Theoretical Analysis of PCA for Heteroscedastic Data

David Hong
Department of EECS
University of Michigan
Ann Arbor, Michigan 48105

Laura Balzano
Department of EECS
University of Michigan
Ann Arbor, Michigan 48105

Jeffrey A. Fessler
Department of EECS
University of Michigan
Ann Arbor, Michigan 48105

I. Introduction

Principal Component Analysis (PCA) is a classical method for estimating a subspace given noisy samples. It is useful in a variety of applications [1], [2], [3] and problems ranging from dimensionality reduction to anomaly detection and the visualization of high dimensional data. Effective use of PCA requires a rigorous understanding of its performance. This work analyzes PCA for samples with heteroscedastic noise, that is, samples that have non-uniform noise variances. In particular, we provide a simple asymptotic prediction of the recovery of a low-dimensional subspace basis from noisy heteroscedastic samples. The prediction enables: (a) easy and efficient calculation of the asymptotic performance, (b) reasoning about the asymptotic performance (with heteroscedasticity such as outliers), and (c) a deeper understanding that PCA has best performance when the noise is homoscedastic (all points share the same noise level).

II. Model for Heteroscedastic Data

We model n heteroscedastic samples $y_1, \dots, y_n \in \mathbb{R}^d$ as

$$\mathbf{Y} = [y_1 \ \cdots \ y_n] = \tilde{\mathbf{U}}\tilde{\mathbf{\Theta}}\tilde{\mathbf{Z}}^{\mathsf{H}} + [\eta_1\varepsilon_1 \ \cdots \ \eta_n\varepsilon_n] \tag{1}$$

where

- $\tilde{\mathbf{U}} = [\tilde{u}_1 \ \cdots \ \tilde{u}_k] \in \mathbb{C}^{d\times k}$ is the true subspace basis,
- $\tilde{\mathbf{\Theta}} = \mathrm{diag}(\tilde{\theta}_1, \dots, \tilde{\theta}_k) \in \mathbb{R}_+^{k\times k}$ are subspace amplitudes,
- $\tilde{\mathbf{Z}} = [\tilde{z}^{(1)} \ \cdots \ \tilde{z}^{(k)}] \in \mathbb{C}^{n\times k}$ has IID random coefficient entries with mean 0 and variance 1,
- $\varepsilon_i \in \mathbb{C}^d$ are IID noise vectors with IID entries with mean 0, variance 1 and bounded fourth moment,
- n_1 samples have noise std. dev. $\eta_i = \sigma_1$, n_2 have $\eta_i = \sigma_2$, ...

III. Main Result

The following theorem (given for a single component model in [4]) provides part of our main result: a simple expression for the asymptotic performance of the i-th principal component $\hat{u}_i$. The paper [5] contains the remainder: simple expressions for asymptotic performance of the i-th score vector $\hat{z}^{(i)}$ and i-th PCA amplitude $\hat{\theta}_i$.

Theorem 1: Suppose the sample-to-dimension ratio $n/d \to c > 0$ and the noise variance proportions $n_\ell/n \to p_\ell$ for $\ell = 1, \dots, L$ as $n, d \to \infty$. Then the i-th principal component $\hat{u}_i$ is such that [5]

$$\begin{aligned} |\langle \hat{u}_i, \mathrm{Span}\{\tilde{u}_j : \tilde{\theta}_j = \tilde{\theta}_i\}\rangle|^2 &\xrightarrow{a.s.} \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)} \\ |\langle \hat{u}_i, \mathrm{Span}\{\tilde{u}_j : \tilde{\theta}_j \neq \tilde{\theta}_i\}\rangle|^2 &\xrightarrow{a.s.} 0, \end{aligned} \tag{2}$$

if $A(\beta_i) > 0$ where β_i is the largest real root of $B_i(x)$ and

$$A(x) := 1 - c\sum_{\ell=1}^{L} \frac{p_\ell \sigma_\ell^4}{(x-\sigma_\ell^2)^2}, \quad B_i(x) := 1 - c\tilde{\theta}_i^2 \sum_{\ell=1}^{L} \frac{p_\ell}{x-\sigma_\ell^2}. \tag{3}$$

We conjecture that when $A(\beta_i) \leq 0$, the above limit (2) is zero.

Figure 1 shows simulations displaying these results for $L = 2$. Figure 2 illustrates the behavior of the largest real root of $B_i(x)$ in relation to the noise variances.

IV. Dependence on Parameters

The main result (Theorem 1) enables us to reason about how PCA depends on the various parameters.

Dependence on sample-to-dimension ratio and amplitudes. Figure 3 shows asymptotic recovery (2) as we sweep over c and $\tilde{\theta}_i$ where $p_1 = 80\%$ of samples have noise variance $\sigma_1^2 = 0.8$ and $p_2 = 20\%$ of samples have noise variance $\sigma_2^2 = 1.8$. As expected, recovery generally improves with increasing c (more samples per dimension) and increasing $\tilde{\theta}_i$ (stronger signal). It also has a similar shape as the homoscedastic case (analyzed in [6]) but with more samples generally needed to achieve the same recovery.

Dependence on sample proportions. The green curves in Figure 1 show asymptotic recovery (2) as $p_2 \in (0, 1)$ varies with $p_1 = 1 - p_2$, $c = 10$, $\tilde{\theta} = (1, 0.8)$, $\sigma_1^2 = 0.1$, and $\sigma_2^2 = 3.25$. Recovery generally degrades with increasing p_2; having more low noise samples is better.

Dependence on noise variances. Figure 4 shows asymptotic recovery (2) as σ_1^2 and σ_2^2 vary with $c = 10$, $p = (0.7, 0.3)$ and $\tilde{\theta}_i = 1$. Recovery generally improves with decreasing noise variances. Interestingly, for a fixed average noise variance (i.e., along the dashed line in the plot), the best recovery occurs when $\sigma_1^2 = \sigma_2^2$. It turns out this is true in general, a fact which we next state.

V. An upper bound on performance

The following theorem provides an upper bound on the asymptotic subspace recovery (2). The bound is a function of the average noise variance and is attained when the noise is homoscedastic.

Theorem 2: The asymptotic subspace recovery (2) is bounded as:

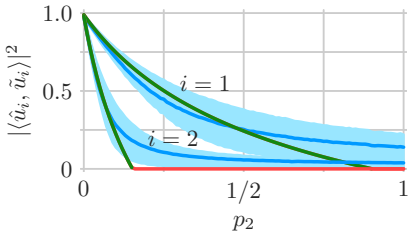
$$\frac{A(\beta_i)}{\beta_i B_i'(\beta_i)} \leq \max\left(0, \frac{c - \bar{\sigma}^4/\tilde{\theta}_i^4}{c + \bar{\sigma}^2/\tilde{\theta}_i^2}\right) \tag{4}$$

where $\bar{\sigma}^2 := \sum_{\ell=1}^{L} p_\ell \sigma_\ell^2$ is the average noise variance. The bound is attained when $\sigma_1^2 = \cdots = \sigma_L^2$ (i.e., the noise is homoscedastic).

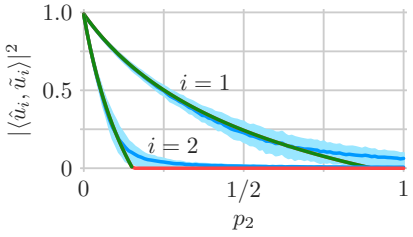
Remark 1: It follows that the asymptotic recovery (2) for a fixed average noise variance is maximized when the noise is homoscedastic. Thus, using average noise variance to predict performance gives an optimistic proxy; actual recovery will be worse than such predictions for heteroscedastic data. Our expression (2) is more accurate.

VI. Conclusion

This work provides a simple expression for the asymptotic recovery of a subspace basis from heteroscedastic samples by PCA. We use the result to gain insights about the performance of PCA as a function of the parameters and find an upper bound that shows the performance for a fixed average noise variance is optimal when the noise is homoscedastic. There are many avenues for future work, including further study of the algebraic structure of the asymptotic recovery (2), analyzing the non-asymptotic recovery, and considering a weighted version of PCA. Preliminary work on weighted PCA suggests that inverse variance weights (i.e., whitening) improve performance but are not optimal; work on this analysis is ongoing.



(a) Results for $d = 10^2$, $n = 10^3$ from 10000 trials.



(b) Results for $d = 10^3$, $n = 10^4$ from 1000 trials.

Fig. 1: Numerical simulation results for $c = 10$, $\tilde{\theta} = (1, 0.8)$, $\sigma_1^2 = 0.1$, and $\sigma_2^2 = 3.25$ where p_2 is swept from 0 to 1 with $p_1 = 1 - p_2$. Simulation mean (blue curve) and interquartile interval (light blue ribbon) shown with the asymptotic recovery (2) (green curve). The region where $A(\beta_i) \leq 0$ is the red horizontal segment with value zero (as conjectured). Going from Figure 1a to Figure 1b, we increase the problem size while keeping the same model parameters. The simulation mean moves closer to the asymptotic prediction and the interquartile range shrinks, indicating concentration to the asymptotic prediction.

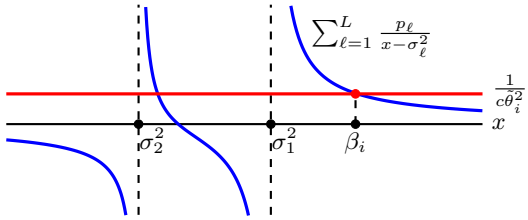


Fig. 2: Location of the largest real root β_i of $B_i(x)$ for two noise variances $\sigma_1^2 = 2$ and $\sigma_2^2 = 0.75$, occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 1$ and the subspace amplitude is $\tilde{\theta}_i = 1$.

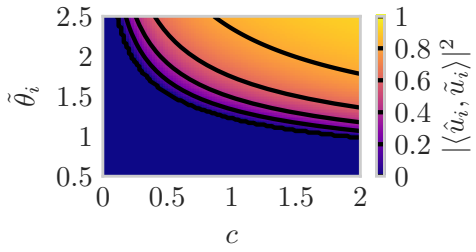


Fig. 3: Asymptotic subspace recovery (2) as a function of sample-to-dimension ratio c and subspace amplitude $\tilde{\theta}_i$ with $p_1 = 80\%$ of samples having noise variance $\sigma_1^2 = 0.8$ and $p_2 = 20\%$ of samples having noise variance $\sigma_2^2 = 1.8$. Contours are overlaid in black and the region where $A(\beta_i) \leq 0$ is shown as zero (as conjectured).

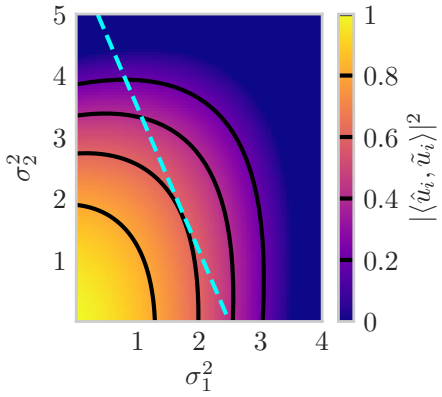


Fig. 4: Asymptotic subspace recovery (2) as a function of noise variances σ_1^2 and σ_2^2 occurring in proportions $p_1 = 70\%$ and $p_2 = 30\%$, where the sample-to-dimension ratio is $c = 10$ and the subspace amplitude is $\tilde{\theta}_i = 1$. Contours are overlaid in black and the region where $A(\beta_i) \leq 0$ is shown as zero (as conjectured). Along the dashed cyan line, the average noise variance is $\bar{\sigma}^2 \approx 1.74$. As observed in Section IV, the best performance occurs when $\sigma_1^2 = \sigma_2^2 = \bar{\sigma}^2$, and this is true in general as discussed in Section V.

References

[1] B. A. Ardekani, J. Kershaw, K. Kashikura, and I. Kanno, "Activation detection in functional MRI using subspace modeling and maximum likelihood estimation," *IEEE Transactions on Medical Imaging*, vol. 18, no. 2, pp. 101–114, 1999.

[2] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM '04, 2004, pp. 219–230.

[3] N. Sharma and K. Saroha, "A novel dimensionality reduction method for cancer dataset using PCA and feature ranking," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, pp. 2261–2264.

[4] D. Hong, L. Balzano, and J. A. Fessler, "Towards a theoretical analysis of pca for heteroscedastic data," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2016.

[5] ——, "Asymptotic Performance of PCA for High-Dimensional Heteroscedastic Data," https://arxiv.org/abs/1703.06610.

[6] I. M. Johnstone and A. Y. Lu, "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 2009.