal detection problem as above (same Gaussian model as Fig. 1). The first plot in figure 3 give the performance of the algorithm as a function of the number of Lanczos steps m. The parameters were chosen to be $p = 2000, n = 2500, \sigma = 1.1$. We know the relation between the error ϵ in the eigenvalue estimation by the Lanczos algorithm and the number of Lanczos steps m from Lemma 3.2. Hence, increasing m is equivalent to decreasing ϵ. We see that for a very few Lanczos steps $m \geq 4$, we get accurate results. This is because, it is well-known that the top eigenvalues computed by Lanczos algorithm converges fast [27]. This superior performance of the Lanczos algorithm was observed in [34] as well for a similar Gaussian signal detection model.

In the second and third plots, we plot the performance of the Krylov subspace method for signal detection as a function of the signal strength (magnitude of λ_q in the middle) and the noise level σ (right), with $p = 2000, n = 2500, m = 5$. We observed that, our Algorithm 1, for $m \geq 4$, performs very well and replicates the

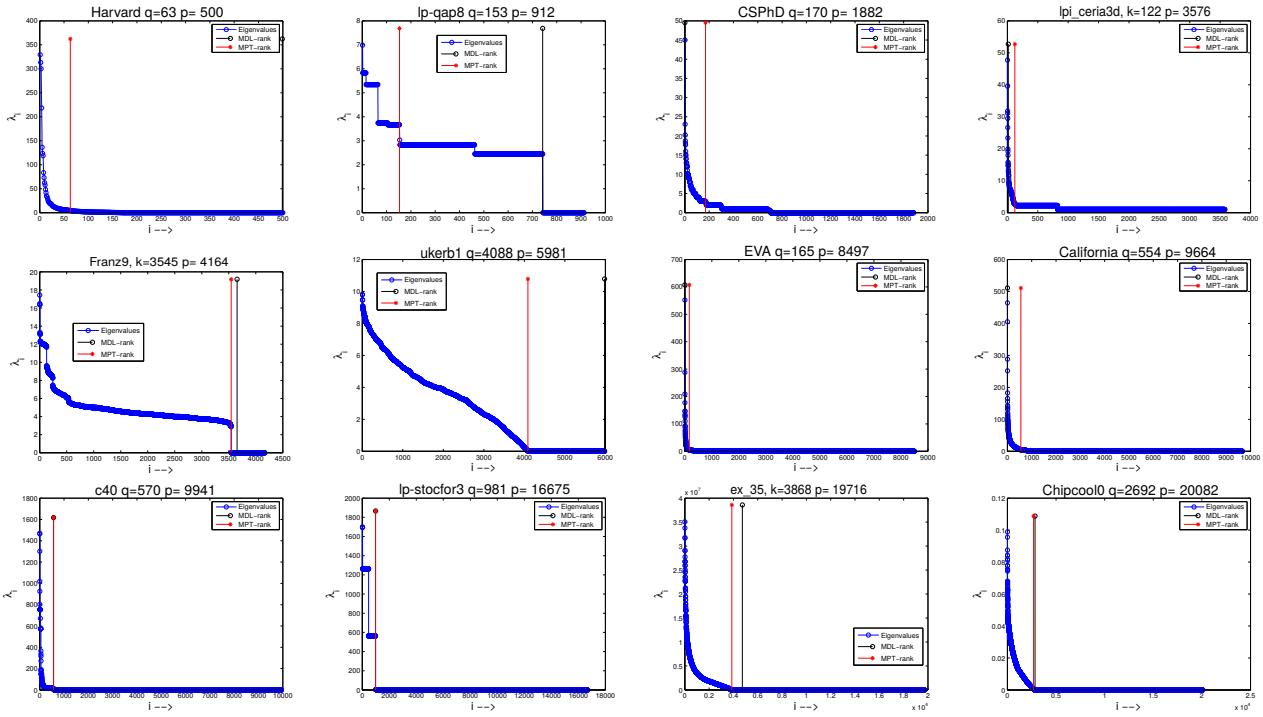


Figure 4: Numerical rank estimation of data matrices by the proposed method (MPT) and MDL, along with the actual spectrum.

results we obtained by the proposed method with exact eigenvalues of the sample covariance matrix (reported in Figure 1).

Data Matrices: In Table 1 of the main paper, we saw the performance of the proposed algorithm on few sparse data matrices. The following results give us more insight into the method's performance. Figure 4 presents the spectrum of twelve matrices obtained from the SuiteSparse database with low numerical rank and gap in the spectrum, along with the rank estimated by the the proposed method (MPT) as a red (star) line and MDL in black (circle). We chose $C_n = \log n$ in all cases and $\sigma = 1$ (except chipcool0 where $\sigma = 0.01$ was chosen). The matrix name, size p and the actual numerical rank q (based on the gap) are given in the title of each plot. We note that the proposed method gives god solution for almost all examples except one case (lp-qap8, second plot, the method chooses a different gap in the spectrum for $\sigma = 1$). The MDL method fails in a few examples and is slightly off in a couple more examples. The matrix lpiceria3d (fourth plot/1st row 1st column) is interesting because the matrix has two distinct eigen-gaps close to zero. Our method selects the first one. These set of experiments show that the proposed method performs very well (determines the rank based on the spectral gap) for general data matrices too, where the distribution assumptions do not hold.