Submitted to the Annals of Applied Statistics

2SDR: APPLYING KRONECKER ENVELOPE PCA TO DENOISE CRYO-EM IMAGES

BY SZU-CHI CHUNG, PO-YAO NIU, SU-YUN HUANG,
WEI-HAU CHANG AND I-PING TU

Institute of Statistical Science, Academia Sinica, Taiwan

November 25, 2019

Principal component analysis (PCA) is arguably the most widely used dimension reduction method for vector type data. When applied to image data, PCA demands the images to be portrayed as vectors. The resulting computation is heavy because it will solve an eigenvalue problem of a huge covariance matrix due to the vectorization step. To mitigate the computation burden, multilinear PCA (MPCA) that generates each basis vector using a column vector and a row vector with a Kronecker product was introduced, for which the success was demonstrated on face image sets [1, 2]. However, when we apply MPCA on the cryo-electron microscopy (cryo-EM) particle images, the results are not satisfactory when compared with PCA. On the other hand, to compare the reduced spaces as well as the number of parameters of MPCA and PCA, Kronecker Envelope PCA (KEPCA) [1, 3, 4] was proposed to provide a PCA-like basis from MPCA. Here, we apply KEPCA to denoise cryo-EM images through a two-stage dimension reduction (2SDR) algorithm. 2SDR first applies MPCA to extract the projection scores and then applies PCA on these scores to further reduce the dimension. 2SDR has two benefits that it inherits the computation advantage of MPCA and its projection scores are uncorrelated as those of PCA. Testing with three cryo-EM benchmark experimental datasets shows that 2SDR performs better than MPCA and PCA alone in terms of the computation efficiency and denoising quality. Remarkably, the denoised particles boxed out from the 2SDR-denoised micrographs allow subsequent structural analysis to reach a high-quality 3D density map. This demonstrates that the high resolution information can be well preserved through this 2SDR denoising strategy.

1. Introduction. Principal component analysis (PCA) is arguably the most popular dimension reduction method for vector type data. There exist many reasons for PCA to gain its popularity. First, its statistical concept is very intuitive that it searches new variables, through orthogonal linear transformation, to capture the most variation of the data. Second, PCA is very friendly to users because most statistical or mathematical packages offer PCA with a simple interface. Third, it has broad applications as long as the data can be presented as a matrix (variable vs sample) form. Many modern data analysis tools that target on large datasets adopt PCA as a built-in device to reduce the dimension of data and thus facilitate the processing, where t-SNE [5] is a typical example.

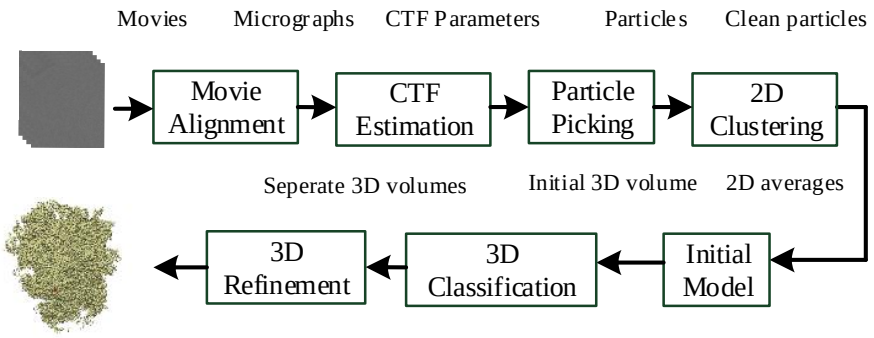


Fig 1: Single particle analysis workflow. **Data Collection** The particle projections are recorded in movie files of which each contains tens of movie frames. **Movie Alignment** This step is to correct the particle motions across frames induced by the electron beams to improve the average performance that makes particles on the micrographs better visualized. **CTF Estimation** CTF estimation is to restore the signal which has undergone a non-uniform transmission over spatial frequencies when passing through the electron microscopy optics. **Particle Picking** Particle picking from the micrographs is to extract the particle images which are the 2D projections of the molecules. These 2D projections contain various views and free orientations of the molecules. **2D Clustering** This step involves aligning particle images through in-plane rotation and shift, as well as grouping images with similar views. 2D clustering has two major goals. First, non-particles or low quality images should be screened out by removing under-performed clusters to create a cleaner particle set. Second, the resulting 2D class average set can be used to generate a 3D initial model. **Initial Model** Based on the high quality 2D class averages, an initial 3D model can be constructed through packages like Prime [6] or Eman2 [7]. **3D Classification** When considering the heterogeneity of the underlying dataset, multiple initial 3D models will be generated through perturbation and the 3D Classification will be performed on these initial 3D models. **3D Refinement** The Euler angle and particle center for each particle image are important parameters to compose the 3D density map and they can be iteratively refined. Each particle image will search its best parameters that conform to the most updated 3D solution and all the new parameters can update a 3D density map.

When applied on an image set, PCA requires the images to be presented as vectors. In this paper, we use ivPCA (image vectorization PCA) to specify this approach. ivPCA would face huge computation complexity because it involves solving an eigenvalue problem of a covariance matrix whose size is the square of the pixel number. Multilinear PCA (MPCA) [1, 2] was thus developed to avoid image vectorization by solving eigenvectors in column and row spaces and applying a Kronecker product to compose the matrix structure. In addition, MPCA has shown better performance than ivPCA on reconstructing a testing dataset of face images [1]. However, when we tested on an image set from single particle cryo-electron microscopy (cryo-EM) experiment which is heavily contaminated by noise, we observed that, compared to ivPCA (if ivPCA's computation barrier was overcome), MPCA was not able to recover a reasonable amount of particle information from the noise even with an exhaustive selection of the ranks.

Compared to ivPCA, MPCA usually adopts a larger number of bases to construct its projection space due to its Kronecker structure. To discuss the efficiency of parameter usage, Kronecker Envelope PCA (KEPCA)[1, 3, 4] was presented to formulate more PCA-like bases from MPCA. We envision that KEPCA has the capability in handling matrix structure data, such as cryo-EM images. In this work, we realize the usage of KEPCA in denoising cryo-EM images through a two-stage dimension reduction (2SDR) algorithm. 2SDR first applies MPCA to extract the projection scores from the image set, and then applies PCA on these scores to further reduce the dimension. In this way, 2SDR inherits the computation advantage of MPCA while its projected dimensionality can be as succinct as that of ivPCA. In the rest of this paper, we use 2SDR to refer the application of KEPCA on images. Since this application is motivated by analyzing single particle cryo-EM image sets that we consider it is beneficial to arm the readers with a brief introduction to single particle cryo-EM.

Single particle cryo-EM analysis has become the mainstream method for solving high-resolution 3D structure density maps of biomolecules. Yet, while targeting on solving the high resolution structure, cryo-EM images are extremely noisy because very low dose of electron is used to mitigate radiation damages[1]. As a result, a typical cryo-EM experiment tends to collect a large number of particle images, usually more than hundreds of thousands, in hope of compensating the noise contamination by averaging. Owing to the current advances of electron microscope detector and automation, the size of a cryo-EM particle image is often larger than a hundred in each direction such that the dimension it faces with is extremely high. The data characters of cryo-EM images, including strong noise contamination, huge dimension and large sample size, make its processing very challenging. We summarize the cryo-EM processing workflow as a flowchart in Figure 1 [8].

[1]The signal to noise ratio is often smaller than 0.1

In the present study, we compare the 2D images reconstructed by 2SDR, ivPCA and MPCA on three cryo-EM benchmark datasets, including two particle image sets and one micrograph set. These datasets represent 3 typical computation loadings. The first is the 70s Ribosome particle image set containing 5000 images with size 130×130 whose computation cost is affordable to a modern notebook for all the three methods and thus its denoised image quality will provide a check point. We show that ivPCA and 2SDR perform much better than MPCA in this case. The second is the 80s Ribosome particle image set containing 105,247 images with size 360×360. Since ivPCA involves intensive computations for this dataset, it fails the task on the regular computers. Yet, 2SDR and MPCA succeed, by which their computation advantages are demonstrated. Importantly, 2SDR has much better performance in image reconstruction than MPCA for the second dataset. The third is the Beta-galactosidase micrograph set containing 15 movie files and each movie contains 16 frames with size 1950×1950 which is a very challenging dimension reduction task due to its extremely high noise, very large dimension. For this challenging dataset, the images reconstructed by 2SDR and MPCA can both clearly reveal particles in the micrographs after applying a standard contrast enhancement method while ivPCA can not. Our simulations also confirm that using minimum Mean Square Error as the criterion, 2SDR can best reconstruct the signal information when data carry heavy noise, which is the case of cryo-EM images.

2. Methods.

2.1. *ivPCA.* Let $X_1, \cdots, X_n$ be n copies of image matrix with size $p \times q$ and let $y_i = \mathrm{vec}(X_i)$, where vec is the operator of matrix vectorization by stacking the matrix into a vector by columns. The statistical model for ivPCA is

$$y = \mu + \Gamma \nu + \varepsilon \quad \in \mathbb{R}^m,$$

where μ is the grand mean, $\nu \in \mathbb{R}^r$ with $r \leq m = pq$ is a random vector, $\Gamma \in \mathscr{O}_{m,r}$ where $\mathscr{O}_{m,r} := \{M \in \mathbb{R}^{m \times r} | M^\top M = I_r\}$, and ε is an error vector modeled to be independent of ν with zero mean and $\mathrm{Cov}(\varepsilon) = \sigma^2 I_m$. The random vector ν has covariance matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_r)$ with descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$.

The estimation $\widehat{\Gamma}$ contains the first r eigenvectors of the sample covariance matrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^\top$, and the images can be reconstructed as

$$\mathrm{vec}(\widehat{X}_i) = \mathrm{vec}(\overline{X}) + \widehat{\Gamma} \cdot \widehat{\nu}_i = \mathrm{vec}(\overline{X}) + P_{\widehat{\Gamma}} \mathrm{vec}(X_i - \overline{X}), \tag{1}$$

where $P_{\widehat{\Gamma}} = \widehat{\Gamma}\widehat{\Gamma}^\top$ also called the projection matrix of $\widehat{\Gamma}$ and $\overline{X}$ is the estimated mean.

ivPCA was introduced in the early development of single particle cryo-EM analysis to reduce the dimension of the 2D projection images to facilitate 2D image

clustering [9, 10]. As the sample size and dimension size both grow rapidly, the computing cost of ivPCA can become a serious burden because image vectorization will lead to a huge covariance matrix. For example, the high resolution project of 80s Ribosome dataset [11] has more than 100,000 particle images with pixel size $p \times q$ where $p = q = 360$ which creates a covariance matrix Σ_n with size $m \times m = 360^2 \times 360^2$ with $m = pq$ for ivPCA. Notice that the computation complexity of solving the first r eigenvectors of Σ_n is $O(m^2 r)$[12]. As a result, solving the first r eigenvector set of ivPCA will encounter the computation complexity of $O(10^{10} \times r)$ with a rough approximation of 360^2 by 10^5. This computational demand becomes a problem for those users who do not have very rich computing resources.

2.2. *MPCA.* An alternative dimension reduction approach that obviates matrix vectorization has been proposed, including the high order singular value decomposition (HOSVD) [13] and MPCA [1, 2]. Both HOSVD and MPCA have been reported to effectively decrease the computation cost and yet reconstruct the image data reasonably well [1]. Yet, if the rank is correctly specified, MPCA is shown to have better asymptotic property than HOSVD [1], thus we only consider MPCA in this paper.

MPCA models the matrix as:

$$\begin{aligned} X &= Z + \mathscr{E} \quad \in \mathbb{R}^{p \times q}, \\ Z &= M + AUB^\top, \end{aligned} \tag{2}$$

where $M \in \mathbb{R}^{p \times q}$ is the grand mean, $U \in \mathbb{R}^{p_0 \times q_0}$ is a random matrix $p_0 \le p$, $q_0 \le q$; $A \in \mathscr{O}_{p,p_0}$, $B \in \mathscr{O}_{q,q_0}$ are orthogonal non-random matrices, $\mathscr{E}$ is the error term which is independent of U and has zero mean and $\mathrm{Cov}(\mathrm{vec}(\mathscr{E})) = \sigma^2 I_m$, $m = pq$.

Ye [2] proposed the generalized low rank approximations of matrices (GLRAM) to estimate A and B. Given (p_0, q_0), $\widehat{A}$ consists of the leading p_0 eigenvectors of the covariance matrix $\sum_{i=1}^n (X_i - \overline{X}) P_{\widehat{B}} (X_i - \overline{X})^\top$, and $\widehat{B}$ consists of the leading q_0 eigenvectors of $\sum_{i=1}^n (X_i - \overline{X})^\top P_{\widehat{A}} (X_i - \overline{X})$, where $P_{\widehat{A}} = \widehat{A}\widehat{A}^\top$ and $P_{\widehat{B}} = \widehat{B}\widehat{B}^\top$. The solution is retained by convergence through iterations, usually less than 10 steps.

The $\widehat{U}_i$ of the above two methods are computed as $\widehat{U}_i = \widehat{A}^\top (X_i - \overline{X}) \widehat{B}$, for $i = 1, \ldots, n$. The images reconstructed by MPCA's bases are as

$$\begin{aligned} &\widehat{X}_i = \overline{X} + \widehat{A}\widehat{U}_i\widehat{B}^\top = \overline{X} + P_{\widehat{A}}(X_i - \overline{X})P_{\widehat{B}} \text{ or} \\ &\mathrm{vec}(\widehat{X}_i) = \mathrm{vec}(\overline{X}) + (\widehat{B} \otimes \widehat{A}) \cdot \mathrm{vec}(\widehat{U}_i) = \mathrm{vec}(\overline{X}) + P_{\widehat{B}\otimes\widehat{A}} \cdot \mathrm{vec}(X_i - \overline{X}), \end{aligned} \tag{3}$$

where the projection matrix $P_{\widehat{B}\otimes\widehat{A}} = (\widehat{B} \otimes \widehat{A})(\widehat{B} \otimes \widehat{A})^\top = (\widehat{B}\widehat{B}^\top) \otimes (\widehat{A}\widehat{A}^\top)$, and $\otimes$ denotes the Kronecker product. (3) implies that $\mathrm{vec}(\widehat{X}_i - \overline{X})$ is reconstructed by the bases $(\widehat{B} \otimes \widehat{A})$ with the scores $\mathrm{vec}(U_i) = (\widehat{B} \otimes \widehat{A})^\top \mathrm{vec}(X_i - \overline{X})$.

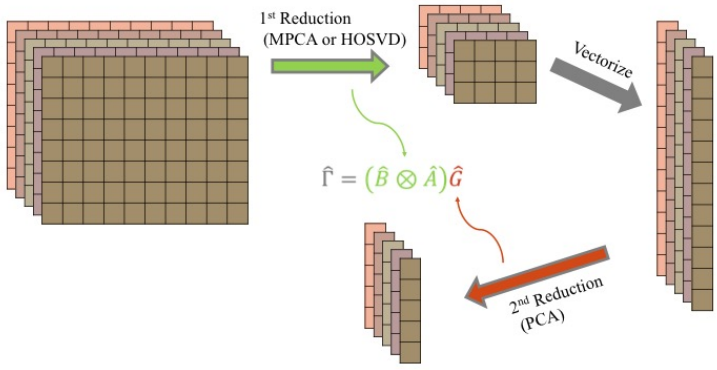


Fig 2: A visual illustration of 2SDR procedure.

Comparing the reconstruction representation of ivPCA and MPCA by their vector forms (1) and (3), we observe that ivPCA employs $\widehat{\Gamma}$ as basis to reconstruct the data and MPCA employs $(\widehat{B} \otimes \widehat{A})$. Note that the pairwise sample correlations among r variables of $\{\widehat{v}_i, 1 \le i \le n\}$ are all zero, i.e., these r variables are uncorrelated, while entries in $\{\widehat{U}_i, 1 \le i \le n\}$ are not. In our experience on many real datasets, the pairwise sample correlations among variables of $\{\widehat{U}_i, 1 \le i \le n\}$ usually are highly correlated. This suggests that $\{\widehat{U}_i, 1 \le i \le n\}$ can be further decorrelated and the dimension be reduced.

2.3. *2SDR.* For 2SDR, matrix data is modeled as

$$\begin{aligned} X &= Z + \mathscr{E} \quad \in \mathbb{R}^{p \times q}, \\ Z &= M + AUB^{\top}, \quad \operatorname{vec}(U) = G\nu, \end{aligned} \tag{4}$$

and its corresponding vector form is

$$\operatorname{vec}(X) = M + (B \otimes A) \cdot G\nu + \operatorname{vec}(\mathscr{E}). \tag{5}$$

Notice that the vector form (5) of 2SDR has been formulated in [1, 3, 4] as the Kronecker envelope PCA model.
We let $\widehat{A}$, $\widehat{B}$ and $\widehat{U}_i$'s be solved through MPCA, and $\widehat{v}_i$'s solved from PCA of the covariance of $\{\operatorname{vec}(\widehat{U}_i), 1 \le i \le n\}$, that is

$$\widehat{U}_i = \widehat{A}^{\top}(X_i - \overline{X})\widehat{B}$$

$$\begin{aligned} \widehat{u}_i &= \mathrm{vec}(\widehat{U}_i) = (\widehat{B}^\top \otimes \widehat{A}^\top)\mathrm{vec}(X_i - \overline{X}) \\ \widehat{v}_i &= \widehat{G}^\top \widehat{u}_i, \end{aligned} \tag{6}$$

where $\widehat{G}$ contains the first r eigenvectors of the covariance matrix $\Sigma_{\widehat{u}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i \widehat{u}_i^\top$. Thus, with the estimate of M by $\overline{X}$, and equation set (6), we have

$$\mathrm{vec}(\widehat{X}_i) = \mathrm{vec}(\overline{X}) + (\widehat{B} \otimes \widehat{A}) \cdot \widehat{G} \cdot \widehat{v} = \mathrm{vec}(\overline{X}) + P_{(\widehat{B}\otimes\widehat{A})\cdot\widehat{G}}\mathrm{vec}(X_i - \overline{X}), \tag{7}$$

where $P_{(\widehat{B}\otimes\widehat{A})\cdot\widehat{G}} = (\widehat{B} \otimes \widehat{A}) \cdot \widehat{G}\widehat{G}^\top (\widehat{B}^\top \otimes \widehat{A}^\top)$. The whole process is summarized in Figure 2.

By comparing equations (1) of ivPCA and (7) of 2SDR, one can see that 2SDR employs $(\widehat{B} \otimes \widehat{A}) \cdot \widehat{G}$ to construct the projection matrix, where $\widehat{G}$ is composed of the eigenvectors of $\Sigma_{\widehat{u}}$, which guarantees the pairwise sample correlations among variables of $\{\widehat{v}_i, 1 \le i \le n\}$ in (6) are all zero. Although 2SDR involves two stages of dimension reduction (MPCA and PCA), the computation cost in solving the eigenvectors is much less than that of ivPCA. For 2SDR, the size of the covariance matrices is $p \times p$, $q \times q$ and followed by $p_0q_0 \times p_0q_0$, where (p_0, q_0) is the reduced size of data after MPCA. For ivPCA, the size of covariance matrix is $pq \times pq$. Recall that the computation cost for solving the first r eigenvectors of a covariance matrix with size $m \times m$ is $O(m^2 r)$ [12]. As for the 80s Ribosome particle image set, the resulting overall complexity for solving eigenvectors of 2SDR is of order $O(10^5 \times p_0 + 10^5 \times q_0)$, where ivPCA is $O(10^{10} \times r)$.[2] A practical problem in implementing 2SDR is the rank selection. Many rank selection approaches for PCA may be used in 2SDR with some modifications. For high dimensional data with moderate or even smaller sample size, SURE method has been shown to have better performance than AIC and BIC, both for PCA [14] and MPCA [15]. For the cryo-EM particle dataset, the number of images usually have similar order of magnitude with features ($O(n) = O(p \times q)$). On the other hand, for the movie files in cryo-EM, the size of samples is much lower than features ($n < p \times q$). Both situations satisfy the condition of SURE, thus we suggest to select the rank of 2SDR according to it. Table 1 gives a summary of different dimension reduction methods we discussed so far.

3. Numerical Study Using Simulation Data.

3.1. *Simulation data.* In this simulation, we tested the performance of ivPCA, MPCA and 2SDR under various settings. We generated the data to follow both model (1), where there is no structure assumption on Γ and model (5), where Γ is

[2] The cost of solving the $p_0q_0 \times p_0q_0$ covariance matrix will be smaller than the cost of solving the other two covariance matrices in this work

TABLE 1
Summary of different dimension reduction methods, the means are omitted for simplicity

	ivPCA	MPCA	2SDR
Model	$\text{vec}(X) = \Gamma\nu + \varepsilon$	$X = AUB^\top + \varepsilon$	$X = AUB^\top + \varepsilon$ $\text{vec}(U) = G\nu$
Eigen space	$\text{span}(\widehat{\Gamma})$	$\text{span}(\widehat{B}\otimes\widehat{A})$	$\text{span}((\widehat{B}\otimes\widehat{A})\cdot\widehat{G})$
Computation complexity	$O(p^2q^2r)$	$O(p^2p_0 + q^2q_0)$	$O((p^2p_0 + q^2q_0) + (p_0q_0)^2r)$
Scores	$\widehat{\Gamma}^\top\text{vec}(X)$	$(\widehat{B}\otimes\widehat{A})^\top\text{vec}(X)$	$\widehat{G}^\top(\widehat{B}\otimes\widehat{A})^\top\cdot\text{vec}(X)$
Reconstructions	$P_{\widehat{\Gamma}}\text{vec}(X)$	$P_{\widehat{B}\otimes\widehat{A}}\text{vec}(X)$	$P_{(\widehat{B}\otimes\widehat{A})\cdot\widehat{G}}\text{vec}(X)$

embedded with a Kronecker envelope structure. Then, n i.i.d. copies of $\nu \sim N(0,\Lambda)$ and $\text{vec}(\mathscr{E}) \sim N(0,\sigma^2 I_m)$ are generated. Parameters including sample size n and the variance of error σ^2 are varied among these settings to reflect various signal-to-noise ratio (SNR) levels of the data. In each setting, 50 iterations of simulations are conducted.

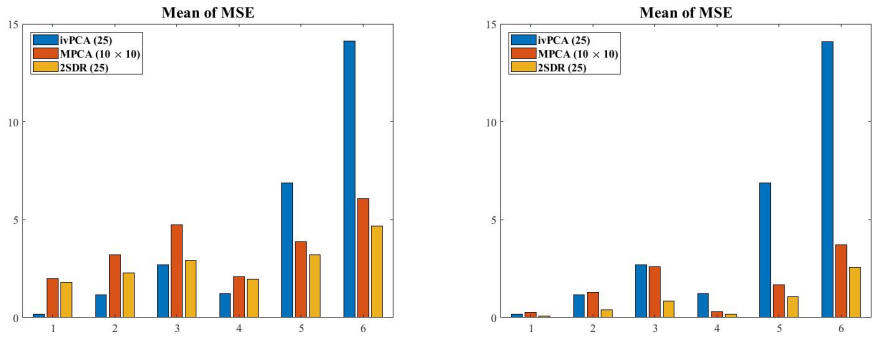
We used Mean Square Error (MSE) to evaluate the performance of each method.

$$MSE = \frac{1}{pqn}\sum_n \|\Gamma\nu_i - \text{vec}(\widehat{X}_i)\|_2^2 \tag{8}$$

where $\text{vec}(\widehat{X}_i) = \text{vec}(\overline{X}) + \widehat{\Gamma}\widehat{\Gamma}^\top\big(\text{vec}(X_i) - \text{vec}(\overline{X})\big)$. MSE averages the difference between the reconstruction and the true underlying signal which is crucial for practical applications.

As can be observed from Figure 3(a), even when the underlying model does not favor 2SDR, 2SDR has smaller MSE than ivPCA in the settings 5 and 6, where both SNR and sample size are comparatively small. On the other hand, 2SDR outperforms MPCA in terms of MSE in all settings of Figure 3(b) when the underlying model follows (5). The difference becomes more obvious when the sample size is small or when the SNR is low. In short, 2SDR is the best choice under model (5). It uses smaller dimension than MPCA, employs fewer parameters and gets finer estimate of Γ than ivPCA, and reaches the lowest MSE among the three.

3.2. *Synthetic cryo-EM images.* In this subsection, synthetic cryo-EM images are used to further examine the effect of the three dimension reduction methods in various SNR values. We downloaded the Relion classification benchmark dataset, E. coli 70s Ribosome, which contains 10000 particle images with box size 130×130. The first 5000 and the second 5000 images of this dataset have different structure conformations. The synthetic dataset of 70s Ribosome is prepared as follows. Firstly, the 3D structure is obtained from the first 5000 images of original experimental dataset using CryoSparc[16] which represents the Ribosome bound with an elongation factor (EF-G). Then, a total of 50 distinct 2D images with



(a) Assuming ivPCA model

(b) Assuming 2SDR model

Fig 3: Reconstruction errors of simulation data. Six simulations under different settings for data generated from model (1)(a) and from model (5)(b). 1, 2, 3: Large sample size where the number of sample is in the same order of pixels which represent the case for a typical particle dataset, 4, 5, 6: Small sample size which is the case for regular movie files; 1, 4: SNR ≈ 0.5; 2, 5: SNR ≈ 0.1; 3, 6: SNR ≈ 0.05. The details of each setting are in the Supplementary Table 1 and Supplementary Table 2.

130×130 pixels were generated by projecting the 3D structure of 70S Ribosome in equally spaced (angle-wise) orientations. Second, 5000 images were generated from these 50 projections with 100 copies for each projection. Third, each image was then convoluted with the electron microscopy contrast transfer function randomly sampled from a set of 50 CTF values. [3] Finally, i.i.d. Gaussian noise $N\left(0, \sigma^2 I_m\right)$ with different σ^2 is added to generate 3 dataset such that the SNR is equal to 0.1, 0.05 and 0.01, respectively.

In each setting, 50 iterations of simulations are conducted using ivPCA, MPCA and 2SDR to perform dimension reduction. As can be observed in Figure 4, 2SDR outperforms the other two methods in terms of the MSE between original clean data and reconstruction data. The difference becomes more clear as the SNR goes low. The comparison of the reconstruction results along with the simulated images are shown in Figure 5 where the SNR is decreased from left-hand side to right-hand side. One can perceive that when SNR equals to 0.1, both ivPCA and 2SDR can preserve the high resolution features of particles while greatly lower the noise. When the noise increase, we observe that the contrast of the reconstructions from

[3] Here, the defocus is randomly sampled from 1.5um to 2um and the astigmatism angle is from 0.2 to 1.4 radian. The electron beam accelerating voltage was set to be 300KeV with spherical aberration Cs = 2mm, amplitude contrast=0.07 and pixel size is 2.82.

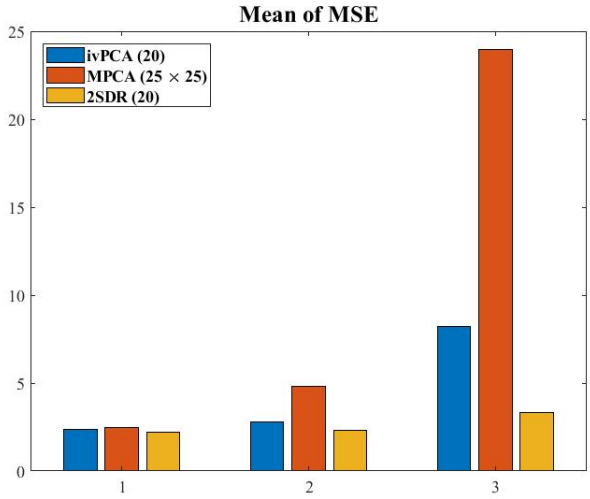


Fig 4: Reconstruction errors of synthetic cryo-EM images. Three simulations under different settings. 1: SNR $= 0.1$; 2: SNR $= 0.05$; 3: SNR $= 0.01$. The details of each setting are in the Supplementary Table 3.

ivPCA clearly decrease and the high resolution information also reduce a lot especially when the SNR drops to 0.01. On the other hand, 2SDR not only maintains the contrast of particles but preserves the high resolution details well even when the SNR decrease to 0.01. Finally, notice that the streak patterns presented in the reconstructions of MPCA decrease the contrast between particle and background and the situation is worsened when SNR decrease. In summary, there are three major observations from this experiment. First, it shows that 2SDR can eliminate the streak patterns that exist in MPCA reconstructions to enhance the contrast. Second, the details of the particles can be recovered by preserving particle signal and reduce the background noise comparing with ivPCA. Finally, the advantage becomes more evident as the SNR decrease which is consistent with the results from simulation dataset.

4. Real Data Applications.

4.1. *2SDR on denoising particle images of 70s Ribosome.* In this section, we utilized the real data collected from cryo-EM experiments to investigate whether or not the usage of 2SDR will benefit the reconstruction as observed in Section 3. We chose the first 5000 experiment images from the 70s Ribosome dataset for testing where Xmipp[17] package is applied for phase correction and Spider[18] package

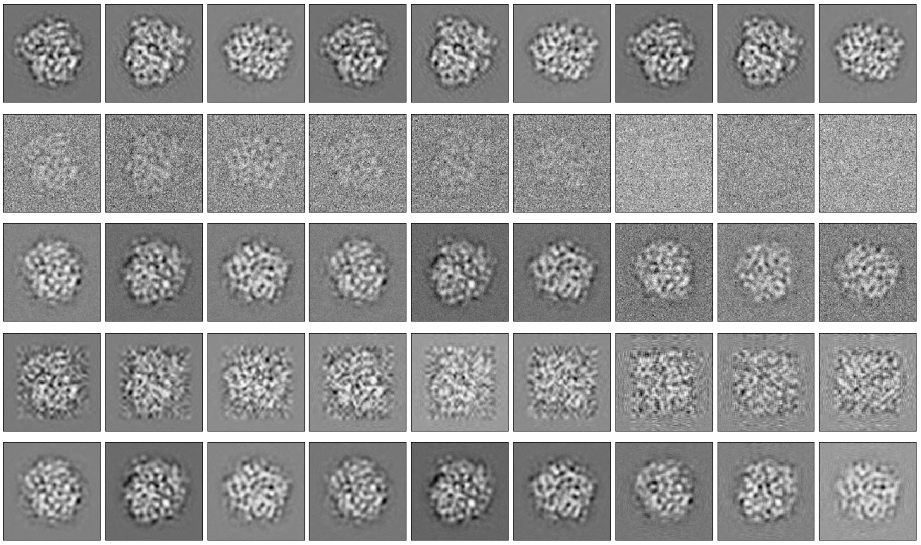


Fig 5: Reconstruction of synthetic 70s Ribosome images by dimension reduction. The first row shows the 9 original clean images. The second row shows the noisy images correspond to different SNR levels. The third row is the result by applying ivPCA to the second row with 20 components. The fourth row is by MPCA with $(p_0, q_0) = (25, 25)$, which contributes 625 basis components. The fifth row is by 2SDR: $(p_0, q_0) = (25, 25)$ for MPCA follow by choosing the first 20 components from PCA. Note that the first 3 columns correspond to SNR$=0.1$, the middle 3 columns correspond to SNR$=0.05$ and the last 3 columns correspond to SNR$=0.01$.

for alignment.

We compared the 3 dimension reduction methods, including ivPCA, MPCA and 2SDR, and presented their reconstructed images of 9 randomly selected particles in Figure 6. The result shows that 2SDR (rows 4) perform much better than MPCA (rows 3), which demonstrates the improvement made by adding the second step.

4.2. *2SDR on denoising particle images of 80s Ribosome.* To further examine our scheme on larger dataset, 80s Ribosome that comes from Relion Benchmark example is used. This dataset contains 105,247 particle images with pixel size 360 by 360. The complexity of solving the complete eigenvector set[4] is in the order of 10^{15} for ivPCA that many current PCA implementations fail to handle such huge amount of work loading due to the limitation of the underlying numerical linear

[4]Recall that the computation complexity of solving the full eigenvector set of $m \times m$ covariance matrix is $O(m^3)$. In addition, a rough approximation of 360^2 by 10^5 is used.

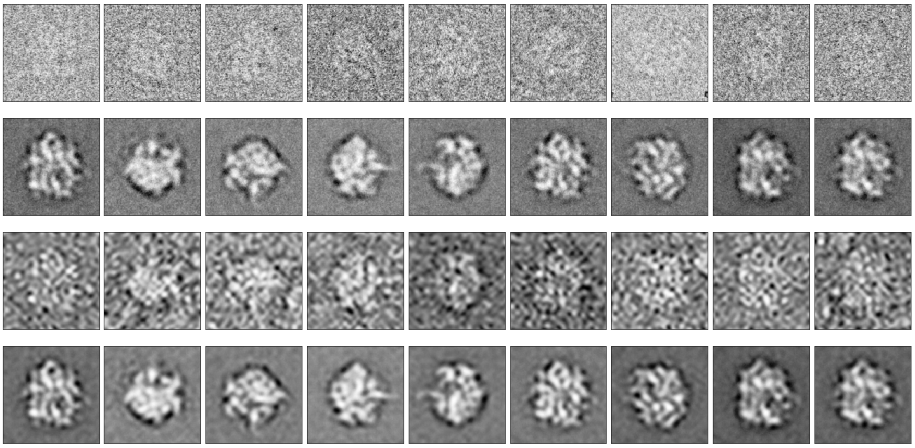


Fig 6: Reconstruction of experiment 70s Ribosome images by dimension reduction. The first row shows the 9 original images after alignment. The second row is by ivPCA with 20 components. The third row is by MPCA with $(p_0, q_0) = (25, 25)$, which contributes 625 basis components. The fourth row is by 2SDR: $(p_0, q_0) = (25, 25)$ for MPCA follow by choosing the first 20 components from PCA.

algebra package LAPACK[19]. [5] In contrast, we can perform 2SDR in the server since the computing complexity has been reduced by several orders of magnitude.

In Figure 7, we show 9 randomly selected images and their reconstruction images by ivPCA, MPCA and 2SDR. The result shows that, while gaining the computation advantage over ivPCA, 2SDR has comparable reconstruction quality as ivPCA. Similarly to Figure 6, 2SDSR performs much better than MPCA.

4.3. *2SDR application on micrographs.* To test the performance of 2DSR on the micrographs, the Beta-galactosidase dataset which contains 15 movie files were downloaded from Relion tutorial data. We use the first movie file to demonstrate the applicability 2SDR on images of extremely large pixels. The first movie has 16 frames with pixel size 1950×1950 whose vectorization leads to the dimension size of 3,802,500.

Figure 8(a)-(c) shows a denoised frame by ivPCA, by MPCA and by 2SDR respectively. The sigma-contrast from Relion[20] which re-distribute and clamp the pixels (i.e. black is 3 standard deviation below the mean and white is 3 standard deviation above the mean) and Gaussian filter[17] are used to further improve the

[5] We used a server equipped with two Intel Xeon CPU E5-2699 v4 at 2.20GHz and 512GB memory to execute ivPCA for this dataset. Though we can not obtain the full eigenvector set, top eigenvectors can still be extracted using algorithms like power method.

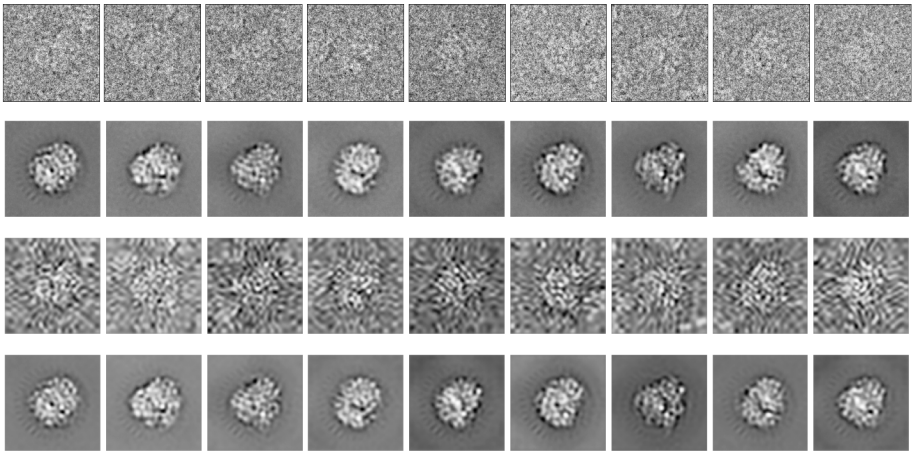


Fig 7: Reconstruction of experiment 80s Ribosome images by dimension reduction. The first row shows the 9 original images after alignment. The second row is by ivPCA with 20 components. The third row is by MPCA with $(p_0, q_0) = (26, 26)$. The fourth row is by 2SDR: $(p_0, q_0) = (26, 26)$ for MPCA follow by choosing the first 20 components from PCA.

contrast as presented in 8(d)-(e). We observe that the de-noise images by 2SDR and MPCA clearly reveals the particle shapes while that by ivPCA does not.

This example suggests that 2SDR might be a powerful tool in de-noising the movie files. Since the particles are clearly revealed in each single frame, we executed an experiment to test if the structure information is lost by applying 2SDR. We did particle picking on each denoised frame and continued processing until the end of the workflow in Figure 1 to generate a 3D map. The power spectrum of original single frame is greatly improved after applying 2SDR as shown in Figure 9(a),(b). Applying CryoSparc[16] package, we parallelly processed the original 15 micrograph movies and the denoised 15×16 movie frames. Their 3D reconstruction models and their corresponding Fourier Shell Correlation (FSC) curves in Figure 9 demonstrate that the denoised data does not lose high resolution information and even has slightly better performance.

5. Discussion. Many current statistical algorithms are designed to be applied on vector data. When facing matrix data or high order tensor data, the naive approach usually employs data vectorization to accommodate the input format. The consequence is that the computation cost becomes a serious challenge due to the rapid growth of the dimension. Using structural modeling on the eigenbasis, 2SDR provides a solution. Our analysis on real data of particle images suggests that 2SDR

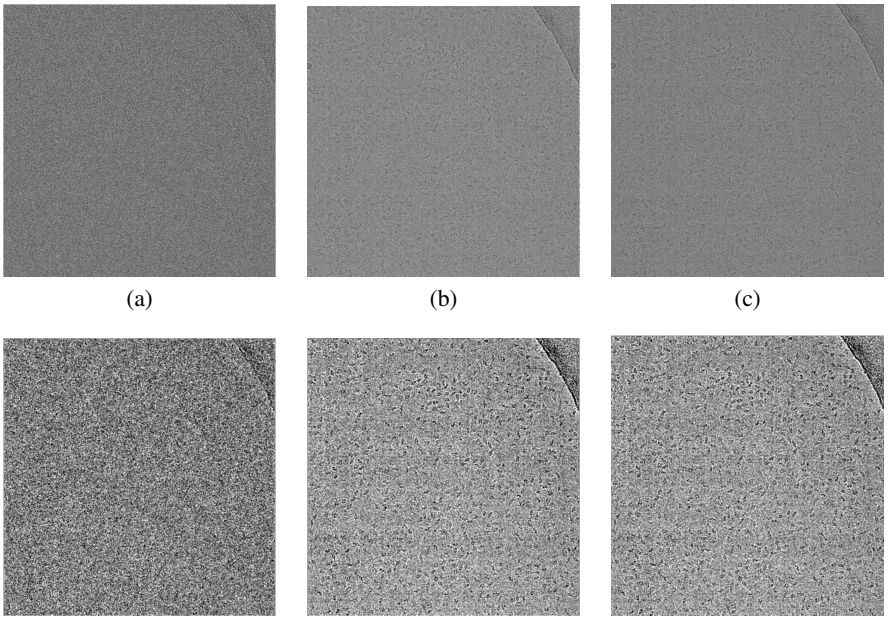


Fig 8: Denoise of raw frame in a movie file of Beta-galactosidase dataset. (a) ivPCA reconstructions using 12 components. (b) MPCA reconstructions with $(p_0, q_0) = (26, 24)$. (c) 2SDR reconstructions with $(p_0, q_0) = (26, 24)$ and choose $r = 12$ components in the second stage. (d-f) are the results after applying the contrast enhancement method on (a-c), respectively.

based on Kronecker envelope model has the advantage of computation cost over ivPCA and preserves the signal information better than MPCA. In terms of denoising performance, our simulation study shows that 2SDR performs better than ivPCA and MPCA when SNR is low, which is the case of cryo-EM images.

One major concern to apply a dimension reduction method on cryo-EM images is the possible loss of high resolution information, which may lead to weakening the final resolution of the 3D density map. For example, those denoising tools based on Fourier transform, like low-pass filtering, tend to lose high resolution information because such information often sits in high frequency components. When using the denoised particle images for the subsequent analysis of 2D clustering, we do observe the high resolution information loss in their 2D cluster averages. Surprisingly, when applying 2SDR on the micrographs, the high resolution information is well preserved.

In summary, this work uses the denosing examples to demonstrate the advan-

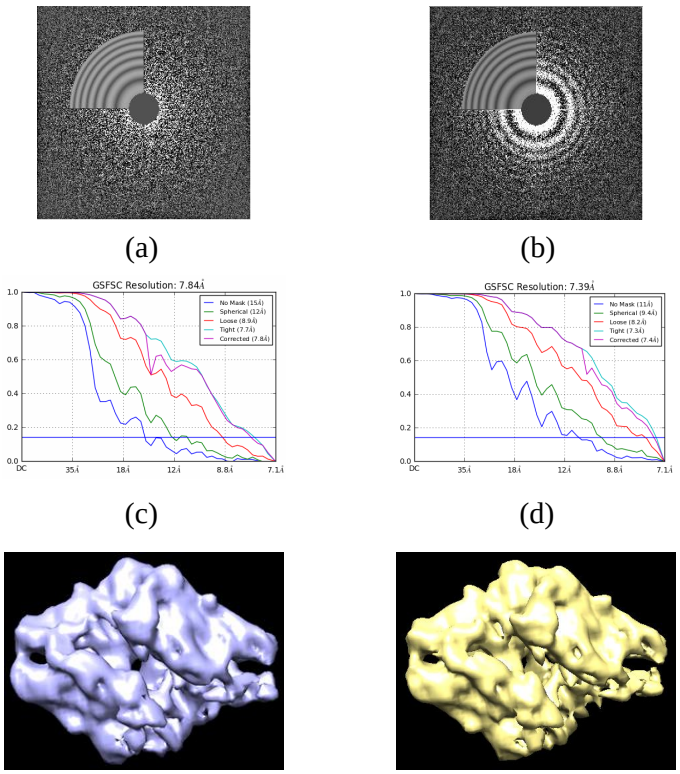


Fig 9: (a) Power spectrum of original frame. (b) Power spectrum of denoised frame through 2SDR. (c) FSC curve using original workflow. (d) FSC curve using single-frame workflow. (e) 3D reconstruction map in original workflow. (f) 3D reconstruction map in single-frame workflow.

tage of 2SDR over ivPCA. Other important cryo-EM analysis such as analyzing discrete heterogeneity[21] and obtaining energy landscape associated with 3D structure[22, 23] which employ ivPCA on input data may also benefit from 2SDR. Furthermore, 2SDR is not limited to the cryo-EM application, it is a general and powerful dimension reduction method that can serve as an alternative to ivPCA when dealing with 2D images of huge number of pixels or higher order tensors.

Acknowledgement. This work is supported by Academia Sinica:AS-GCS-108-08 and MOST:106-2118-M-001-001 -MY2. The authors are with Institute of Statistical Science, Academia Sinica, Taiwan.

References.

[1] Hung Hung, Pei-Shien Wu, I-Ping Tu, and Su-Yun Huang. On multilinear principal component analysis of order-two tensors. *Biometrika*, 99:569–583, 2012.

[2] J. Ye. Generalized low rank aapproximation of matrices. *Machine Learning*, 61:167–191, 2005.

[3] Ting-Li Chen, Su-Yun Huang, Hung Hung, and I-Ping Tu. An introduction to multilinear principal component analysis. *Journal of the Chinese Statistical Association*, 52(1):24–43, 2014.

[4] James R Schott. Tests for Kronecker envelope models in multilinear principal components analysis. *Biometrika*, 101(4):978–984, 2014.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[6] Hans Elmlund, Dominika Elmlund, and Samy Bengio. Prime: probabilistic initial 3d model generation for single-particle cryo-electron microscopy. *Structure*, 21(8):1299–1306, 2013.

[7] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of Structural Biology*, 157(1):38–46, 2007.

[8] Yifan Cheng, Nikolaus Grigorieff, Pawel A Penczek, and Thomas Walz. A primer to single-particle cryo-electron microscopy. *Cell*, 161(3):438–449, 2015.

[9] Marin Van Heel and Joachim Frank. Use of multivariates statistics in analysing the images of biological macromolecules. *Ultramicroscopy*, 6(1):187–194, 1981.

[10] Joachim Frank. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press, 2006.

[11] Wilson Wong, Xiao-chen Bai, Alan Brown, Israel S Fernandez, Eric Hanssen, Melanie Condron, Yan Hong Tan, Jake Baum, and Sjors HW Scheres. Cryo-em structure of the plasmodium falciparum 80s ribosome bound to the anti-protozoan drug emetine. *Elife*, 3:e03080, 2014.

[12] Victor Y Pan and Zhao Q Chen. The complexity of the matrix eigenproblem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 507–516. ACM, 1999.

[13] L. De Lathauwer, B. De Moor, and J. Vandewalle. On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21:1324–1342, 2000b.

[14] Magnus O Ulfarsson and Victor Solo. Rank selection in noist pca with sure and random matrix theory. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3317–3320. IEEE, 2008.

[15] I-Ping Tu, Su-Yun Huang, and Dai-Ni Hsieh. The generalized degrees of freedom of multilinear principal component analysis. *Journal of Multivariate Analysis*, 173:26–37, 2019.

[16] Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker. cryosparc: algorithms for rapid unsupervised cryo-em structure determination. *Nature Methods*, 14(3):290, 2017.

[17] C.O. Sorzano, R. Marabini, J. Velazquez-Muriel, J.R. Bilbao-Castro, S.H. Scheres, J.M. Carazo, and A. Pascual-Montano. Xmipp: a new generation of an open-source image processing package for electron microscopy. *Journal of Structural Biology*, 148(2):194–204, 2004.

[18] J. Frank, M. Radermachera, P. Penczeka, J. Zhua, Y. Li, M. Ladjadj, and A Leitha. Spider and web: Processing and visualization of images in 3d electron microscopy and related fields. *Journal of Structural Biology*, 116(1):190–199, 1996.

[19] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, Alan McKenney, and Danny Sorensen. *LAPACK Users' guide*, volume 9. Siam, 1999.

[20] Sjors HW Scheres. Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519–530, 2012.

[21] Pawel A Penczek, Marek Kimmel, and Christian MT Spahn. Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-em images. *Structure*, 19(11):1582–1590, 2011.

[22] David Haselbach, Jil Schrader, Felix Lambrecht, Fabian Henneberg, Ashwin Chari, and Holger Stark. Long-range allosteric regulation of the human 26s proteasome by 20s proteasome-targeting cancer drugs. *Nature Communications*, 8:15578, 2017.

[23] David Haselbach, Ilya Komarov, Dmitry E Agafonov, Klaus Hartmuth, Benjamin Graf, Olexandr Dybkov, Henning Urlaub, Berthold Kastner, Reinhard Lührmann, and Holger Stark. Structure and conformational dynamics of the human spliceosomal bact complex. *Cell*, 172(3): 454–464, 2018.

SUPPLEMENTARY TABLE 1
Detail settings and results in Figure 3 (a).

Parameters	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=4,n=1000,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=20,n=1000,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	0.1832	1.9866	1.8082	1.1698	3.1934	2.2839
Variance of MSE	0.0000	0.0002	0.0002	0.0001	0.0004	0.0003
	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=40,n=1000,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=4,n=100,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	2.7059	4.7402	2.9088	1.2264	2.0770	1.9598
Variance of MSE	0.0004	0.0004	0.0004	0.0001	0.0032	0.0033
	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=20,n=100,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=40,n=100,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	6.8816	3.8925	3.1991	14.1142	6.0710	4.6598
Variance of MSE	0.0022	0.0033	0.0033	0.0048	0.0036	0.0036

SUPPLEMENTARY TABLE 2
Detail settings and results in Figure 3 (b).

Parameters	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=4,n=1000,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=20,n=1000,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	0.1830	0.2561	0.0740	1.1690	1.2908	0.3926
Variance of MSE	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=40,n=1000,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=4,n=100,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	2.7069	2.6111	0.8378	1.2269	0.3079	0.1760
Variance of MSE	0.0002	0.0002	0.0001	0.0001	0.0000	0.0000
	$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=20,n=100,\lambda_{26-i}=10i$			$p=40,q=40,p_0=10,q_0=10,$ $r=25,\sigma^2=40,n=100,\lambda_{26-i}=10i$		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	6.8775	1.6672	1.0801	14.1062	3.7130	2.5575
Variance of MSE	0.0032	0.0006	0.0004	0.0069	0.0030	0.0028

SUPPLEMENTARY TABLE 3
Detail settings and results in Figure 4. $p = 130, q = 130, p_0 = 25, q_0 = 25, r = 20$ *in all settings.*

Parameters	SNR= 0.1			SNR= 0.05		
Method	ivPCA	MPCA	2SDR	ivPCA	MPCA	2SDR
Mean of MSE	2.3876	2.4838	2.2089	2.7755	4.8036	2.3163
Variance of MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	SNR= 0.01					
Method	ivPCA	MPCA	2SDR			
Mean of MSE	8.2121	23.9534	3.3121			
Variance of MSE	0.0016	0.0005	0.0001			