MISALIGNED PRINCIPAL COMPONENTS ANALYSIS: APPLICATION TO GENE EXPRESSION TIME SERIES ANALYSIS

Arnau Tibau-Puig[1], *Ami Wiesel*[2], *Raj Rao Nadakuditi*[1], *Alfred O. Hero III*[1]

(1) University of Michigan - Department of Electrical Engineering
1301 Beal Avenue, Ann Arbor, MI, 48109-2122 USA
(2) The Hebrew University of Jerusalem - School of Computer Science and Engineering
55 Ross Building, Jerusalem, 91904 Israel

ABSTRACT

Principal Component Analysis (PCA) is a widely applied method for extracting structure from samples of high dimensional biological data. Often there exist misalignments between different samples and this can cause severe problems in PCA if not properly taken into account. For example, subject-dependent temporal differences in gene expression response to a treatment will create relative time shifts in the samples that decohere the PCA analysis. Depending on the characteristics of the underlying signal, the sensitivity of PCA to such misalignments is severe, leading to a phase transition phenomenon that can be studied using the spectral theory of autocorrelation matrices. With this as motivation, we propose a new method of PCA, called MisPCA, that explicitly accounts for the effects of misalignments in the samples. We illustrate MisPCA on clustering longitudinal temporal gene expression data.

1. INTRODUCTION

Principal Component Analysis (PCA) [1] is a widely used technique for dimensionality-reduction of high dimensional data, with applications in pattern recognition [2], blind channel estimation [3] and network-traffic anomaly detection [**?**]. In all these applications, PCA can be used to separate the latent features corresponding to signal from the random fluctuations of noise. The fundamental assumption underlying this approach is that the signal lies in a lower dimensional subspace, while the noise is random and isotropic; spreading its power across all directions in the observation space.

Unfortunately, in many cases, despite the appropriateness of the low-dimensional subspace model, measurement limitations can lead to observations revealing different signal subspaces. This occurs for example when the sampling times across observations can not be synchronized appropriately, due to technical limitations or to different temporal latencies of the phenomena under study. In previous work, we considered an Order Preserving Factor Analysis (OPFA) model that accounted for order-preserving circular shifts in each factor and we demonstrated its effectiveness for extracting order-preserving factors from misaligned data [4]. Here, we propose an alternative approach to OPFA that applies to misaligned data without order restrictions and is applicable to larger sample sizes.

In this paper, we consider the limitations of PCA for the problem of estimating a rank-1 signal subspace from high-dimensional misaligned data. We introduce a modified version of PCA, called Misaligned PCA (MisPCA), that simultaneously aligns the data and estimates the aligned signal subspace. For signal subspaces of rank greater than one, a deflation procedure is applied to sequentially estimate succesive principal components.

The paper is divided into two parts. First, we propose a simple approximation of the combinatorial MisPCA estimation problem that considerably improves the PCA estimate whenever misalignments are present. Second, building on recent results in random matrix theory [5, 6], we derive high-dimensional asymptotic results that characterize the minimum SNR necessary to detect and estimate the signal from the sample covariance.

This paper is organized as follows. Section 2 introduces the misaligned signal model. We give algorithms for Misaligned PCA in Section 3. Section 4 studies the statistical effects of misalignments on the sample covariance. We present numerical results and a gene expression data analysis application in Section 5 and we conclude the paper in Section 6.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscript ^{T} denotes the transpose operator. Given a symmetric matrix $\boldsymbol{X}$, $\lambda_i(\boldsymbol{X})$ and $\boldsymbol{v}_i(\boldsymbol{X})$ refer to its i-th eigenvalue and eigenvector, respectively. $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ denote the trace and determinant operators. $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

[1]The work of A. Tibau-Puig and A.O. Hero was supported in part by DARPA under the PHD program. The views, opinions, and findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

2. PROBLEM FORMULATION

We consider the following discrete-time, circularly misaligned, rank-1 signal model,

$$x_i[k] \quad = a_i h[k-d_i] + \epsilon_i[k], \quad i = 1, \cdots, n. \qquad (1)$$

Here $h[k]$ is an unknown real sequence of length equal to p and indexed by k, and the integer valued elements of the vector $\boldsymbol{d} \in \{0, \cdots, d_{\max}\}^n$ parameterize the amount of circular shift in each observation, with $d_{\max} < p$. For each $i = 1, \cdots, n$, the random variables a_i are i.i.d, zero-mean Gaussian and the p-length sequences $\epsilon_i[k]$ are i.i.d., zero-mean Gaussian white processes. To simplify the notation, we will further assume that $E\left[\epsilon_i^2[k]\right] = \sum_{k=1}^{p} h^2[k] = 1$, and we define the Signal-to-Noise Ratio (SNR) as:

$$\text{SNR} = \frac{E\left[a_i^2\right]}{E\left[\epsilon_i^2[k]\right]}.$$

The problem considered in this paper is that of estimating the signal sequence $h[k]$ from a collection of observations obeying model (1). For convenience, we will write (1) in vector form:

$$\boldsymbol{x}_i \quad = a_i \boldsymbol{C}_{d_i} \boldsymbol{h} + \boldsymbol{\epsilon}_i, \quad i = 1, \cdots, n,$$

where $\boldsymbol{x}_i$, $\boldsymbol{h}$ and $\boldsymbol{\epsilon}_i$ are p-dimensional real vectors, and $\boldsymbol{C}_{d_i}$ is a $p \times p$ circular shift matrix with shift equal to d_i:

$$\left[\boldsymbol{C}_{d_i}\right]_{k,l} = \begin{cases} 1 & \text{if } k = (d_i + l) \bmod p \\ 0 & \text{otherwise.} \end{cases}$$

Using the properties of a_i and $\boldsymbol{\epsilon}_i$ we can conclude that $\boldsymbol{x}_i$ follows a multivariate Gaussian distribution with zero mean and covariance:

$$\boldsymbol{\Sigma}_i = E\left[\boldsymbol{x}_i \boldsymbol{x}_i^T\right] = \text{SNR}\, \boldsymbol{C}_{d_i} \boldsymbol{h}\boldsymbol{h}^T \boldsymbol{C}_{d_i}^T + \boldsymbol{I}_p. \qquad (2)$$

3. ALGORITHMS

In general, the covariance matrix of each observation is not the same for all $i = 1, \cdots, n$. However, equation (2) reflects an underlying rank-1 structure corresponding to the signal $\boldsymbol{h}$. In this section we propose to exploit this fact by estimating $\boldsymbol{h}$ from the joint likelihood of the misaligned data $\{\boldsymbol{x}_i\}_{i=1}^n$. The log-likelihood function is:

$$l(\boldsymbol{h}, \boldsymbol{d}, \text{SNR}) = c - \sum_{i=1}^{n} \text{tr}\left(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T\right) - \sum_{i=1}^{n} \log\det \boldsymbol{\Sigma}_i$$

where c denotes a constant independent of the relevant parameters. Using the Sherman-Morrison-Woodbury matrix inversion formula,

$$l(\boldsymbol{h}, \boldsymbol{d}, \text{SNR}) = c + n\frac{\text{SNR}}{1+\text{SNR}} \boldsymbol{h}^T \boldsymbol{S}(\boldsymbol{d})\, \boldsymbol{h} - n\log(\text{SNR}+1),$$

where, for any $\boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n$, possibly different from $\boldsymbol{d}$, we define the $p \times p$ matrix:

$$\boldsymbol{S}(\boldsymbol{\tau}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{C}_{\tau_i}^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{C}_{\tau_i}. \qquad (3)$$

This quantity can be interpreted as an *aligned sample covariance matrix*, with alignment parameter equal to $\boldsymbol{\tau}$. When $\boldsymbol{\tau} = \boldsymbol{0}$, this coincides with the sample covariance.

Maximizing $l(\boldsymbol{h}, \boldsymbol{d}, \text{SNR})$ under the constraint $||\boldsymbol{h}||_2 = 1$ for a fixed SNR leads to the Misaligned Principal Component Analysis (MisPCA) solution:

$$\begin{aligned} \lambda^{\text{MisPCA}} \quad &= \max \quad \lambda_1(\boldsymbol{S}(\boldsymbol{\tau})) \\ &\text{s.t.} \quad \boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n, \end{aligned} \qquad (4)$$

which consists of finding the alignment vector $\boldsymbol{\tau}$ that maximizes the leading eigenvalue of the aligned covariance $\boldsymbol{S}(\boldsymbol{\tau})$. The optimal alignment is denoted by $\boldsymbol{d}^{\text{MisPCA}}$, and the corresponding MisPCA signal estimate is given by:

$$\boldsymbol{h}^{\text{MisPCA}} = \boldsymbol{v}_1\left(\boldsymbol{S}\left(\boldsymbol{d}^{\text{MisPCA}}\right)\right).$$

To estimate the SNR, it suffices to maximize $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \text{SNR}\right)$ under the constraint $\text{SNR} \geq 0$. The optimum occurs at (see Appendix A.1):

$$\text{SNR}^{\text{MisPCA}} = \begin{cases} 0 & \text{if } \lambda^{\text{MisPCA}} < 1 \\ \lambda^{\text{MisPCA}} - 1 & \text{otherwise.} \end{cases} \qquad (5)$$

Unfortunately, the MisPCA problem (4) is combinatorial, and exhaustive search is prohibitive even for small n. Here we consider two simple approximate solutions to (4). The first approximation ignores the misalignments altogether, i.e. solving (4) with $\boldsymbol{d} = \boldsymbol{0}$. This leads to the usual PCA estimate of $\boldsymbol{h}$:

$$\boldsymbol{h}^{\text{PCA}} = \boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{0})). \qquad (6)$$

The second approximation, alternatively estimates $\boldsymbol{d}$ and $\boldsymbol{h}$. At each iteration $t > 1$, we compute:

$$\begin{aligned} \boldsymbol{d}_t^{\text{A-MisPCA}} \quad &= \quad \arg \max_{\boldsymbol{\tau} \in \{0,\cdots,d_{\max}\}^n} \boldsymbol{h}_{t-1}^{\text{A-MisPCA}\,T} \boldsymbol{S}(\boldsymbol{\tau})\, \boldsymbol{h}_{t-1}^{\text{A-MisPCA}} \\ \boldsymbol{h}_t^{\text{A-MisPCA}} \quad &= \quad \boldsymbol{v}_1\left(\boldsymbol{S}\left(\boldsymbol{d}_t^{\text{A-MisPCA}}\right)\right) \end{aligned}$$

where we set $\boldsymbol{h}^0$ to an initial estimate of $\boldsymbol{h}$ and stop the algorithm when the change in likelihood is sufficiently small. We call this procedure Alternating MisPCA (A-MisPCA).

4. STATISTICS OF THE MISALIGNED COVARIANCE

The performance of the algorithms presented in the last section depend on the statistics of the leading eigenvalue and

eigenvector of the random matrix $\boldsymbol{S}(\boldsymbol{\tau})$, for a fixed, deterministic $\boldsymbol{\tau}$. In this section, we use recent asymptotic results on the spectrum of large random matrices [5, 6] to characterize the asymptotic behavior of $\lambda_1(\boldsymbol{S}(\boldsymbol{\tau}))$ and $\boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{\tau}))$ in the following setting: We assume that the number of variables $p = p_n$ grows linearly with the number of samples n so that, as n tends to infinity,

$$\lim_{n\to\infty} \frac{p_n}{n} = c > 0. \tag{7}$$

Note that this includes the possibility of p_n being larger than the number of observations n. Before we proceed to state the main result, we will need to define the following quantities. For any $\boldsymbol{t} \in \{0, \cdots, p-1\}^n$, define the function $\boldsymbol{s}(\boldsymbol{t}) : \{0, \cdots, p-1\}^n \to \{0, \cdots, n\}^p$, with coordinates given by:

$$s_i(\boldsymbol{t}) = \frac{|\{j \in \{1, \cdots, n\} \ : \ t_j = i-1\}|}{n} \tag{8}$$

where $|S|$ denotes the cardinality of a set S. (One can interpret $\boldsymbol{s}(\boldsymbol{t}) = [s_1(\boldsymbol{t}), \cdots, s_p(\boldsymbol{t})]$ as a histogram of the values in $\boldsymbol{t}$.) In addition, for any $\boldsymbol{h} \in \mathbb{R}^p$, we define the $p \times p$ autocorrelation matrix $\boldsymbol{R}_h$ of $\boldsymbol{h}$ as:

$$[\boldsymbol{R}_h]_{i,j} = \boldsymbol{h}^T \boldsymbol{C}_{i-j} \boldsymbol{h} \tag{9}$$

Finally, the expected value of $\boldsymbol{S}(\boldsymbol{\tau})$ is given by:

$$\boldsymbol{\Sigma}(\boldsymbol{\tau}) := E[\boldsymbol{S}(\boldsymbol{\tau})] = \text{SNR}\, \boldsymbol{H} \text{diag}\, \boldsymbol{s}(\boldsymbol{d} -_p \boldsymbol{\tau})\, \boldsymbol{H}^T + \boldsymbol{I}_p,$$

where $\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h} & \boldsymbol{C}_1\boldsymbol{h} & \cdots & \boldsymbol{C}_{p-1}\boldsymbol{h} \end{bmatrix}$, $\boldsymbol{d}$ denotes the true alignment parameter with which the data was generated, and $-_p$ indicates a modulo p subtraction.

The following result shows that the leading eigenpair of $\boldsymbol{S}(\boldsymbol{\tau})$ matches that of $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ only if the SNR is higher than a phase transition SNR which depends on the unknown parameters of the model, $\boldsymbol{h}$ and $\boldsymbol{d}$.

Theorem 4.1 *Let $\boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n$ and $\boldsymbol{S}(\boldsymbol{\tau})$ be the $p_n \times p_n$ aligned sample covariance evaluated at $\boldsymbol{\tau}$, defined in (3). Let SNR, $\boldsymbol{h}$ and $\boldsymbol{d}$ be the true model parameters as defined in Section II. Then, assuming (7), as $p_n, n \to \infty$,*

$$\lambda_1(\boldsymbol{S}(\boldsymbol{\tau})) \overset{a.s.}{\to} \begin{cases} (\text{SNR}\gamma + 1)\left(1 + \frac{c}{\text{SNR}\gamma}\right) & \text{SNR} > \frac{\sqrt{c}}{\gamma} \\ (1+\sqrt{c})^2 & \text{otherwise}, \end{cases}$$

and:

$$|\langle \boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{\tau})), \boldsymbol{w}\rangle|^2 \overset{a.s.}{\to} \begin{cases} \frac{(\text{SNR}\gamma)^2 - c}{(\text{SNR}\gamma)^2 + c\text{SNR}\gamma} & \text{SNR} > \frac{\sqrt{c}}{\gamma} \\ 0 & \text{otherwise}, \end{cases}$$

where $\overset{a.s.}{\to}$ denotes almost sure convergence, $\boldsymbol{w} = \boldsymbol{v}_1(\boldsymbol{\Sigma}(\boldsymbol{\tau}))$, and c is defined in (7). Here, γ is the gain/loss due to misalignments (i.e. $\boldsymbol{\tau}$ being different from $\boldsymbol{d}$), and is given by:

$$\gamma = \lambda_1\left(\text{diag}\left(\boldsymbol{s}(\boldsymbol{d} -_p \boldsymbol{\tau})\right)^{\frac{1}{2}} \boldsymbol{R}_h \text{diag}\left(\boldsymbol{s}(\boldsymbol{d} -_p \boldsymbol{\tau})\right)^{\frac{1}{2}}\right), \tag{10}$$

where $\boldsymbol{s}(\boldsymbol{t})$ and $\boldsymbol{R}_h$ are defined in (8) and (9), respectively.

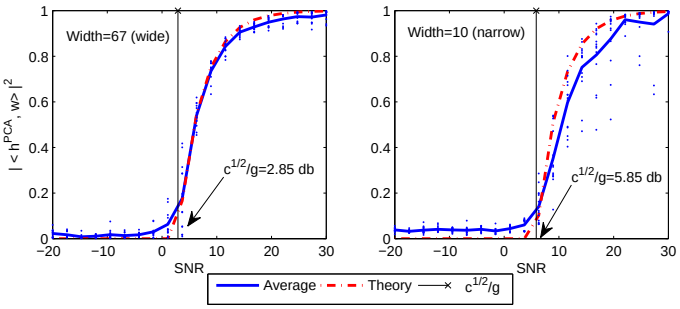

Fig. 1. Predicted and average values of $|\langle \boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{0})), \boldsymbol{v}_1(\boldsymbol{\Sigma}(\boldsymbol{0}))\rangle|^2$ for $\boldsymbol{h} \in \mathbb{R}^{200}$ equal to a rectangular pulse with width 67 and 10. The phase transition SNR predicted by Theorem 4.1, denoted by $\frac{c^{1/2}}{g}$ in the figure, is higher for narrow signals, which are less robust to misalignments.

See Appendix A.2 for a proof. This result is better understood graphically. Figure 1 shows the average $|\langle \boldsymbol{v}_1(\boldsymbol{S}(\boldsymbol{0})), \boldsymbol{w}\rangle|^2$ computed over 50 random realizations generated with model (1) for a signal of dimension $p = 200$, $n = 200$ samples and two choices of $\boldsymbol{h}$, with $d_{\max} = 100$. Notice that the empirical results accurately match the asymptotic theory.

Theorem 4.1 determines a "no-hope" regime for PCA and MisPCA. Consider for instance the PCA estimate, where $\boldsymbol{\tau} = \boldsymbol{0}$, and uniformly distributed misaligments $\boldsymbol{d}$ so that $\boldsymbol{s}(\boldsymbol{d}) = \frac{1}{d_{\max}}\boldsymbol{1}$. Then Theorem 4.1 implies that if the SNR is lower than

$$\frac{d_{\max}}{\lambda_1(\boldsymbol{R}_h)}\sqrt{c}, \tag{11}$$

the PCA estimate, defined in (6), is orthogonal to the leading eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{0})$, which contains partial information about the underlying signal $\boldsymbol{h}$. The scalar accompanying $\sqrt{c}$ in (11) can be interpreted as a tradeoff between the magnitude of the misalignments and the smoothness of the signal $\boldsymbol{h}$.

More generally, if $\text{SNR} \leq \frac{\sqrt{c}}{\gamma}$ for any $\boldsymbol{\tau} \in \{0, \cdots, d_{\max}\}^n$, then the first part of Theorem 4.1 asserts that the MisPCA objective in (4) is almost surely uninformative:

$$\lambda_1(\boldsymbol{S}(\boldsymbol{\tau})) \overset{a.s.}{\to} (1+\sqrt{c})^2 \quad \text{as } n \to \infty,$$

and hence there is little hope for recovering $\boldsymbol{d}$ and $\boldsymbol{h}$.

5. EXPERIMENTS

In this section, we present numerical results that demonstrate the benefit of using the A-MisPCA algorithm described in Section III.

5.1. Numerical comparison of MisPCA Algorithms

We compare the PCA and A-MisPCA approximations described in Section II. As a benchmark, we compute the Oracle-PCA, which assumes knowledge of $\boldsymbol{d}$ and consists of performing PCA on $\boldsymbol{S}(\boldsymbol{d})$.

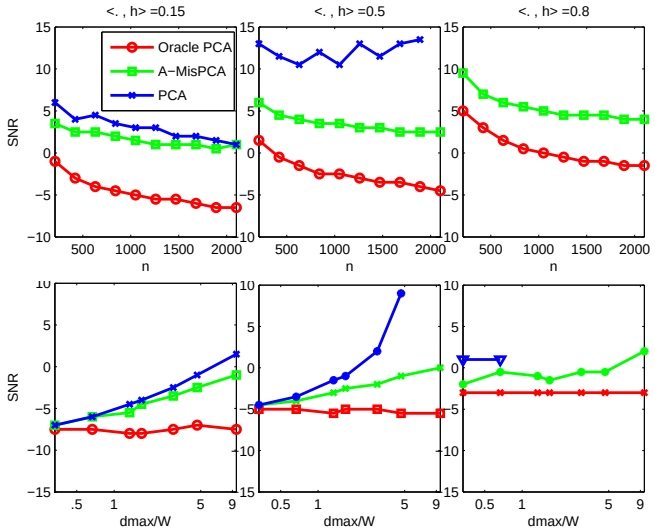

Fig. 2. Estimated SNR levels needed for each of the algorithms to attain a level of fidelity ρ, defined as $\left|\langle \boldsymbol{h}^{\text{algo}}, \boldsymbol{h}\rangle\right| \geq \rho$, for $\rho = .15$ and $\rho = .7$, as a function of the number of samples n, and as a function of the ratio $\frac{d_{\max}}{W}$, where $d_{\max}$ is the maximum delay and W is the time width of the rectangular signal $\boldsymbol{h}$. Since PCA is biased, it fails to attain the fidelity level in several regimes.

In our experiments, we estimate the minimum SNR needed for each algorithm to attain a certain level of fidelity with respect to the generative $\boldsymbol{h}$, here a rectangular signal of width W. The top plots of Figure 2 show the results as a function of the number of samples n with $\frac{d_{\max}}{W} = 5$. The bottom plots of the same figure show the results as a function of $\frac{d_{\max}}{W}$ for fixed $n = 100$. These results demonstrate the advantage of A-MisPCA over PCA in almost every regime. Only when $\frac{d_{\max}}{W} \leq 1$ does PCA fare better than A-misPCA. In that regime, the misalignments are small compared to the width of the rectangular signal and hence affect little the PCA estimate.

5.2. Application to longitudinal gene expression data clustering

In this section we apply our methodology to the study of an influenza challenge study which is part of the (DARPA) Predicting Health and Disease program [7]. This dataset consists of a collection of 272 microarray samples of dimension 12023 genes obtained from 17 individuals. All of these subjects were inoculated with influenza A H3N2Wisconsin and $n = 16$ blood samples were extracted before and after inoculation at prespecified time points. Finally, the clinicians on the team established which of these subjects developed symptoms, based on a standardized symptom scoring method. In previous work, we showed that the trajectories of the gene expression values for different subjects are misaligned with respect to one another [4].

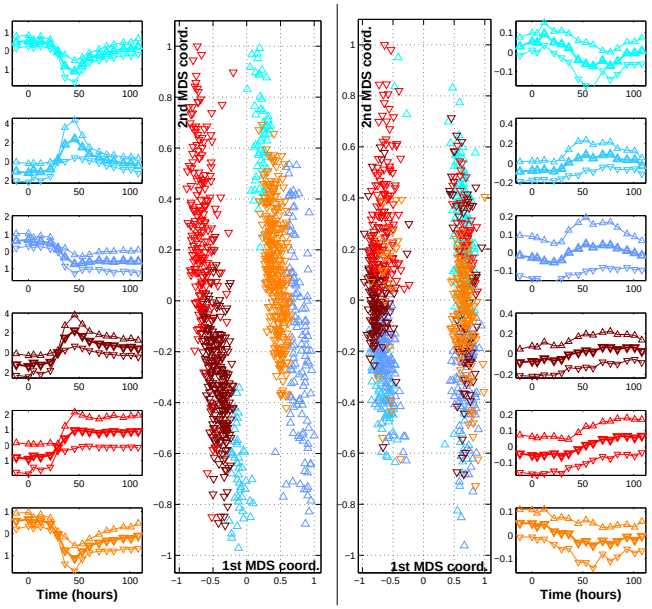

Fig. 3. Hierarchical Clustering results obtained after MisPCA and PCA-based dimensionality reduction. The leftmost and the right most panels show the centroids ($+/-$ standard deviations) after MisPCA and PCA, respectively. The middle panels correspond to a 2-dimensional embedding of the data projected on the MisPC's (left) and the PC's (right).

An important problem in the analysis of temporal gene expression data is that of performing *temporal clustering*, which consists in identifying groups of genes with similar temporal pattern. These genes are likely to be part of a biological pathway and their temporal responses relate to the mechanistics of the process under study. In this section, we use A-MisPCA as a dimensionality reduction tool prior to clustering, and we show its advantage with respect to dimensionality reduction using standard PCA. For this purpose, we compute the first k Misaligned Principal Components (MisPC's) using a deflation heuristic. At each step, we compute the A-MisPCA estimate on the residual obtained after projecting the data to the previously obtained MisPC's. The number k of Principal Components ($k = 4$) is chosen as to minimize the cross validation error, using the cross-validation procedure described in [4]. We apply the same methodology to obtain a rank-4 PCA decomposition. As is common in gene-expression data analysis, we apply an Analysis-of-Variance pre-processing step to select 1000 genes exhibiting high temporal variability. The clustering results, obtained with a hierarchical clustering algorithm[1], are shown in Figure 3. The MisPCA-based centroids, shown on the leftmost panel, have on average 30% less variance that those obtained using PCA. The second and the

[1]The hierarchical clustering algorithm is used with standardized Euclidean distance and complete linkage. Different choices of the number of clusters were explored and 6 was shown to give the most interpretable results.

third pannel show a 2-dimensional embedding, computed using Multidimensional Scaling (MDS), of the projection of the data on the MisPC's and the Principal Components (PC's). It is clear that the clusters corresponding to up-regulated genes (low-to-high variation) are better separated from the down-regulated ones (high-to-low variation) in the MisPCA-based projections.

6. CONCLUSIONS

We have introduced a new method of PCA that compensates for potential circular shifts in the observed data. We have proposed an approximate algorithm to solve the Misaligned PCA problem and have shown its advantage over other approximations. Our methodology can be used to enhance the clustering of misaligned data to obtain centroids that capture more definite latent temporal features from gene expression time series.

7. REFERENCES

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[2] R. W. Picard and T. Kabir, "Finding similar patterns in large image databases," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1993, vol. 5, pp. 161 –164.

[3] E. Moulines, P. Duhamel, J.F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *Signal Processing, IEEE Transactions on*, vol. 43, no. 2, pp. 516–525, 2002.

[4] A. Tibau-Puig, A. Wiesel, A.K. Zaas, C.W. Woods, G.S. Ginsburg, G. Fleury, and A.O. Hero, "Order-preserving factor analysis: Application to longitudinal gene expression," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4447 –4458, 2011.

[5] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," *Statistica Sinica*, vol. 17, no. 4, pp. 1617, 2007.

[6] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 227, no. 1, pp. 494 – 521, 2011.

[7] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Øien, M. T. McClain, J. B. Varkey, B. Nicholson, L. Carin, S Kingsmore, C. W. Woods, G. S. Ginsburg, and A. O. Hero, III, "Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection," *PLoS Genetics*, vol. 7, 2011.

A. APPENDIX

A.1. Derivation of $\text{SNR}^{\text{MisPCA}}$

It is easy to verify that if $\lambda^{\text{MisPCA}} < 1$, then $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \text{SNR}\right)$ is monotonically decreasing over $\text{SNR} \geq 0$. Otherwise, it has a positive stationary point at:

$$\text{SNR}^{\circ} = \lambda^{\text{MisPCA}} - 1.$$

The second derivative of $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \text{SNR}\right)$ with respect to SNR is negative at SNR°, hence SNR° is at least a local maxima. It is easy to check that $l\left(\boldsymbol{h}^{\text{MisPCA}}, \boldsymbol{d}^{\text{MisPCA}}, \text{SNR}\right)$ is strictly increasing over $0 \leq \text{SNR} < \text{SNR}^{\circ}$ and strictly decreasing over $\text{SNR}^{\circ} > \text{SNR}$, thus the local maxima is also a global maxima. This finalizes the proof of (5).

A.2. Proof of Theorem 4.1

The eigenvalue decomposition of $\boldsymbol{\Sigma}(\boldsymbol{\tau})$ is denoted by $\boldsymbol{Q}_{\Sigma}\boldsymbol{\Delta}_{\Sigma}\boldsymbol{Q}_{\Sigma}^{T}$, where $\boldsymbol{Q}_{\Sigma}$ is a unitary matrix containing its eigenvectors and $\boldsymbol{\Delta}_{\Sigma}$ is a diagonal matrix containing its eigenvalues:

$$[\boldsymbol{\Delta}_{\Sigma}]_{i,i} = \begin{cases} \text{SNR}\,\lambda_i\left(\boldsymbol{H}\text{diag}\,\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right)\boldsymbol{H}^{T}\right)+1 & 1 \leq i \leq r \\ 1 & r < i \leq p \end{cases}$$

where $r = \text{rank}\left(\boldsymbol{H}\text{diag}\,\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right)\boldsymbol{H}^{T}\right) \leq 2d_{\max}$. A well-known property of the eigenvalues of Grammian matrices allows us to conclude that $\lambda_i\left(\boldsymbol{H}\text{diag}\,\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right)\boldsymbol{H}^{T}\right)$ is equal to:

$$\lambda_i\left(\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right)\right)^{\frac{1}{2}}\boldsymbol{R}_h\text{diag}\left(\boldsymbol{s}\left(\boldsymbol{d}_{-p}\boldsymbol{\tau}\right)\right)^{\frac{1}{2}}\right)$$

In addition, using properties of the Gaussian distribution, we can write:

$$\boldsymbol{S}(\boldsymbol{\tau}) = \boldsymbol{Q}_{\Sigma}\tilde{\boldsymbol{S}}\boldsymbol{Q}_{\Sigma}^{T}.$$

where $\tilde{\boldsymbol{S}} = \frac{\boldsymbol{Z}\boldsymbol{Z}^{T}}{n}$ and each column of the $p \times n$ matrix $\boldsymbol{Z}$ follows a zero-mean multivariate Gaussian distribution with covariance $\boldsymbol{\Delta}_{\Sigma}$. The result for $\lambda_1\left(\boldsymbol{S}(\boldsymbol{\tau})\right)$ follows from observing that $\lambda_1\left(\boldsymbol{S}(\boldsymbol{\tau})\right) = \lambda_1\left(\tilde{\boldsymbol{S}}\right)$ and applying Theorems 1 and 2 from [5]. The result concerning $\boldsymbol{v}_1\left(\boldsymbol{S}(\boldsymbol{\tau})\right)$ follows from observing that:

$$\begin{aligned} \left\langle \boldsymbol{v}_1\left(\boldsymbol{S}(\boldsymbol{\tau})\right), \boldsymbol{v}_1\left(\boldsymbol{\Sigma}(\boldsymbol{\tau})\right)\right\rangle &= \left\langle \boldsymbol{Q}_{\Sigma}\boldsymbol{v}_1\left(\tilde{\boldsymbol{S}}\right), \boldsymbol{v}_1\left(\boldsymbol{\Sigma}(\boldsymbol{\tau})\right)\right\rangle \\ &= \left\langle \boldsymbol{v}_1\left(\tilde{\boldsymbol{S}}\right), \boldsymbol{e}_1\right\rangle, \end{aligned}$$

where $\boldsymbol{e}_1$ denotes the vector of all zeros except for a 1 in the first coordinate, and applying Theorem 4 from [5]. See [6] for an alternative derivation and insight into the origin of the phase transition.