PROBABILISTIC PCA FOR HETEROSCEDASTIC DATA

David Hong, Laura Balzano, and Jeffrey A. Fessler

Department of EECS, University of Michigan, Ann Arbor, Michigan, USA

ABSTRACT

Principal Component Analysis (PCA) is a standard dimensionality reduction technique, but it treats all samples uniformly, making it suboptimal for heterogeneous data that are increasingly common in modern settings. This paper proposes a PCA variant for samples with heterogeneous noise levels, i.e., heteroscedastic noise, that naturally arise when some of the data come from higher quality sources than others. The technique handles heteroscedasticity by incorporating it in the statistical model of a probabilistic PCA. The resulting optimization problem is an interesting nonconvex problem related to but not seemingly solved by singular value decomposition, and this paper derives an expectation maximization (EM) algorithm. Numerical experiments illustrate the benefits of using the proposed method to combine samples with heteroscedastic noise in a single analysis, as well as benefits of careful initialization for the EM algorithm.

***Index Terms*—** Principal component analysis, heterogeneous data, maximum likelihood estimation, latent factors.

1. INTRODUCTION

Principal Component Analysis (PCA) is a ubiquitous method for unsupervised dimensionality reduction of high-dimensional data [1]. This paper proposes a PCA variant based on the probabilistic PCA framework [2] that extends PCA to handle modern data that are increasingly heterogeneous and messy. In particular, we consider samples with heteroscedastic noise, that is, data where some samples are noisier than others. Such data arises naturally when samples are obtained under varied conditions. For example, spectrophotometric data in [3] are averaged over increasingly long windows of time; samples from shorter windows are nosier. Similarly, atmospheric noise in the astronomical data of [4] varies across nights. Conventional PCA can perform poorly on such data since it treats all samples uniformly, as analyzed in [5].

A common approach to accounting for heteroscedastic noise is to use a weighted PCA [1, Section 14.2.1] that gives cleaner samples more weight. In particular, one often weights samples by inverse noise variance, i.e., samples with half as much noise get twice as much weight. This choice whitens the noise and can be interpreted as a maximum likelihood weighting [6], but as analyzed in [7], is not aggressive enough when the signal-to-noise ratio is small. Optimal weights are derived in [7] that lie between inverse noise variance weights and their squares. However, optimal weights differ depending on which principal component one wants to recover, making them most suitable for recovering individual components.

This paper focuses on recovering multiple components simultaneously, and derives a probabilistic PCA (PPCA) approach that estimates latent factors by maximum likelihood estimation of a heteroscedastic factor model (1). In the homoscedastic setting, the maximum likelihood estimate is given by applying PCA with shrinkage [2], and can be computed via singular value decomposition (SVD) of the data. The heteroscedastic setting involves solving a related nonconvex problem but seems not to have a direct SVD solution. Section 3 derives an expectation maximization (EM) algorithm that extends the homoscedastic variants in [2, Appendix B] and [8]. Numerical experiments in Sections 4 to 6 illustrate the benefits of the proposed heteroscedastic PPCA method and its initialization.

2. HETEROSCEDASTIC PROBABILISTIC PCA

As in the PPCA [2] derivation of classical PCA, we model n independent samples $y_1, \dots, y_n \in \mathbb{R}^d$ as

$$y_i = \mathbf{F} z_i + \varepsilon_i, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{d \times k}$ is a deterministic factor matrix to estimate, $z_i \sim \mathcal{N}(0, \mathbf{I}_k)$ are random coefficients, $\varepsilon_i \sim \mathcal{N}(0, \eta_i^2 \mathbf{I}_d)$ are random noise, and η_i^2 is the ith noise variance. Unlike [2], we allow noise variances to vary across samples. Equivalently,

$$y_i \sim \mathcal{N}(0, \mathbf{F}\mathbf{F}' + \eta_i^2 \mathbf{I}_d). \tag{2}$$

PPCA estimates latent factors $\mathbf{F}$ by maximizing the log-likelihood, dropping the $\ln(2\pi)^{-d/2}$ constant:

$$\mathcal{L}(\mathbf{F}) \coloneqq \frac{1}{2} \sum_{i=1}^{n} \Big\{ \ln \det(\mathbf{F}\mathbf{F}' + \eta_i^2 \mathbf{I}_d)^{-1} - y_i' (\mathbf{F}\mathbf{F}' + \eta_i^2 \mathbf{I}_d)^{-1} y_i \Big\}. \tag{3}$$

D. Hong and L. Balzano were supported in part by ARO YIP award W911NF1910027. L. Balzano was also supported in part by NSF CAREER award CCF-1845076. D. Hong, L. Balzano and J. A. Fessler were supported in part by NSF BIGDATA award IIS-1838179.

When the noise is homoscedastic, i.e., $\eta_1^2 = \cdots = \eta_n^2 = \sigma^2$, this nonconvex problem can be solved via eigendecomposition of the sample covariance matrix [2, Section 3.2], but the same is not true in general.

Writing (3) in terms of $\mathbf{U} \in \mathbb{R}^{d\times k}$ and $\theta_1, \ldots, \theta_k$, the left singular vectors and values of $\mathbf{F}$, and simplifying yields

$$\mathcal{L}(\mathbf{F}) = c + \frac{1}{2}\sum_{i=1}^{n}\left\{ y_i'\mathbf{U}\mathbf{W}_i\mathbf{U}'y_i - \sum_{j=1}^{k}\log(\theta_j^2+\eta_i^2)\right\}, \quad (4)$$

where c is constant with respect to $\mathbf{F}$, and each $\mathbf{W}_i$ is a diagonal matrix with entries $(\theta_1^2/\eta_i^2)/(\theta_1^2+\eta_i^2), \ldots, (\theta_k^2/\eta_i^2)/(\theta_k^2+\eta_i^2)$. Maximizing (4) with respect to $\mathbf{U}$ is a generalized weighted PCA problem with weighting matrices $\mathbf{W}_i$ that depend on the latent factor singular values, but unlike weighted PCA, it does not seem to be solved by eigendecomposition.

3. ALGORITHM: EXPECTATION MAXIMIZATION

This section derives an EM algorithm for PPCA with heteroscedastic noise in the style of [2, Appendix B], where the complete data includes samples $y_1, \ldots, y_n$ and coefficients $z_1, \ldots, z_n$. First, write the complete data log-likelihood

$$\mathcal{L}_{\mathrm{c}}(\mathbf{F}) \coloneqq -\sum_{i=1}^{n}\left(\frac{\|y_i - \mathbf{F}z_i\|_2^2}{2\eta_i^2} + \frac{\|z_i\|_2^2}{2}\right), \quad (5)$$

where (5) drops the constants $\ln(2\pi\eta_i)^{-d/2}$ and $\ln(2\pi)^{-k/2}$.

For the E-step, take the expectation of (5) with respect to the conditionally independent distributions (from Bayes' rule)

$$z_i|y_1, \ldots, y_n, \mathbf{F}_t \overset{\text{ind}}{\sim} \mathcal{N}\Big(\mathbf{M}_{t,i}\mathbf{F}_t'y_i, \eta_i^2\mathbf{M}_{t,i}\Big), \quad (6)$$

where $\mathbf{F}_t$ is the current iterate and $\mathbf{M}_{t,i} \coloneqq (\mathbf{F}_t'\mathbf{F}_t + \eta_i^2\mathbf{I}_k)^{-1}$, yielding

$$\begin{aligned}\bar{\mathcal{L}}(\mathbf{F}; \mathbf{F}_t) &\coloneqq \sum_{i=1}^{n}\left[\frac{1}{\eta_i^2}y_i'\mathbf{F}\mathbb{E}z_i - \frac{1}{2\eta_i^2}\operatorname{tr}\{\mathbf{F}'\mathbf{F}\mathbb{E}(z_iz_i')\}\right] \quad (7)\\ &= \sum_{i=1}^{n}\left[\frac{1}{\eta_i^2}y_i'\mathbf{F}\bar{z}_{t,i} - \frac{1}{2\eta_i^2}\operatorname{tr}\{\mathbf{F}'\mathbf{F}(\bar{z}_{t,i}\bar{z}_{t,i}' + \eta_i^2\mathbf{M}_{t,i})\}\right],\end{aligned}$$

where $\bar{z}_{t,i} \coloneqq \mathbf{M}_{t,i}\mathbf{F}_t'y_i$, (7) drops all terms that are constant with respect to $\mathbf{F}$, and the expectations $\mathbb{E}$ are all with respect to $z_1, \ldots, z_n|y_1, \ldots, y_n, \mathbf{F}_t$ as given in (6).

For the M-step, maximize (7) with respect to $\mathbf{F}$, e.g., by completing the square, to obtain the next EM iterate

$$\mathbf{F}_{t+1} = \mathbf{T}_t\mathbf{S}_t^{-1}, \quad (8)$$

where

$$\mathbf{T}_t \coloneqq \sum_{i=1}^{n}\frac{1}{\eta_i^2}y_i\bar{z}_{t,i}', \quad \mathbf{S}_t \coloneqq \sum_{i=1}^{n}\frac{1}{\eta_i^2}\bar{z}_{t,i}\bar{z}_{t,i}' + \mathbf{M}_{t,i}. \quad (9)$$

3.1. Grouping samples with a common noise variance

Samples often share noise variances, e.g., when they come from the same source or sensor, introducing useful structure into the update (8). Suppose that $\eta_i^2 \in \{\sigma_1^2, \ldots, \sigma_L^2\}$ for all $i \in \{1, \ldots, n\}$, where n_1 of the samples have noise variance $\eta_i^2 = \sigma_1^2$, n_2 have noise variance $\eta_i^2 = \sigma_2^2$, and so on. That is, each sample has one of L distinct noise variances $\sigma_1^2, \ldots, \sigma_L^2$.

Collecting samples in (9) that share noise variance yields

$$\mathbf{T}_t = \sum_{\ell=1}^{L}\frac{1}{\sigma_\ell^2}\mathbf{Y}_\ell\bar{\mathbf{Z}}_{t,\ell}', \quad \mathbf{S}_t = \sum_{\ell=1}^{L}\frac{1}{\sigma_\ell^2}\bar{\mathbf{Z}}_{t,\ell}\bar{\mathbf{Z}}_{t,\ell}' + n_\ell\mathbf{N}_{t,\ell}, \quad (10)$$

where $\mathbf{Y}_\ell \coloneqq (y_i : \eta_i^2 = \sigma_\ell^2) \in \mathbb{R}^{d\times n_\ell}$ is a matrix whose n_ℓ columns are the samples with noise variance σ_ℓ^2,

$$\bar{\mathbf{Z}}_{t,\ell} \coloneqq (\bar{z}_{t,i} : \eta_i^2 = \sigma_\ell^2) = \mathbf{N}_{t,\ell}\mathbf{F}_t'\mathbf{Y}_\ell \in \mathbb{R}^{k\times n_\ell}, \quad (11)$$

and $\mathbf{N}_{t,\ell} \coloneqq (\mathbf{F}_t'\mathbf{F}_t + \sigma_\ell^2\mathbf{I}_k)^{-1}$. Note that

$$\mathbf{Y}_\ell\bar{\mathbf{Z}}_{t,\ell}' = \mathbf{\Lambda}_\ell\mathbf{F}_t\mathbf{N}_{t,\ell}, \quad \bar{\mathbf{Z}}_{t,\ell}\bar{\mathbf{Z}}_{t,\ell}' = \mathbf{N}_{t,\ell}\mathbf{F}_t'\mathbf{\Lambda}_\ell\mathbf{F}_t\mathbf{N}_{t,\ell}, \quad (12)$$

where $\mathbf{\Lambda}_\ell \coloneqq \mathbf{Y}_\ell\mathbf{Y}_\ell' \in \mathbb{R}^{d\times d}$ only needs to be computed once during initialization. Thus, using (12) to compute the update (8) can be more efficient, especially when n_ℓ is large.

3.2. Initialization by homoscedastic PPCA

We initialize the latent factors $\mathbf{F}$ using the homoscedastic PPCA solution [2, Section 3.2]:

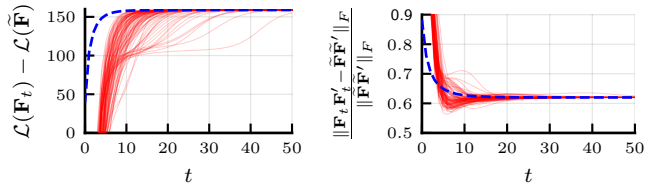
$$\mathbf{F}_0 \coloneqq \mathbf{V}\operatorname{diag}\left(\sqrt{\lambda_1 - \bar{\lambda}}, \ldots, \sqrt{\lambda_k - \bar{\lambda}}\right), \quad (13)$$

where the columns of $\mathbf{V} \in \mathbb{R}^{d\times k}$ are the k principal eigenvectors of the sample correlation matrix $(y_1y_1' + \cdots + y_ny_n')/n$, $\lambda_1, \ldots, \lambda_k$ are the corresponding k principal eigenvalues, and $\bar{\lambda}$ is the average of the remaining $d-k$ eigenvalues.

A benefit of this initialization is that it guarantees all iterates will have likelihood at least as large as homoscedastic PPCA since the EM algorithm never decreases the likelihood. Furthermore, the homoscedastic PPCA solution is likely close to the heteroscedastic PPCA solution when samples have relatively low noise or relatively homogeneous noise variance. Section 4 describes numerical experiments assessing the performance of this initialization.

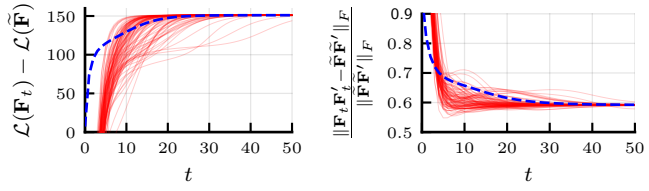
4. INITIALIZATION EXPERIMENTS

This section compares (13) with random initializations. We generate $n = 10^3$ samples in $d = 10^2$ dimensions from $k = 3$ factors according to the model (1). True latent factors are generated as $\widetilde{\mathbf{F}} = \widetilde{\mathbf{U}}\operatorname{diag}(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3)$, where $\tilde{\theta}_1^2 = 4$, $\tilde{\theta}_2^2 = 2$, $\tilde{\theta}_3^2 = 1$, and $\widetilde{\mathbf{U}} = (\tilde{u}_1, \ldots, \tilde{u}_k) \in \mathbb{R}^{d\times k}$ is uniformly drawn from the set of $d \times k$ matrices with orthonormal columns. The data have heteroscedastic noise: $n_1 = 200$ samples have



(a) Objective function (b) Normalized estimation error

Fig. 1: Example data realization where homoscedastic PPCA initialization (blue dashed curves) generally converged faster than random initialization (red solid curves).



(a) Objective function (b) Normalized estimation error

Fig. 2: Example data realization where homoscedastic PPCA initialization (blue dashed curves) converged slower than some random initializations (red solid curves).

noise variance $\sigma_1^2 = 1$ and $n_2 = 800$ samples have noise variance $\sigma_2^2 = 4$. Each random initialization $\mathbf{F}_0$ is generated as $\mathbf{Q}\,\mathrm{diag}(\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3)$, where $\mathbf{Q}$ is drawn uniformly at random from the set of $d \times k$ matrices with orthonormal columns. Namely, random initializations have the true latent covariance eigenvalues (unknown in practice) with random eigenvectors.

Fig. 1 shows the objective function (3) and normalized estimation error[1] $\|\mathbf{F}_t\mathbf{F}_t' - \widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'\|_F / \|\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'\|_F$ over iterations t when initialized by homoscedastic PPCA (13) and by 100 random initializations for an example data realization. Random initializations generally start at worse objective function values and estimation errors, and take longer to converge. Some intermediate iterates from random initializations pass closer to the true latent factors $\widetilde{\mathbf{F}}$, but they are not minima and EM moves away from them. Fig. 2 shows a data realization (generated in the same way) where some randomly initialized runs converge faster. It is otherwise similar to Fig. 1, however, and initialization with homoscedastic PPCA still converges as fast as many of the randomly initialized runs.

These examples were fairly representative in our testing. Homoscedastic PPCA is generally a better initialization, and the resulting iterates typically converge at a rate competitive to those from random initializations. Interestingly, EM seems to always converge to a global minima even though the objective (3) is nonconvex, suggesting that (3) may have special structure similar to homoscedastic PPCA [2, Sections A.2-A.3]. Characterizing it is an interesting area of future work.

[1]Since the objective (3) depends on $\mathbf{F}$ only through $\mathbf{F}\mathbf{F}'$ we can only hope to recover $\widetilde{\mathbf{F}}$ up to right multiplication by a $k \times k$ orthogonal matrix.

5. COMPARISON WITH HOMOSCEDASTIC PPCA

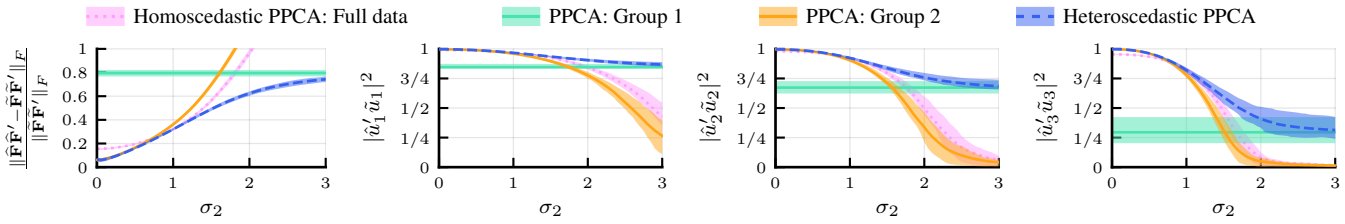
This section illustrates the benefit of accounting for heteroscedasticity in PPCA. We consider $n_1 = 200$ samples (group 1) with noise standard deviation $\sigma_1 = 1$ and $n_2 = 800$ samples (group 2) with noise standard deviation σ_2. To get a range of heteroscedastic settings, we sweep σ_2 from 0 to 3. Data is otherwise generated as in Section 4. For each σ_2 in the sweep, we generate 100 data realizations and evaluate recovery of the underlying true latent factors $\widetilde{\mathbf{F}}$ by estimates $\widehat{\mathbf{F}}$ obtained from heteroscedastic PPCA on the full data and from homoscedastic PPCA on: a) the full data, b) only group 1, and c) only group 2. The heteroscedastic PPCA is initialized as in Section 3.2 and run for 1000 iterations to obtain $\widehat{\mathbf{F}} = \mathbf{F}_{1000}$.

Fig. 3a shows mean normalized estimation errors $\|\widehat{\mathbf{F}}\widehat{\mathbf{F}}' - \widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'\|_F / \|\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'\|_F$ as curves with ribbons for the associated interquartile intervals. Likewise, Figs. 3b to 3d show mean and interquartile component recoveries $|\hat{u}_1'\tilde{u}_1|^2, \ldots, |\hat{u}_3'\tilde{u}_3|^2$, where $\tilde{u}_1, \ldots, \tilde{u}_3 \in \mathbb{R}^d$ and $\hat{u}_1, \ldots, \hat{u}_3 \in \mathbb{R}^d$ are the principal eigenvectors of $\widetilde{\mathbf{F}}\widetilde{\mathbf{F}}'$ and $\widehat{\mathbf{F}}\widehat{\mathbf{F}}'$, respectively. Lower is better for estimation error, and higher is better for component recoveries. We first compare the homoscedastic PPCA's.

When σ_2 is sufficiently smaller than $\sigma_1 = 1$, homoscedastic PPCA performs best when applied to only group 2 data since doing so excludes noisier data. Using the full data has the advantage of incorporating more samples, but including noisier group 1 samples is a bigger downside in this regime. Using only group 1 data performs worst since this dataset is both smallest and noisiest. As $\sigma_2 \to 0$, group 2 data become noiseless, and using only this data yields perfect component recoveries. On the other hand, homoscedastic PPCA on the full data and on only group 1 still incorporate the noisy group 1 samples and do not achieve perfect recovery.

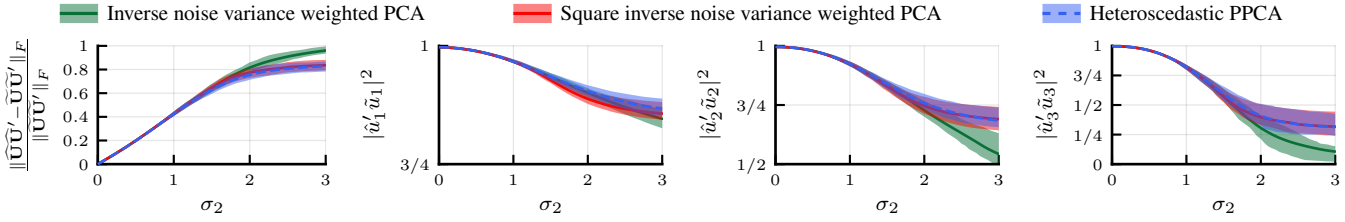
As σ_2 increases, homoscedastic PPCA performance using either the full data or only group 2 degrades since the samples they use become noisier. Moreover, the benefit of using only group 2 samples diminishes, and using the full data becomes better than using only group 2. The performance of using only group 1 remains the same; its performance depends on the size n_1 and noise level σ_1 of group 1, not σ_2. When $\sigma_2 = \sigma_1$ and noise is homoscedastic, using the largest full dataset performs best, followed by using only group 2 then using only group 1. When $\sigma_2 > \sigma_1$, excluding the cleaner group 1 data gives no benefit, but using only group 1 initially remains worse than using only group 2 because it has $n_1/n_2 = 1/4$ as many samples. As σ_2 continues to grow, full data and group 2 homoscedastic PPCA performances eventually degrade beyond using only group 1, and using only group 1 yields the best homoscedastic PPCA performance.

Heteroscedastic PPCA (dashed blue) uses the full data but accounts for their heteroscedastic noise. When σ_2 is small, it seems to match the best performing homoscedastic PPCA that uses only group 2. As σ_2 increases, its performance degrades like full data or group 2 homoscedastic PPCA, but it does so



(a) Normalized factor estimation error (b) Component 1 recovery (c) Component 2 recovery (d) Component 3 recovery

Fig. 3: Comparison with homoscedastic PPCA.



(a) Normalized subspace error (b) Component 1 recovery (c) Component 2 recovery (d) Component 3 recovery

Fig. 4: Comparison with weighted PCA.

more slowly, matching then outperforming both. Eventually, its performance approaches homoscedastic PPCA using only group 1. Throughout, heteroscedastic PPCA either matches or outperforms the three homoscedastic PPCA approaches. It appropriately balances the two groups of data, and seems to always benefit from including all samples even if very noisy.

6. COMPARISON WITH WEIGHTED PCA

This section repeats the experiments of Section 5 but with weighted PCA [1, Section 14.2.1]. Weighted PCA estimates principal components $\widehat{\mathbf{U}} = (\hat{u}_1, \ldots, \hat{u}_k) \in \mathbb{R}^{d\times k}$ as the principal eigenvectors of the weighted sample covariance matrix

$$\frac{1}{n}\sum_{i=1}^{n} \omega_i^2 y_i y_i' = \frac{1}{n}\sum_{\ell=1}^{L} w_\ell^2 \sum_{i:\eta_i=\sigma_\ell} y_i y_i' = \frac{1}{n}\sum_{\ell=1}^{L} w_\ell^2 \mathbf{\Lambda}_\ell,$$

where we give group 1 samples a weight of w_1^2 and group 2 samples a weight of w_2^2. We consider inverse noise variance weights $w_\ell^2 = 1/\sigma_\ell^2$ that effectively rescale samples to make the noise homoscedastic, and square inverse noise variance weights $w_\ell^2 = 1/\sigma_\ell^4$ that can be more effective in low signal-to-noise ratio regimes since they more aggressively downweight noisier samples [7].

Fig. 4 shows the mean and interquartile principal subspace estimation errors $\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}' - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}'\|_F / \|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}'\|_F$ and component recoveries $|\hat{u}_1'\tilde{u}_1|^2, \ldots, |\hat{u}_3'\tilde{u}_3|^2$. When σ_2 is small, both weighted PCA approaches perform similarly to heteroscedastic PPCA; they account for heteroscedasticity and benefit from using the full data. As σ_2 grows, inverse noise variance weighted PCA becomes worse than heteroscedastic PPCA, especially in the weaker components 2 and 3, since it does not downweight noisier group 2 samples enough. Square inverse noise variance weights, on the other hand, remain comparable to heteroscedastic PPCA, with similar subspace estimation error throughout the sweep. Component recoveries are worse for moderate σ_2, but this gap shrinks as σ_2 grows large and both methods likely rely primarily on cleaner group 1 data.

The performance of weighted PCA and heteroscedastic PPCA compared with the homoscedastic PPCA approaches underscores the benefit of accounting for heteroscedasticity. By appropriately combining heterogeneous samples in one analysis, these methods make better use of all available data.

7. CONCLUSION

This paper proposes a PPCA for data with heteroscedastic noise and derives an EM algorithm to compute the factor estimate. Numerical experiments show the benefits of initializing EM with homoscedastic PPCA. Experiments also illustrate how the proposed heteroscedastic PPCA outperforms inverse and square inverse weighted PCA, as well as homoscedastic PPCA applied to all available samples or to only those sharing a noise variance.

Other optimization approaches that utilize the alternative log-likelihood form (4) is an avenue of ongoing work. Orthonormality constraints in the resulting problem suggest approaches based, e.g., on manifold optimization. Another area for future work is characterizing the objective (3) to understand why the EM algorithm seemed to always converge to global minima. Finally, extensions of this approach to incorporate other aspects of messy data, notably missing/unobserved data, is an important avenue for further work.

8. REFERENCES

[1] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 2002.

[2] Michael E. Tipping and Christopher M. Bishop, "Probabilistic Principal Component Analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, Aug. 1999.

[3] Robert N. Cochran and Frederick H. Horne, "Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments," *Analytical Chemistry*, vol. 49, no. 6, pp. 846–853, May 1977.

[4] O. Tamuz, T. Mazeh, and S. Zucker, "Correcting systematic effects in a large set of photometric light curves," *Monthly Notices of the Royal Astronomical Society*, vol. 356, no. 4, pp. 1466–1470, Feb. 2005.

[5] David Hong, Laura Balzano, and Jeffrey A. Fessler, "Asymptotic performance of PCA for high-dimensional heteroscedastic data," *Journal of Multivariate Analysis*, vol. 167, pp. 435–452, Sept. 2018.

[6] Gale Young, "Maximum likelihood estimation and factor analysis," *Psychometrika*, vol. 6, no. 1, pp. 49–53, Feb. 1941.

[7] David Hong, Jeffrey A. Fessler, and Laura Balzano, "Optimally Weighted PCA for High-Dimensional Heteroscedastic Data," 2018, In preparation.

[8] Sam T. Roweis, "EM Algorithms for PCA and SPCA," in *Advances in Neural Information Processing Systems 10*, M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds., pp. 626–632. MIT Press, 1998.