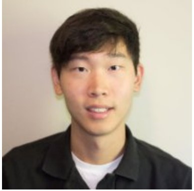
Optimally Weighted PCA for High-Dimensional Heteroscedastic Data

Laura Balzano
work of PhD student David Hong, joint with Jeff Fessler

Department of EECS, University of Michigan

Simons Workshop for Robust High-Dimensional Statistics
1 November 2018

Collaborators



David Hong

Jeff Fessler

Modern data: High-Dimensional & Heteroscedastic

Modern data are often high-dimensional, but we assume some low-dimensional underlying structure exists.

millions of voxels (hundreds per patch)

thousands of detector elements

thousands of locations

---

http://www.medicalnewstoday.com/articles/153201.php
https://www.nasa.gov/multimedia/imagegallery/iotd.html
http://www.livescience.com/27992-portable-pollution-sensors-improve-data-nsf-ria.html

Modern data: High-Dimensional & Heteroscedastic

Modern data are often corrupted by **heteroscedastic** noise.

millions of voxels
(hundreds per patch)

varying radiation levels

thousands of
detector elements

varying atmosphere

thousands of locations

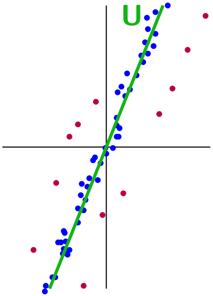
varying sensor quality

http://www.medicalnewstoday.com/articles/153201.php
https://www.nasa.gov/multimedia/imagegallery/iotd.html
http://www.livescience.com/27992-portable-pollution-sensors-improve-data-nsf-ria.html

Modern data: High-Dimensional & Heteroscedastic

Model samples $y_1, \dots, y_n \in \mathbb{C}^d$ as

$$\begin{aligned}\mathbf{Y} &= [y_1 \ \cdots \ y_n] \\ &= \mathbf{U}\boldsymbol{\Theta}\mathbf{Z}^\mathsf{H} + [\eta_1\varepsilon_1 \ \cdots \ \eta_n\varepsilon_n]\end{aligned}$$



Modern data: High-Dimensional & Heteroscedastic

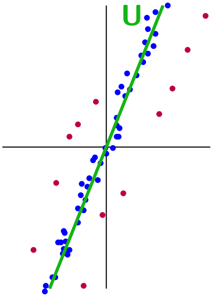
Model samples $y_1, \ldots, y_n \in \mathbb{C}^d$ as

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

components $[u_1 \ \cdots \ u_k]$

IID random scores (mean 0, var. 1)

amplitudes $\mathrm{diag}(\theta_1, \ldots, \theta_k)$



Modern data: High-Dimensional & Heteroscedastic

Model samples $y_1, \dots, y_n \in \mathbb{C}^d$ as

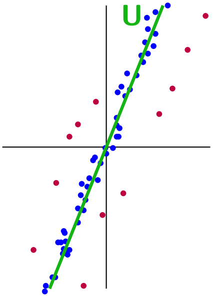
$$\begin{aligned}\mathbf{Y} &= [y_1 \ \cdots \ y_n] \\ &= \mathbf{U\Theta Z}^\mathsf{H} + [\eta_1 \varepsilon_1 \ \cdots \ \eta_n \varepsilon_n]\end{aligned}$$

noise std. dev.

IID noise
(mean 0, var. 1)

n_1 samples have $\eta_j = \sigma_1$
n_2 samples have $\eta_j = \sigma_2$
$\vdots$
n_L samples have $\eta_j = \sigma_L$

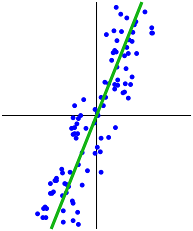
Samples have heteroscedastic noise.



80% $\eta_j^2 = 0.1$,
20% $\eta_j^2 = 1.9$

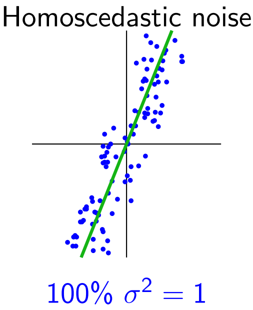
Existing results for (unweighted) PCA

Homoscedastic noise



$100\%\ \sigma^2 = 1$

Existing results for (unweighted) PCA

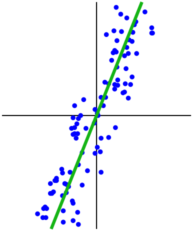


2004(-2009): Johnstone and Lu
2007: Paul
2008: Nadler
2012: Benaych-Georges and Nadakuditi

$$|\langle \hat{u}, u \rangle|^2 \xrightarrow{a.s.} \frac{c - (\sigma/\theta)^4}{c + (\sigma/\theta)^2}$$

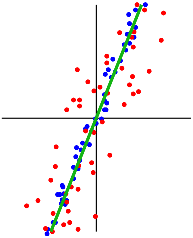
Existing results for (unweighted) PCA

Homoscedastic noise



$100\%\ \sigma^2 = 1$

Heteroscedastic noise



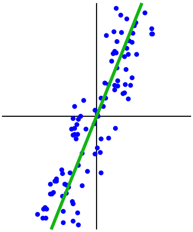
$50\%\ \sigma_1^2 = 0.1$, $50\%\ \sigma_2^2 = 1.9$

2004(-2009): Johnstone and Lu
2007: Paul
2008: Nadler
2012: Benaych-Georges and Nadakuditi

$$|\langle \hat{u}, u \rangle|^2 \xrightarrow{a.s.} \frac{c - (\sigma/\theta)^4}{c + (\sigma/\theta)^2}$$

Existing results for (unweighted) PCA

Homoscedastic noise

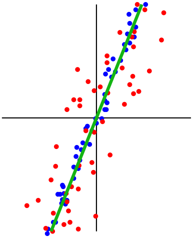


$100\% \ \sigma^2 = 1$

2004(-2009): Johnstone and Lu
2007: Paul
2008: Nadler
2012: Benaych-Georges and Nadakuditi

$$|\langle \hat{u}, u \rangle|^2 \xrightarrow{a.s.} \frac{c - (\sigma/\theta)^4}{c + (\sigma/\theta)^2}$$

Heteroscedastic noise



$50\% \ \sigma_1^2 = 0.1$, $50\% \ \sigma_2^2 = 1.9$

2018: Hong, Balzano and Fessler

$$|\langle \hat{u}, u \rangle|^2 \xrightarrow{a.s.} \frac{A(\beta)}{\beta B'(\beta)}$$

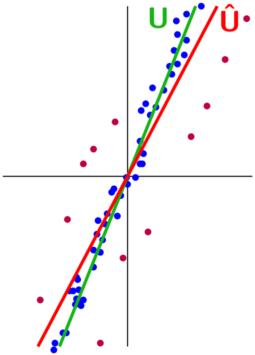
where A, B are rational functions and β is the largest real root of B.

Weighted PCA: Dimensionality Reduction

Weighted PCA finds components $\hat{u}_1, \ldots, \hat{u}_k$ that minimize:

$$\hat{\mathbf{U}} := [\hat{u}_1 \ \cdots \ \hat{u}_k]$$

$$= \operatorname*{argmin}_{\substack{\tilde{\mathbf{U}} \in \mathbb{C}^{d \times k} \\ \tilde{\mathbf{U}}^{\mathsf{H}} \tilde{\mathbf{U}} = \mathbf{I}}} \min_{\tilde{z}_j \in \mathbb{C}^k} \sum_{j=1}^{n} \gamma_j^2 \| y_j - \tilde{\mathbf{U}} \tilde{z}_j \|_2^2$$

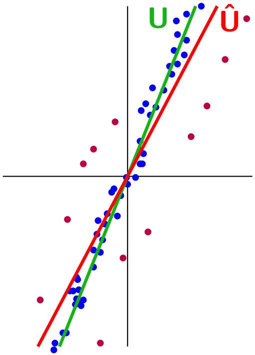


Weighted PCA: Dimensionality Reduction

Weighted PCA finds components $\hat{u}_1, \ldots, \hat{u}_k$ that minimize:

$$\hat{\mathbf{U}} := [\hat{u}_1 \ \cdots \ \hat{u}_k]$$

$$= \operatorname*{argmin}_{\substack{\tilde{\mathbf{U}} \in \mathbb{C}^{d \times k} \\ \tilde{\mathbf{U}}^{\mathsf{H}} \tilde{\mathbf{U}} = \mathbf{I}}} \min_{\tilde{z}_j \in \mathbb{C}^k} \sum_{j=1}^{n} \gamma_j^2 \| y_j - \tilde{\mathbf{U}} \tilde{z}_j \|_2^2$$



The solution is the first k eigenvectors of the weighted sample covariance

$$\frac{1}{n} \sum_{j=1}^{n} \gamma_j^2 y_j y_j^{\mathsf{H}} = \frac{1}{n} \sum_{j=1}^{n} (\gamma_j y_j)(\gamma_j y_j)^{\mathsf{H}}$$

Weighted PCA: Dimensionality Reduction

Weighted PCA finds components $\hat{u}_1, \dots, \hat{u}_k$ that minimize:

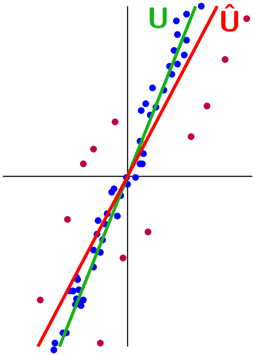
$$\hat{\mathbf{U}} := [\hat{u}_1 \ \cdots \ \hat{u}_k]$$

$$= \underset{\substack{\tilde{\mathbf{U}} \in \mathbb{C}^{d \times k} \\ \tilde{\mathbf{U}}^{\mathsf{H}} \tilde{\mathbf{U}} = \mathbf{I}}}{\operatorname{argmin}} \ \min_{\tilde{z}_j \in \mathbb{C}^k} \ \sum_{j=1}^{n} \gamma_j^2 \| y_j - \tilde{\mathbf{U}} \tilde{z}_j \|_2^2$$

The solution is the first k eigenvectors of the weighted sample covariance

$$\frac{1}{n} \sum_{j=1}^{n} \gamma_j^2 y_j y_j^{\mathsf{H}} = \frac{1}{n} \sum_{j=1}^{n} (\gamma_j y_j)(\gamma_j y_j)^{\mathsf{H}}$$

Weight by group: $\gamma_j^2 \in \{w_1^2 \cdots , w_L^2\}$.



Main Question: How should we weight?

Given data properties...

- samples per dimension n/d
- subspace amplitudes $\theta_1, \ldots, \theta_k$
- noise variances $\sigma_1^2, \ldots, \sigma_L^2$
- proportions $n_1/n, \ldots, n_L/n$.

what choice of weights $w_1, \ldots, w_L$ is best?

Main Question: How should we weight?

Given data properties...

- samples per dimension n/d
- subspace amplitudes $\theta_1, \ldots, \theta_k$
- noise variances $\sigma_1^2, \ldots, \sigma_L^2$
- proportions $n_1/n, \ldots, n_L/n$.

what choice of weights $w_1, \ldots, w_L$ is best?

Common choices:

- binary 0/1, $w_\ell = 0$ or 1 "throw away the noisier data"
- inverse noise var., $w_\ell^2 \propto 1/\sigma_\ell^2$ "whitening"

Main Question: How should we weight?

Given data properties...

- samples per dimension n/d
- subspace amplitudes $\theta_1, \dots, \theta_k$
- noise variances $\sigma_1^2, \dots, \sigma_L^2$
- proportions $n_1/n, \dots, n_L/n$.

what choice of weights $w_1, \dots, w_L$ is best?

Common choices:

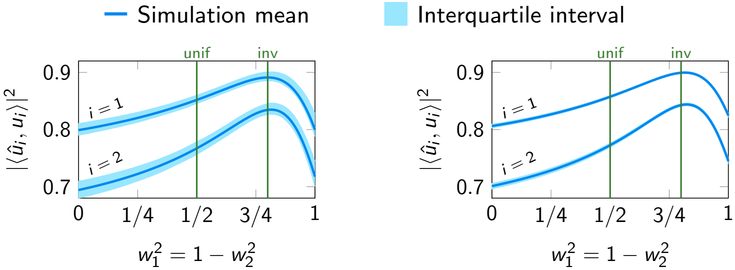
- binary 0/1, $w_\ell = 0$ or 1 "throw away the noisier data"
- inverse noise var., $w_\ell^2 \propto 1/\sigma_\ell^2$ "whitening"
- is there an optimal choice?

An illustrative example with a sneak peek

- $c = 1$ sample per dim
- $\theta_1^2 = 25, \theta_2^2 = 16$ amplitudes
- $p_1 = 20\%$ noise var. $\sigma_1^2 = 1$
- $p_2 = 80\%$ noise var. $\sigma_2^2 = 4$

An illustrative example with a sneak peek

- $c = 1$ sample per dim
- $\theta_1^2 = 25, \theta_2^2 = 16$ amplitudes
- $p_1 = 20\%$ noise var. $\sigma_1^2 = 1$
- $p_2 = 80\%$ noise var. $\sigma_2^2 = 4$



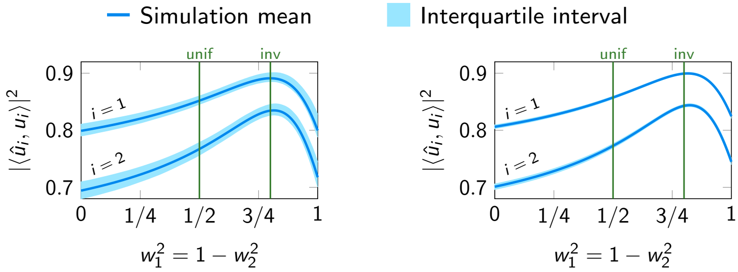
$n = 10^3$ samples
$d = 10^3$ dims.
1000 trials

increasing n, d ⟶ with $n/d = c$

$n = 10^4$ samples
$d = 10^4$ dims.
500 trials

An illustrative example with a sneak peek

- $c = 1$ sample per dim
- $\theta_1^2 = 25, \theta_2^2 = 16$ amplitudes
- $p_1 = 20\%$ noise var. $\sigma_1^2 = 1$
- $p_2 = 80\%$ noise var. $\sigma_2^2 = 4$



Behavior in simulation seems to concentrate in high dimensions...
We will predict that (asymptotic) high dimensional behavior.

Theory: Asymptotic Weighted PCA performance

Theorem (Recovery of components [Hong et al., 2018])

If the sample-to-dimension ratio $n/d \to c > 0$ and the proportions of different quality data $n_\ell/n \to p_\ell$ for $\ell = 1, \ldots, L$ as $n, d \to \infty$, then the i^{th} WPCA component $\hat{u}_i$ converges as

$$\sum_{j:\theta_j=\theta_i} |\langle \hat{u}_i, u_j \rangle|^2 \xrightarrow{a.s.} r_i^{(u)} = \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)}, \quad \sum_{j:\theta_j \neq \theta_i} |\langle \hat{u}_i, u_j \rangle|^2 \xrightarrow{a.s.} 0$$

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

$$A(x) := 1 - c \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^4 \sigma_\ell^4}{(x - w_\ell^2 \sigma_\ell^2)^2}, \quad B_i(x) := 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^2}{x - w_\ell^2 \sigma_\ell^2}.$$

Theory: inverse noise variance

Recall

$$A(x) := 1 - c \sum_{\ell=1}^{L} p_\ell \frac{w_\ell^4 \sigma_\ell^4}{(x - w_\ell^2 \sigma_\ell^2)^2}, \quad B_i(x) := 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^2}{x - w_\ell^2 \sigma_\ell^2}.$$

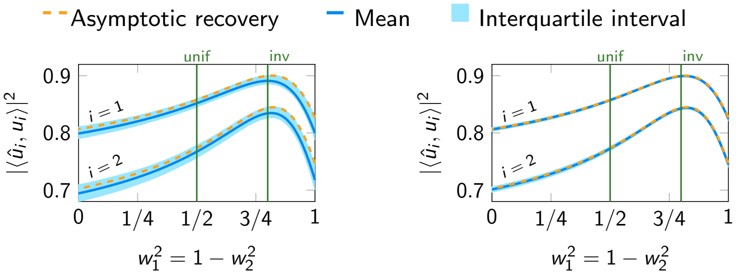
Consider $w_\ell^2 = \bar{\sigma}^2/\sigma_\ell^2$. Then
$A(x) = 1 - \frac{c\bar{\sigma}^4}{(x-\bar{\sigma}^2)^2} \implies \alpha = \bar{\sigma}^2(1+\sqrt{c})$, and
$B_i(x) = 1 - \frac{c\theta_i^2}{x-\bar{\sigma}^2} \implies \beta_i = \bar{\sigma}^2 + c\theta_i^2$.

$$\sum_{j:\theta_j=\theta_i} |\langle \hat{u}_i, u_j \rangle|^2 \xrightarrow{a.s.} r_i^{(u)} = \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)} = \frac{c - \bar{\sigma}^4/\theta_i^4}{c + \bar{\sigma}^2/\theta_i^2}.$$

amplitudes

Asymptotic results compared to finite simulation

- $c = 1$ sample per dim
- $\theta_1^2 = 25, \theta_2^2 = 16$ amplitudes
- $p_1 = 20\%$ noise var. $\sigma_1^2 = 1$
- $p_2 = 80\%$ noise var. $\sigma_2^2 = 4$



$n = 10^3$ samples	increasing n, d	$n = 10^4$ samples
$d = 10^3$ dims.	$\longrightarrow$	$d = 10^4$ dims.
1000 trials	with $n/d = c$	500 trials

Theory so far

- We can predict the asymptotic recovery of unweighted PCA with heteroscedastic data
- We can predict the asymptotic recovery of weighted PCA, given fixed weights
- Idea: Choose the weights by maximizing the asymptotic recovery of the principal components.

Asymptotically Optimal Weights

Theorem (Weight Design [Hong et al., 2018])

The asymptotic recovery of the i^{th} component is maximized by weights

$$w_\ell^2 = \frac{1}{\sigma_\ell^2} \frac{1}{\theta_i^2 + \sigma_\ell^2}$$

The optimal weights:

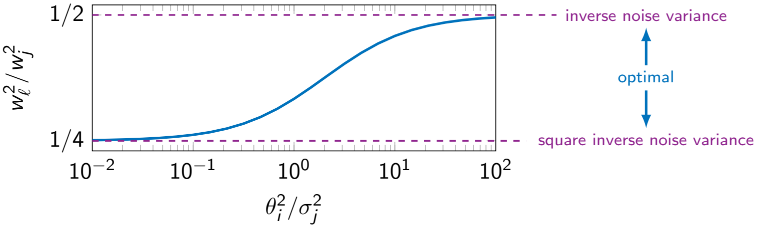
- **do not depend** on samples per dimension or on proportions
- **do depend** on subspace amplitudes - the weights are different for components with different amplitudes

Asymptotically Optimal Weights

Theorem (Weight Design [Hong et al., 2018])

The asymptotic recovery of the i^{th} component is maximized by weights

$$w_\ell^2 = \frac{1}{\sigma_\ell^2} \frac{1}{\theta_i^2 + \sigma_\ell^2}$$

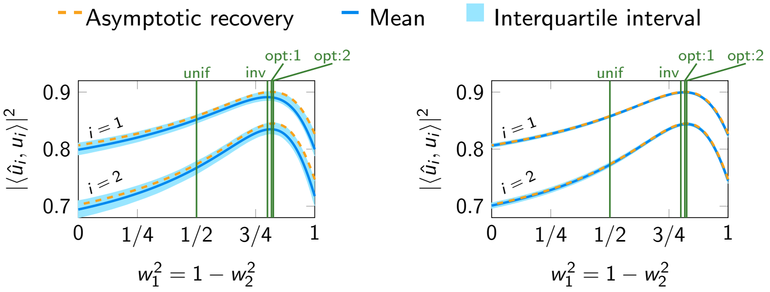


Inverse noise variance is <u>not</u> optimal!

Optimal weights downweight noisy samples more.

Optimal weighting

- $c = 1$ sample per dim
- $\theta_1^2 = 25, \theta_2^2 = 16$ amplitudes
- $p_1 = 20\%$ noise var. $\sigma_1^2 = 1$
- $p_2 = 80\%$ noise var. $\sigma_2^2 = 4$



$n = 10^3$ samples
$d = 10^3$ dims.
1000 trials

increasing n, d $\longrightarrow$ with $n/d = c$

$n = 10^4$ samples
$d = 10^4$ dims.
500 trials

Sub-optimal weighting: two case studies

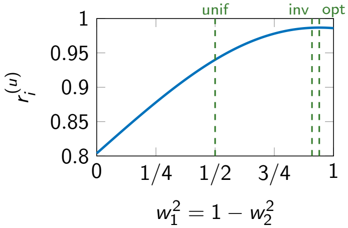
Both cases: $c = 10$ samples per dim; $\theta_i^2 = 1.5$ amplitude;
p_1 at $\sigma_1^2 = 0.1$ (cleaner), p_2 at $\sigma_2^2 = 1$ (noisier).

Sub-optimal weighting: two case studies

Both cases: $c = 10$ samples per dim; $\theta_i^2 = 1.5$ amplitude;
p_1 at $\sigma_1^2 = 0.1$ (cleaner), p_2 at $\sigma_2^2 = 1$ (noisier).

Recall: $r_i^{(u)}$ is our prediction for $|\langle \hat{u}_i, u_i \rangle|^2$.

Case 1: $p_1 = p_2 = 50\%$

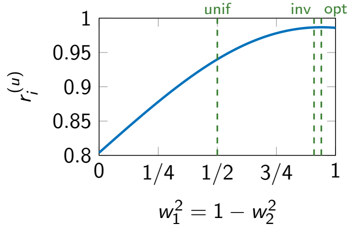


Unweighted PCA on only clean data is near optimal.

Sub-optimal weighting: two case studies

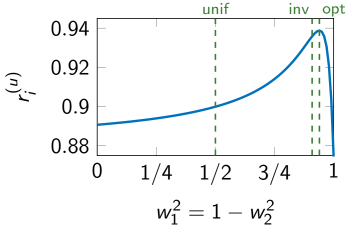
Both cases: $c = 10$ samples per dim; $\theta_i^2 = 1.5$ amplitude;
p_1 at $\sigma_1^2 = 0.1$ (cleaner), p_2 at $\sigma_2^2 = 1$ (noisier).

Case 1: $p_1 = p_2 = 50\%$



Unweighted PCA on only clean data is near optimal.

Case 2: $p_1 = 5\%, p_2 = 95\%$

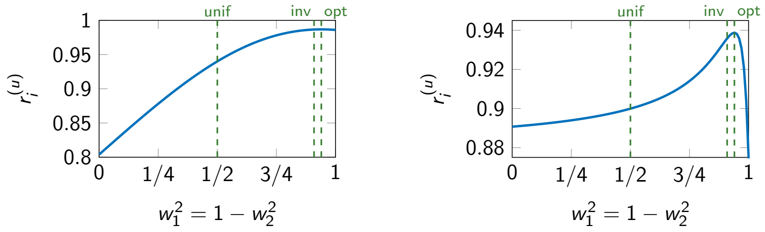


Weighting significantly better than unweighted PCA.

Sub-optimal weighting: two case studies

Both cases: $c = 10$ samples per dim; $\theta_i^2 = 1.5$ amplitude;
p_1 at $\sigma_1^2 = 0.1$ (cleaner), p_2 at $\sigma_2^2 = 1$ (noisier).

Case 1: $p_1 = p_2 = 50\%$ **Case 2:** $p_1 = 5\%, p_2 = 95\%$



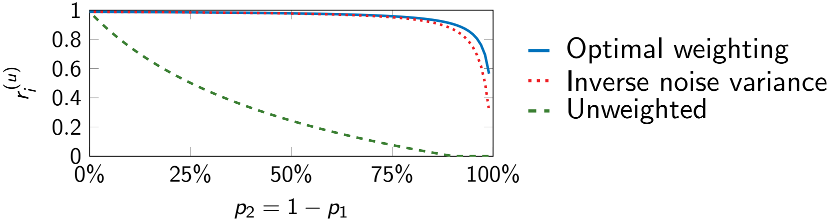
Unweighted PCA on only clean data is near optimal.

Weighting significantly better than unweighted PCA.

Optimal weights are same but recovery "curve" is quite different!

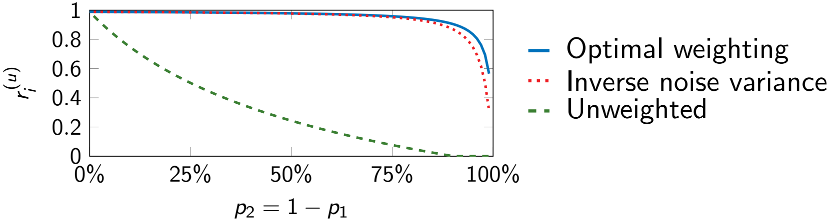
Impact of contamination on component recovery

- $c = 10$ samples per dim
- $\theta_i^2 = 1$ amplitude
- p_1 at noise var. $\sigma_1^2 = 0.1$ (clean)
- p_2 at noise var. $\sigma_2^2 = 3.25$ (noisy)



Impact of contamination on component recovery

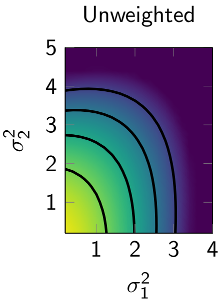
- $c = 10$ samples per dim
- $\theta_i^2 = 1$ amplitude
- p_1 at noise var. $\sigma_1^2 = 0.1$ (clean)
- p_2 at noise var. $\sigma_2^2 = 3.25$ (noisy)



Weighting makes PCA more robust to contamination in the data.
Optimal weighting is even more robust for extreme contamination.

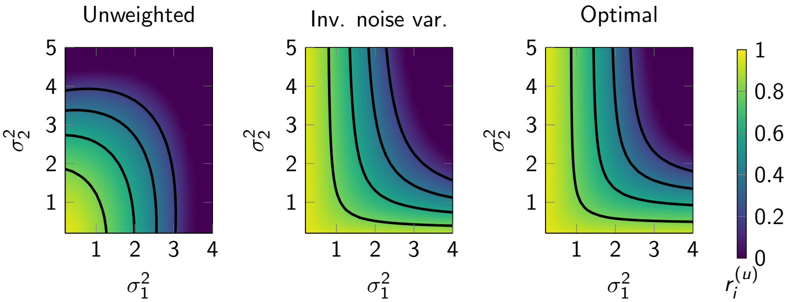
Impact of noise variances on component recovery

- $c = 10$ samples per dim
- $\theta_i^2 = 1$ amplitude
- $p_1 = 70\%$ at noise var. σ_1^2
- $p_2 = 30\%$ at noise var. σ_2^2



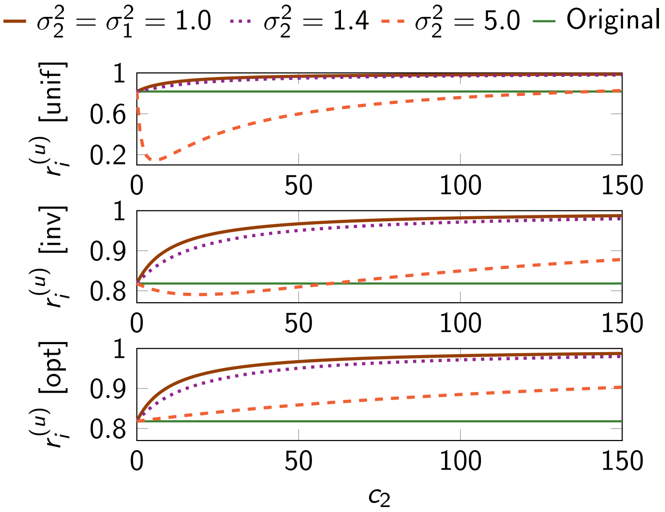
Impact of noise variances on component recovery

- $c = 10$ samples per dim
- $\theta_i^2 = 1$ amplitude
- $p_1 = 70\%$ at noise var. σ_1^2
- $p_2 = 30\%$ at noise var. σ_2^2



Unweighted PCA is most sensitive to largest noise variance.
Weighted PCA is most sensitive to smallest noise variance.

Impact of adding noisy data on component recovery



Discussion: Maximum Likelihood

$$y_i = \mathbf{U\Theta}z_i^{\mathrm{H}} + \eta_i \varepsilon_i$$

Suppose z_i and ε_i are all iid normal $\mathcal{N}(0,1)$. Then the maximum likelihood estimator for $\mathbf{U}$ is given as

$$\max_{U \in \mathcal{G}(d,k)} \sum_{i=1}^{n} y_i^T U \underbrace{\begin{bmatrix} \frac{1}{\eta_i^2}\frac{\theta_1^2}{\theta_1^2+\eta_i^2} & & \\ & \ddots & \\ & & \frac{1}{\eta_i^2}\frac{\theta_k^2}{\theta_k^2+\eta_i^2} \end{bmatrix}}_{\Gamma_i} U^T y_i = \sum_{i=1}^{n} y_i^T U \Gamma_i U^T y_i \, .$$

Conclusion

In summary, this work shows:

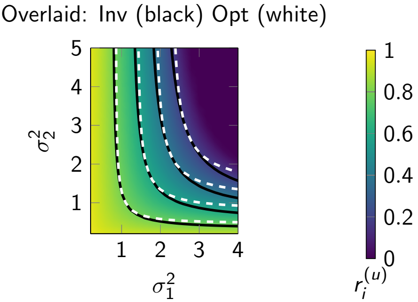
- analysis of the asymptotic performance of weighted PCA for high-dimensional and heteroscedastic data
- weights that optimize asymptotic recovery of the principal components
- numerical experiments illustrating practicality of asymptotics
- interesting cases/regimes where other weights are near-optimal
- how weighting changes the impact of data properties

References I

Thank you!

Hong, D., Fessler, J., and Balzano, L. (2018).
Optimally weighted PCA for high-dimensional heteroscedastic data.
Preprint at https://arxiv.org/abs/1810.12862.

Overlaid predictions



Back to impact

Theory: Asymptotic Weighted PCA performance

Theorem (Recovery of amplitudes [Hong et al., 2018])

Suppose the sample-to-dimension ratio $n/d \to c > 0$ and the proportions of different quality data $n_\ell/n \to p_\ell$ for $\ell = 1, \dots, L$ as $n, d \to \infty$. Then the i^{th} WPCA amplitude $\hat{\theta}_i$ converges as

$$\hat{\theta}_i^2 \xrightarrow{a.s.} r_i^{(\theta)} = \frac{1}{c} \max\{\alpha, \beta_i\} C(\max\{\alpha, \beta_i\}), \tag{1}$$

where α and β_i are, respectively, the largest real roots of $A(x)$ and $B_i(x)$ (given before) and

$$C(x) = 1 + c \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^2 \sigma_\ell^2}{x - w_\ell^2 \sigma_\ell^2}. \tag{2}$$

Theory: inverse noise variance

Recall

$$A(x) := 1 - c \sum_{\ell=1}^{L} p_\ell \frac{w_\ell^4 \sigma_\ell^4}{(x - w_\ell^2 \sigma_\ell^2)^2}, \quad B_i(x) := 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^2}{x - w_\ell^2 \sigma_\ell^2}.$$

Consider $w_\ell^2 = \frac{\bar{\sigma}^2}{\sigma_\ell^2}$. Then

$A(x) = 1 - \frac{c\bar{\sigma}^4}{(x - \bar{\sigma}^2)^2} \implies \alpha = \bar{\sigma}^2(1 + \sqrt{c})$, and

$B_i(x) = 1 - \frac{c\theta_i^2}{x - \bar{\sigma}^2} \implies \beta_i = \bar{\sigma}^2 + c\theta_i^2$.

$$\hat{\theta}_i^2 \xrightarrow{a.s.} r_i^{(\theta)} = \begin{cases} \theta_i^2 \left(1 + \frac{\bar{\sigma}^2}{c\theta_i^2}\right)\left(1 + \frac{\bar{\sigma}^2}{\theta_i^2}\right) & \text{if } c\theta_i^4 > \bar{\sigma}^4, \\ \bar{\sigma}^2(1 + 1/\sqrt{c})^2 & \text{otherwise.} \end{cases}$$

$$\sum_{j:\theta_j = \theta_i} |\langle \hat{u}_i, u_j \rangle|^2 \xrightarrow{a.s.} r_i^{(u)} = \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)} = \frac{c - \bar{\sigma}^4/\theta_i^4}{c + \bar{\sigma}^2/\theta_i^2}.$$

Theory: Base case, no noise

Recall

$$A(x) := 1 - c\sum_{\ell=1}^{L} \frac{p_\ell w_\ell^4 \sigma_\ell^4}{(x - w_\ell^2\sigma_\ell^2)^2}, \quad B_i(x) := 1 - c\theta_i^2 \sum_{\ell=1}^{L} \frac{p_\ell w_\ell^2}{x - w_\ell^2 \sigma_\ell^2}.$$

Consider the noise free case: All $\sigma_i = 0$. Then $A(x) = C(x) = 1$ $\forall x$, and $B_i(x) = 1 - \frac{c\theta_i^2 \sum_\ell p_\ell w_\ell^2}{x} \implies \beta_i = c\theta_i^2 \sum_\ell p_\ell w_\ell^2$.

$$\hat{\theta}_i^2 \xrightarrow{a.s.} \frac{1}{c}\beta_i = \theta_i^2 \sum_\ell p_\ell w_\ell^2$$

$$\sum_{j:\theta_j=\theta_i} |\langle \hat{u}_i, u_j \rangle|^2 \xrightarrow{a.s.} \frac{A(\beta_i)}{\beta_i B_i'(\beta_i)} = 1$$