Probabilistic PCA for Heteroscedastic Data

David Hong

Kyle Gilman

Laura Balzano

Jeffrey A. Fessler

EECS, University of Michigan

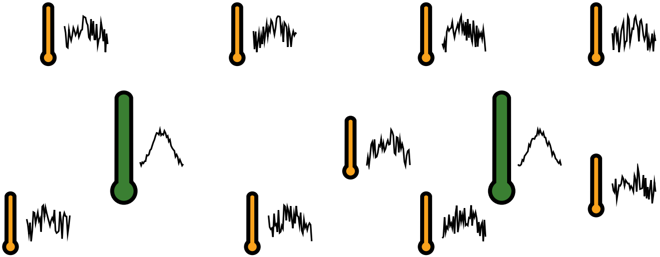
UM BIGDATA 2020-01-15

Outline

Modern data: High-Dimensional & Heteroscedastic

Modern datasets increasingly contain data of varying quality.

For example: mixture of a few high quality (costly) sensors with many low quality (inexpensive) sensors.



Question: How should we account for heterogeneous quality?

Modern data: High-Dimensional & Heteroscedastic

Many settings: medical imaging, astronomy, sensor networks, ...

varying radiation levels

varying atmosphere

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

Many settings: medical imaging, astronomy, sensor networks, ...

varying radiation levels
(millions of voxels)

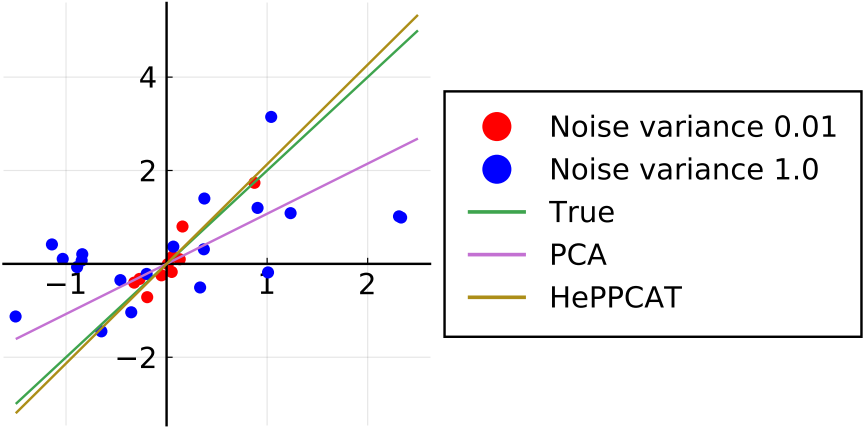
varying atmosphere
(thousands of pixels)

varying sensor quality
(thousands of locations)

Much modern data is high-dimensional...with low-dimensional structure.
A standard (well-understood?) tool: Principal Component Analysis

How should we account for heterogeneous quality in PCA?

`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`

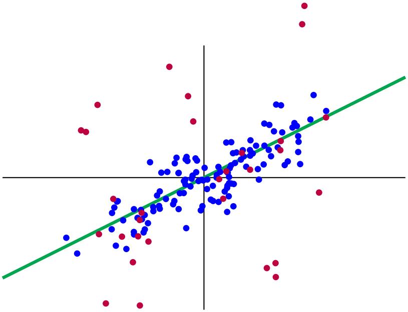


HePPCAT: Heteroscedastic probabilistic PCA technique

Modern data: High-Dimensional & Heteroscedastic

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

$$y_i = \boldsymbol{F} z_i + \sigma_i \varepsilon_i.$$



Modern data: High-Dimensional & Heteroscedastic

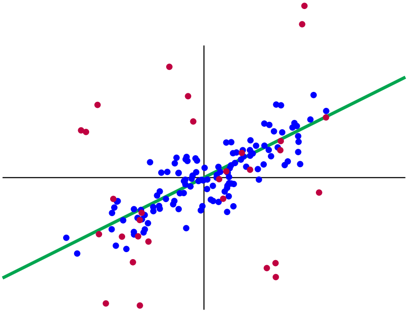
Model samples $y_1, \ldots, y_n \in \mathbb{R}^d$ as

$$y_i = \boldsymbol{F} z_i + \sigma_i \varepsilon_i .$$

k latent factors

$d \times k$ matrix

coefficients

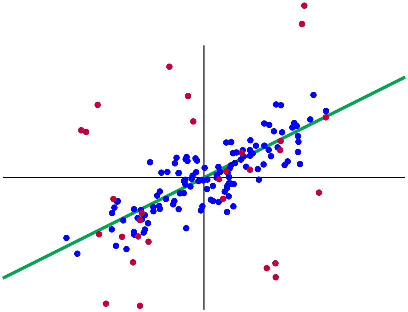


Modern data: High-Dimensional & Heteroscedastic

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

$$y_i = \boldsymbol{F} z_i + \sigma_i \varepsilon_i .$$

noise std. dev. noise



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

Modern data: High-Dimensional & Heteroscedastic

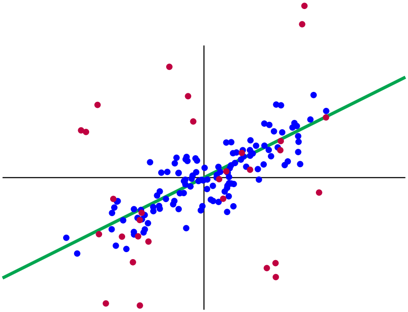
Model samples $y_1, \ldots, y_n \in \mathbb{R}^d$ as

$$y_i = \boldsymbol{F} z_i + \sigma_i \varepsilon_i .$$

noise std. dev. noise

In matrix form

$$\begin{aligned} \boldsymbol{Y} &= [y_1, \ldots, y_n] \\ &= \boldsymbol{F}\,[z_1, \ldots, z_n] + [\sigma_1 \varepsilon_1, \ldots, \sigma_n \varepsilon_n]. \end{aligned}$$



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

Modern data: High-Dimensional & Heteroscedastic

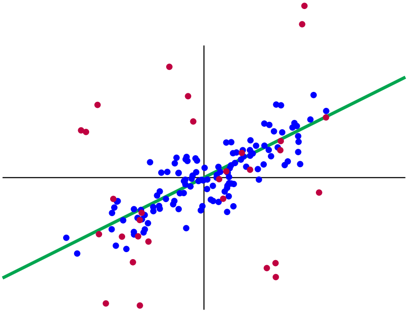
Model samples $y_1, \dots, y_n \in \mathbb{R}^d$ as

$$y_i = \boldsymbol{F} z_i + \sigma_i \varepsilon_i.$$

noise std. dev. noise

In matrix form

$$\begin{aligned} \boldsymbol{Y} &= [y_1, \dots, y_n] \\ &= \boldsymbol{F}\,[z_1, \dots, z_n] + [\sigma_1 \varepsilon_1, \dots, \sigma_n \varepsilon_n]. \end{aligned}$$



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

Samples have heteroscedastic noise.
Homoscedastic if $\sigma_1 = \cdots = \sigma_n$.

Weighted FA / PCA / low-rank approximation

Recall:

$$y_i = \boldsymbol{F} z_i + \varepsilon_i \sim \mathcal{N}(\boldsymbol{F} z_i, \sigma_i^2 \boldsymbol{I})$$

Negative log-likelihood:

$$L(\boldsymbol{F}, \boldsymbol{Z}) = \sum_i \frac{1}{2\sigma_i^2} \|y_i - \boldsymbol{F} z_i\|_2^2.$$

Change of variables: $\tilde{y}_i = y_i / \sigma_i, \ \tilde{z}_i = z_i / \sigma_i$

$$\min_{\boldsymbol{F}, \boldsymbol{X}} \|\tilde{\boldsymbol{Y}} - \boldsymbol{F} \tilde{\boldsymbol{Z}}\|_{\mathrm{F}}^2 \implies \hat{\boldsymbol{F}} = \text{ first } k \text{ left singular vectors of } \tilde{\boldsymbol{Y}}$$

This is "sample-weighted" low-rank approximation / factorization.
aka ML factor analysis (Young, 1941)
See Hong et al. arXiv 1810.12862 for optimally weighted PCA

Outline

Review: Probabilistic PCA for homoscedastic noise

Model: $y_i = \boldsymbol{F} z_i + \sigma \varepsilon_i \in \mathbb{R}^d$ with $z_i \sim \mathcal{N}(0, \boldsymbol{I}_k)$ and $\varepsilon_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$

Review: Probabilistic PCA for homoscedastic noise

Model: $y_i = \boldsymbol{F} z_i + \sigma \varepsilon_i \in \mathbb{R}^d$ with $z_i \sim \mathcal{N}(0, \boldsymbol{I}_k)$ and $\varepsilon_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$

Equivalently, $y_i \sim \mathcal{N}(0, \boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)$.

Review: Probabilistic PCA for homoscedastic noise

Model: $y_i = \boldsymbol{F} z_i + \sigma \varepsilon_i \in \mathbb{R}^d$ with $z_i \sim \mathcal{N}(0, \boldsymbol{I}_k)$ and $\varepsilon_i \sim \mathcal{N}(0, \boldsymbol{I}_d)$

Equivalently, $y_i \sim \mathcal{N}(0, \boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)$.

Approach: Maximize the log-likelihood of $\boldsymbol{F}$ (dropping constants)

$$\mathcal{L}(\boldsymbol{F}) := \frac{1}{2} \sum_{i=1}^{n} \left\{ \ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} - y_i'(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} y_i \right\}$$

$$= \frac{n}{2} \left\{ \ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} - \frac{1}{n} \sum_{i=1}^{n} y_i'(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} y_i \right\}$$

$$= \frac{n}{2} \left[\ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{ \frac{1}{n} \sum_{i=1}^{n} y_i y_i' (\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} \right\} \right].$$

Data appears only through the sample correlation matrix!

Review: Probabilistic PCA for homoscedastic noise

$$\mathcal{L}(\boldsymbol{F}) = \frac{n}{2}\left[\ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^{n} y_i y_i' (\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1}\right\}\right].$$

Review: Probabilistic PCA for homoscedastic noise

$$\mathcal{L}(\boldsymbol{F}) = \frac{n}{2}\left[\ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2\boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^{n} y_i y_i' (\boldsymbol{F}\boldsymbol{F}' + \sigma^2\boldsymbol{I}_d)^{-1}\right\}\right].$$

Fact (Tipping & Bishop 1999): The likelihood is maximized by

$$\hat{\boldsymbol{F}} = \boldsymbol{V}\operatorname{diag}\left(\sqrt{\lambda_1 - \bar{\lambda}}, \ldots, \sqrt{\lambda_k - \bar{\lambda}}\right),$$

where in terms of the sample correlation matrix $\frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}'$:

- $\boldsymbol{V} \in \mathbb{R}^{d \times k}$ contains the k principal eigenvectors,
- $\lambda_1, \ldots, \lambda_k$ are the k principal eigenvalues, and
- $\bar{\lambda}$ is the average of the remaining eigenvalues.

Review: Probabilistic PCA for homoscedastic noise

$$\mathcal{L}(\boldsymbol{F}) = \frac{n}{2}\left[\ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^{n} y_i y_i' (\boldsymbol{F}\boldsymbol{F}' + \sigma^2 \boldsymbol{I}_d)^{-1}\right\}\right].$$

Fact (Tipping & Bishop 1999): The likelihood is maximized by

$$\hat{\boldsymbol{F}} = \boldsymbol{V} \operatorname{diag}\left(\sqrt{\lambda_1 - \bar{\lambda}}, \ldots, \sqrt{\lambda_k - \bar{\lambda}}\right),$$

where in terms of the sample correlation matrix $\frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}'$:

- $\boldsymbol{V} \in \mathbb{R}^{d \times k}$ contains the k principal eigenvectors,
- $\lambda_1, \ldots, \lambda_k$ are the k principal eigenvalues, and
- $\bar{\lambda}$ is the average of the remaining eigenvalues.

Clean solution: Eigendecomposition + shrinkage

Review: Probabilistic PCA for homoscedastic noise

$$\mathcal{L}(\boldsymbol{F}) = \frac{n}{2}\left[\ln\det(\boldsymbol{F}\boldsymbol{F}' + \sigma^2\boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^{n} y_i y_i' (\boldsymbol{F}\boldsymbol{F}' + \sigma^2\boldsymbol{I}_d)^{-1}\right\}\right].$$

Fact (Tipping & Bishop 1999): The likelihood is maximized by

$$\hat{\boldsymbol{F}} = \boldsymbol{V}\operatorname{diag}\left(\sqrt{\lambda_1 - \bar{\lambda}}, \ldots, \sqrt{\lambda_k - \bar{\lambda}}\right),$$

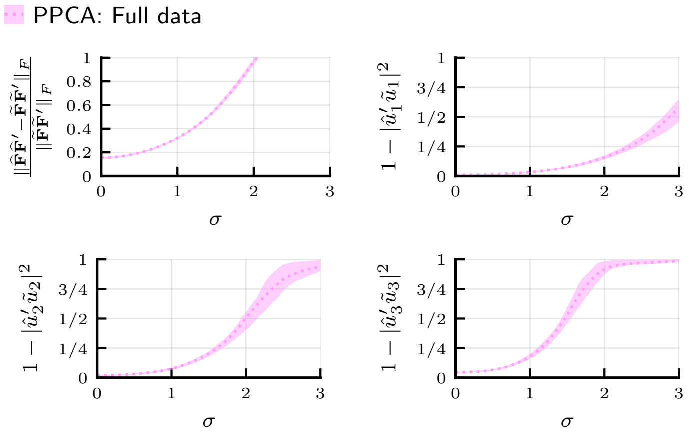
where in terms of the sample correlation matrix $\frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}'$:

- $\boldsymbol{V} \in \mathbb{R}^{d\times k}$ contains the k principal eigenvectors,
- $\lambda_1, \ldots, \lambda_k$ are the k principal eigenvalues, and
- $\bar{\lambda}$ is the average of the remaining eigenvalues.

Clean solution: Eigendecomposition + shrinkage
Components $\boldsymbol{V}$ *are same as in ordinary PCA*

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

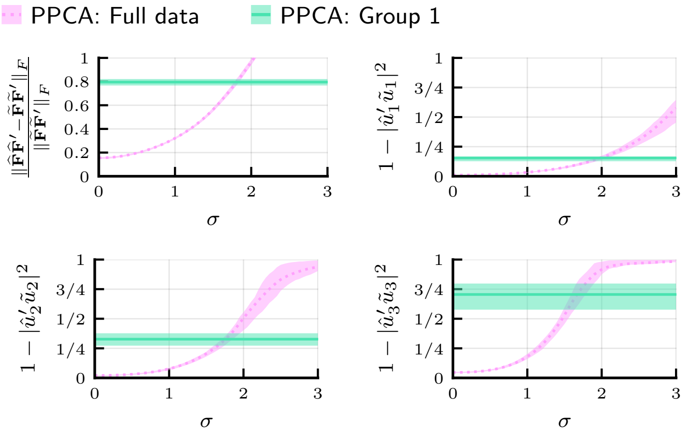
$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + \sigma_i^2 \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3]\ \mathrm{diag}(4, 2, 1),$$



(Homoscedastic) PPCA degrades with heteroscedasticity

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

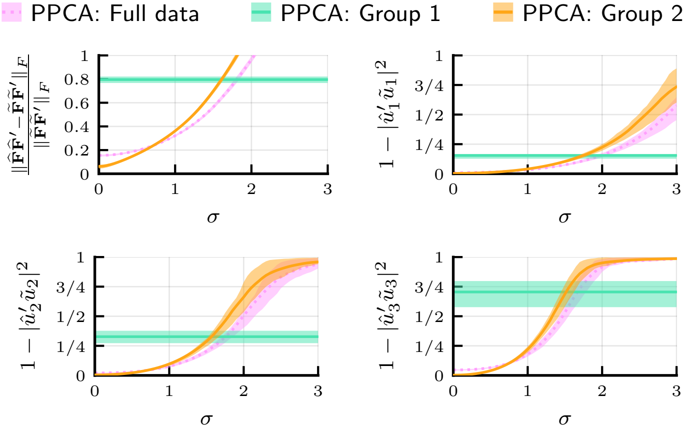
$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + \sigma_i^2 \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$



(Homoscedastic) PPCA degrades with heteroscedasticity

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

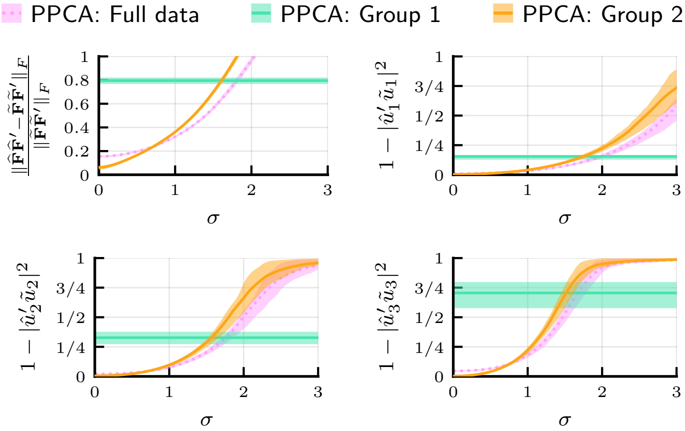
$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + \sigma_i^2 \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$



(Homoscedastic) PPCA degrades with heteroscedasticity

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + \sigma_i^2 \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$



PPCA degrades as data becomes heteroscedastic.
PPCA may work better using only part of the data!?

Outline

Deriving a heterocedastic PPCA

This talk: Develop a heteroscedastic PPCA algorithm.

Version 1: Assume noise variances $\{v_i = \sigma_i^2\}$ are known.

Deriving a heterocedastic PPCA

This talk: Develop a heteroscedastic PPCA algorithm.
Version 1: Assume noise variances $\{v_i = \sigma_i^2\}$ are known.

Challenge: Log-likelihood no longer separates so nicely:

$$\mathcal{L}(\boldsymbol{F}) = \frac{1}{2} \sum_{i=1}^{n} \left\{ \ln\det(\boldsymbol{F}\boldsymbol{F}' + v_i \boldsymbol{I}_d)^{-1} - y_i'(\boldsymbol{F}\boldsymbol{F}' + v_i \boldsymbol{I}_d)^{-1} y_i \right\} \quad (1)$$

Apparently no closed-form for ML estimate of $\boldsymbol{F}$.

Deriving a heterocedastic PPCA

This talk: Develop a heteroscedastic PPCA algorithm.
Version 1: Assume noise variances $\{v_i = \sigma_i^2\}$ are known.

Challenge: Log-likelihood no longer separates so nicely:

$$\mathcal{L}(\boldsymbol{F}) = \frac{1}{2} \sum_{i=1}^{n} \left\{ \ln \det(\boldsymbol{F}\boldsymbol{F}' + v_i \boldsymbol{I}_d)^{-1} - y_i'(\boldsymbol{F}\boldsymbol{F}' + v_i \boldsymbol{I}_d)^{-1} y_i \right\} \tag{1}$$

Apparently no closed-form for ML estimate of $\boldsymbol{F}$.

Approach: Instead, derive an Expectation Maximization (EM) algorithm using the complete data likelihood with complete data $\{(y_i, z_i)\}$

$$\mathcal{L}_{\mathrm{c}}(\boldsymbol{F}) := -\sum_{i=1}^{n} \left(\frac{\|y_i - \boldsymbol{F} z_i\|_2^2}{2 v_i} + \frac{\|z_i\|_2^2}{2} \right).$$

(Tipping & Bishop 1999) derived it for homoscedastic PPCA.

EM for heterocedastic PPCA

$$\mathcal{L}_c(\boldsymbol{F}) := -\sum_{i=1}^{n} \left(\frac{\|y_i - \boldsymbol{F} z_i\|_2^2}{2v_i} + \frac{\|z_i\|_2^2}{2} \right).$$

E-Step: Expectation with respect to $z_1, \dots, z_n | y_1, \dots, y_n, \boldsymbol{F}_t$ is

$$\bar{\mathcal{L}}(\boldsymbol{F}; \boldsymbol{F}_t) = \sum_{i=1}^{n} \left[\frac{1}{v_i} y_i' \boldsymbol{F} \bar{z}_{t,i} - \frac{1}{2v_i} \operatorname{tr}\{\boldsymbol{F}' \boldsymbol{F} (\bar{z}_{t,i} \bar{z}_{t,i}' + v_i \mathrm{M}_{t,i})\} \right],$$

up to constants where

$$\mathrm{M}_{t,i} := (\boldsymbol{F}_t' \boldsymbol{F}_t + v_i \boldsymbol{I}_k)^{-1}, \qquad \bar{z}_{t,i} := \mathrm{M}_{t,i} \boldsymbol{F}_t' y_i.$$

M-step: $\bar{\mathcal{L}}(\boldsymbol{F}; \boldsymbol{F}_t)$ is maximized with respect to $\boldsymbol{F}$ by

$$\boldsymbol{F}_{t+1} = \left(\sum_{i=1}^{n} \frac{1}{v_i} y_i \bar{z}_{t,i}' \right) \left(\sum_{i=1}^{n} \frac{1}{v_i} \bar{z}_{t,i} \bar{z}_{t,i}' + \mathrm{M}_{t,i} \right)^{-1}.$$

EM for heterocedastic PPCA

$$\mathcal{L}_{\mathrm{c}}(\boldsymbol{F}) := -\sum_{i=1}^{n}\left(\frac{\|y_i - \boldsymbol{F}z_i\|_2^2}{2v_i} + \frac{\|z_i\|_2^2}{2}\right).$$

E-Step: Expectation with respect to $z_1, \ldots, z_n | y_1, \ldots, y_n, \boldsymbol{F}_t$ is

$$\bar{\mathcal{L}}(\boldsymbol{F}; \boldsymbol{F}_t) = \sum\nolimits_{i=1}^{n}\left[\frac{1}{v_i} y_i' \boldsymbol{F} \bar{z}_{t,i} - \frac{1}{2v_i} \operatorname{tr}\{\boldsymbol{F}'\boldsymbol{F}(\bar{z}_{t,i}\bar{z}_{t,i}' + v_i \mathrm{M}_{t,i})\}\right],$$

up to constants where

$$\mathrm{M}_{t,i} := (\boldsymbol{F}_t'\boldsymbol{F}_t + v_i \boldsymbol{I}_k)^{-1}, \qquad \bar{z}_{t,i} := \mathrm{M}_{t,i}\boldsymbol{F}_t' y_i.$$

M-step: $\bar{\mathcal{L}}(\boldsymbol{F}; \boldsymbol{F}_t)$ is maximized with respect to $\boldsymbol{F}$ by

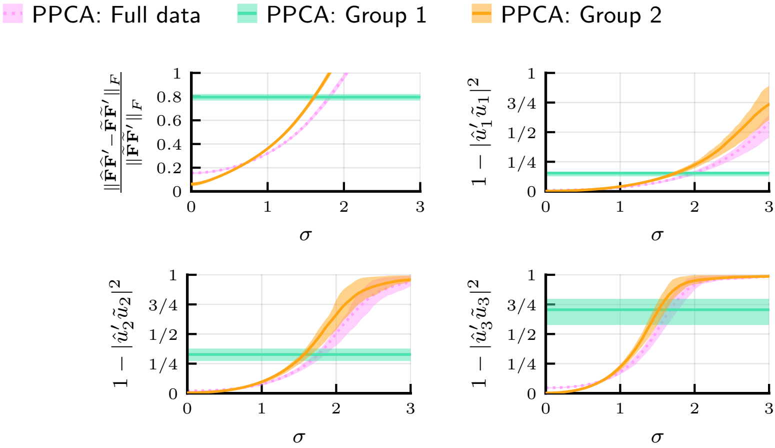
$$\boldsymbol{F}_{t+1} = \left(\sum\nolimits_{i=1}^{n} \frac{1}{v_i} y_i \bar{z}_{t,i}'\right)\left(\sum\nolimits_{i=1}^{n} \frac{1}{v_i} \bar{z}_{t,i}\bar{z}_{t,i}' + \mathrm{M}_{t,i}\right)^{-1}.$$

Now the factor estimates $\hat{\boldsymbol{F}}$ differ from PCA eigenvectors.

Heteroscedastic PPCA effectively uses all data

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

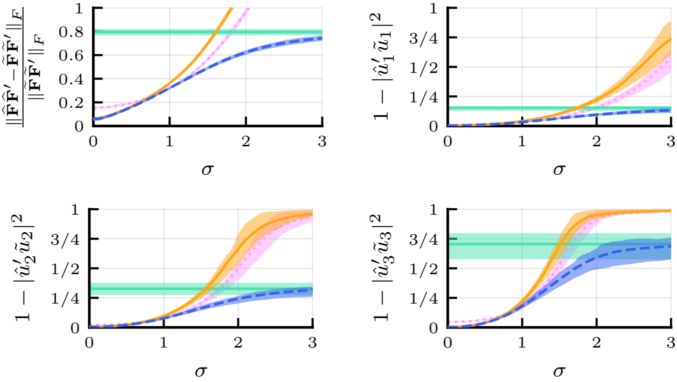
$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + v_i \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$



Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + v_i \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$

PPCA: Full data PPCA: Group 1 PPCA: Group 2 HeteroPPCA

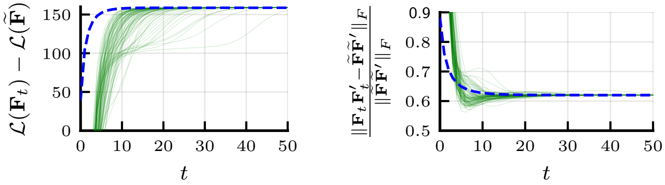


But how to initialize for this nonconcave problem?

Setup: $n_1 = 200$ samples with $v_i = 1$, $n_2 = 800$ samples with $v_i = 4$.

$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + v_i \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1)^{1/2},$$

Rand init (true eigvals, rand eigvecs) vs. Homoscedastic PPCA init



Interesting phenomenon:
iterates do not seem to hit bad local maxima! Why?

What about doing a weighted PCA?

Idea: Give noisier samples less weight in PCA.

Simple tweak: Replace the sample covariance with a *weighted* version

$$\frac{1}{n}\sum_{i=1}^{n} y_i y_i' \qquad \longrightarrow \qquad \frac{1}{n}\sum_{i=1}^{n} \omega_i^2 y_i y_i',$$

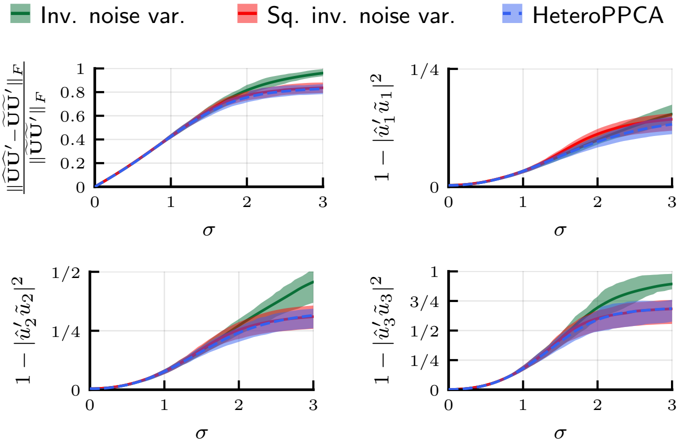
Choices for weights:

- unweighted: $\omega_i^2 = 1$
- inverse noise variance $\omega_i^2 = 1/\sigma_i^2$
  - rescales data to make noise homoscedastic
  - corresponds to MLE for low-rank signal $\boldsymbol{F}[z_1, \ldots, z_n]$
- square inverse noise variance $\omega_i^2 = 1/\sigma_i^4$
  - more aggressive down-weighting, can help for low SNRs

High-dim asymptotics studied in (Hong, Fessler, Balzano 2019).

Setup: $n_1 = 200$ samples with $\sigma_i = 1$, $n_2 = 800$ samples with $\sigma_i = \sigma$.

$$y_i \sim \mathcal{N}(0, \boldsymbol{F}_* \boldsymbol{F}_*' + \sigma_i^2 \boldsymbol{I}_{100}), \qquad \boldsymbol{F}_* = [\tilde{u}_1, \ldots, \tilde{u}_3] \operatorname{diag}(4, 2, 1),$$



Conclusions for Version 1 (known noise variances)

- PPCA for data with heteroscedastic noise that uses all the data effectively, no matter how noisy
- practical EM algorithm
- interesting phenomenon: do not seem to get stuck in bad local maxima!

Conclusions for Version 1 (known noise variances)

- PPCA for data with heteroscedastic noise that uses all the data effectively, no matter how noisy
- practical EM algorithm
- interesting phenomenon: do not seem to get stuck in bad local maxima!

Ongoing/future: (from Dec. 2019 CAMSAP talk)

- joint estimation of noise variance
- optimization landscape
- other optimization approaches (manifold optimization?, minorize maximize?)
- analysis of asymptotic performance

Outline

HePPCAT Version 2: unknown noise variances

Model $n_1 + \cdots + n_L = n$ data samples in $\mathbb{R}^d$ from L noise level groups:

$$\boldsymbol{y}_{\ell,i} = \boldsymbol{F}\boldsymbol{z}_{\ell,i} + \boldsymbol{\varepsilon}_{\ell,i}, \quad i \in \{1, \dots, n_\ell\},\ \ell \in \{1, \dots, L\}, \tag{2}$$

$\boldsymbol{F} \in \mathbb{R}^{d \times k}$: deterministic factor matrix to estimate
$\boldsymbol{z}_{\ell,i} \sim \mathcal{N}(0_k, \boldsymbol{I}_k)$: random coefficients
$\boldsymbol{\varepsilon}_{\ell,i} \sim \mathcal{N}(0_d, v_\ell \boldsymbol{I}_d)$: noise vectors
$v_1, \dots, v_L$: deterministic noise variances to estimate.

Equivalently:

$$\boldsymbol{y}_{\ell,i} \sim \mathcal{N}(0_d, \boldsymbol{F}\boldsymbol{F}' + v_\ell \boldsymbol{I}_d)$$

Joint log-likelihood

Joint log-likelihood for factors $\boldsymbol{F}$ and variances $\boldsymbol{v}$ (dropping a constant):

$$\mathcal{L}(\boldsymbol{F}, \boldsymbol{v}) \triangleq \frac{1}{2} \sum_{\ell=1}^{L} \left[n_\ell \ln \det(\boldsymbol{F}\boldsymbol{F}' + v_\ell \boldsymbol{I}_d)^{-1} - \operatorname{tr}\left\{ \boldsymbol{Y}_\ell' (\boldsymbol{F}\boldsymbol{F}' + v_\ell \boldsymbol{I}_d)^{-1} \boldsymbol{Y}_\ell \right\} \right],$$

$\boldsymbol{Y}_\ell \triangleq [\boldsymbol{y}_{\ell,1}, \ldots, \boldsymbol{y}_{\ell,n_\ell}] \in \mathbb{R}^{d \times n_\ell}$ for $\ell \in \{1, \ldots, L\}$:
data matrices for each of the L groups.

Similar to earlier log-likelihood (1).

Apparently no closed form ML estimates for $\boldsymbol{F}$ and $\boldsymbol{v}$
No closed-form ML estimate of $\boldsymbol{F}$ given $\boldsymbol{v}$
No closed-form ML estimate of $\boldsymbol{v}$ given $\boldsymbol{F}$
$\Longrightarrow$ Alternating ascent algorithms

$\boldsymbol{F}$ update : EM update similar to Version 1 HePPCAT
$\boldsymbol{v}$ update : separates into L univariate maximizations (over $v_\ell \geq 0$)

$$\mathcal{L}_\ell(v_\ell) \triangleq -\sum_{j=0}^{k} \left\{ \alpha_j \ln(\gamma_j + v_\ell) + \frac{\beta_j}{\gamma_j + v_\ell} \right\}, \tag{3}$$

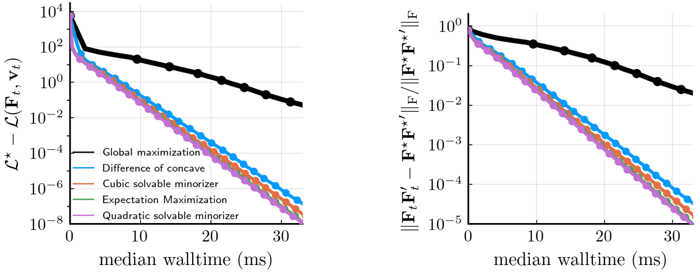
where $\alpha_0 \triangleq d - k$, $\beta_0 \triangleq \|(\boldsymbol{I}_d - \boldsymbol{U}_t \boldsymbol{U}_t')\boldsymbol{Y}_\ell\|_{\mathrm{F}}^2 / n_\ell$, $\gamma_0 \triangleq 0$,

$$j \geq 1: \qquad \alpha_j \triangleq 1, \qquad \beta_j \triangleq \|\boldsymbol{Y}_\ell' \boldsymbol{\mu}_{t,j}\|_2^2 / n_\ell, \qquad \gamma_j \triangleq \lambda_{t,j},$$

and $\boldsymbol{U}_t = [\boldsymbol{\mu}_{t,1}, \ldots, \boldsymbol{\mu}_{t,k}]$ and $\boldsymbol{\lambda}_t = (\lambda_{t,1}, \ldots, \lambda_{t,k})$ are the eigenvectors and eigenvalues of $\boldsymbol{F}_t \boldsymbol{F}_t'$ at iteration t.

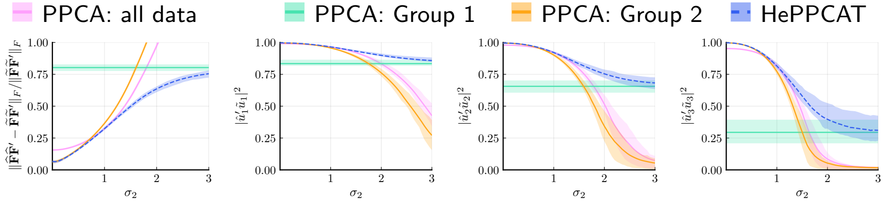
Univariate but non-convex.
Methods developed and investigated:
root finding, EM, minorize-maximize (MM), difference of concave...

Algorithm examples



(a) Convergence w.r.t. $\mathcal{L}$. (b) Convergence w.r.t. $\boldsymbol{F}$.

Figure: Comparison of update methods for $n = 10^3$ samples in $d = 10^2$ dimensions with $k = 3$ underlying factors $\boldsymbol{\lambda}_* = (4, 2, 1)$. The noise is heteroscedastic: the first $n_1 = 200$ samples have noise variance $\tilde{v}_1 = 1$ and the remaining $n_2 = 800$ have $\tilde{v}_2 = 4$. Walltimes are medians taken over 100 runs of the algorithm to reduce the effect of experimental noise. Markers are placed every five iterations.



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

Figure: Comparison with homoscedastic PPCA applied on: i) full data, ii) only group 1, i.e, the $n_1 = 200$ samples with noise variance $\tilde{v}_1 = 1$, and iii) only group 2, i.e., the $n_2 = 800$ samples with noise variance $\tilde{v}_2 = \sigma_2^2$. Lower is better in (a), and higher is better in (b)-(d). HePPCAT outperforms the homoscedastic methods on all four metrics.

Here, problem size is large enough that HePPCAT with unknown variance work essentially as well as with known variance.

Effect of block size on variance estimates

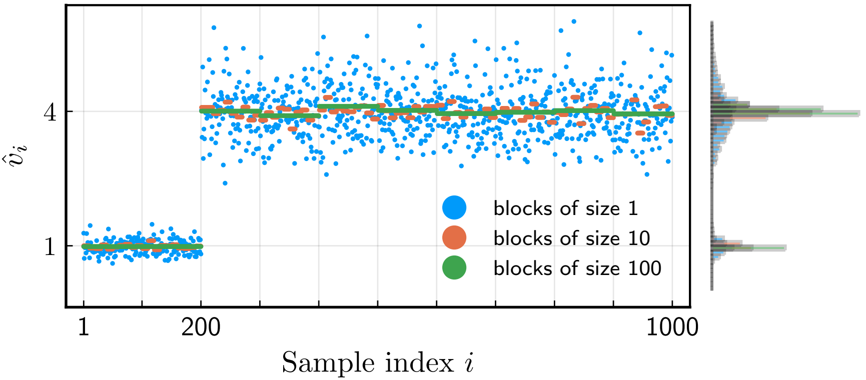


Figure: Estimated noise variances for various block sizes.

Effect of block size on factor estimates

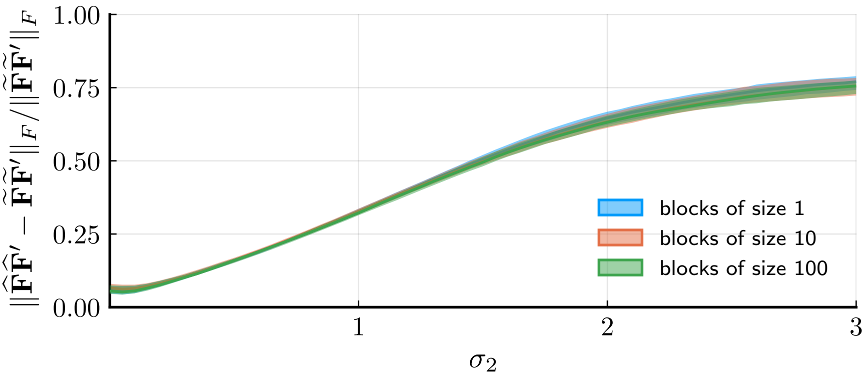
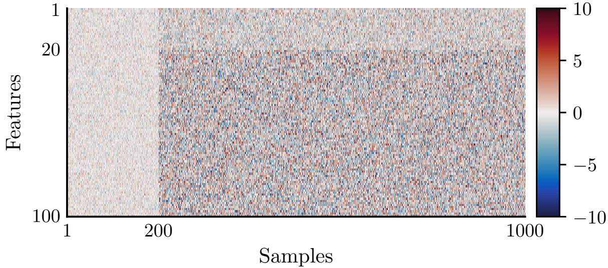


Figure: Normalized factor estimation error (median and interquartile intervals) for varying block sizes.

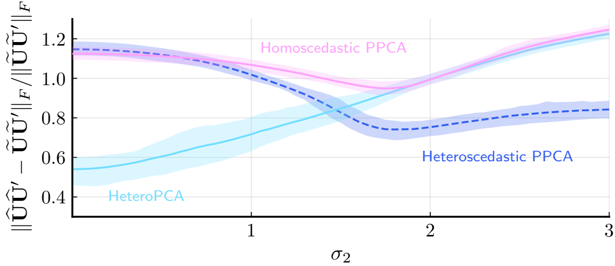
The three block sizes lead to practically identical performance here.

Heteroscedasticity across features *and* samples

Single data realization (heatmap).



Normalized error in subspace estimation (lower is better).



Zhang, Cai, Wu. 2018
ArXiv 1810.08316

Apparently favorable convergence

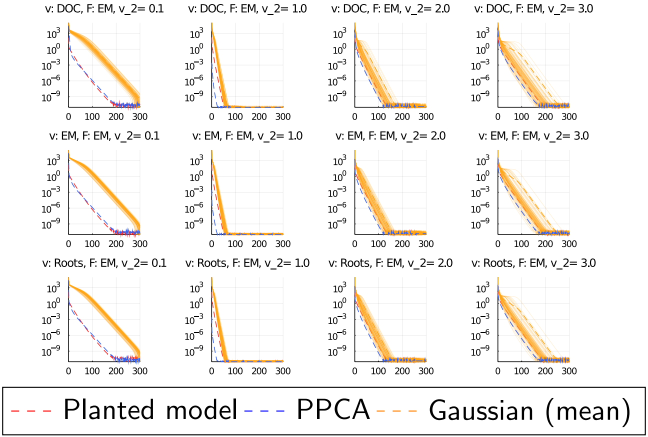


Figure: Convergence gaps of each algorithm to the maximum converged log-likelihood per heteroscedastic noise experiment. $n = [200, 800]$ and $v_1 = 1$. Three different types of initializations.

Outline

Introduction

Weighted PCA

Homoscedastic PPCA (review)

Heterocedastic PPCA: known variances (2019)

Heterocedastic PPCA: unknown variances (2020)

Extensions (2021)

Extensions

- Incremental / online algorithms

Extensions

- Incremental / online algorithms
- Acceleration via momentum

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)
- Nonnegative PCA, sparse PCA, robust PCA, etc.,
  all generalized for heteroscedastic noise

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)
- Nonnegative PCA, sparse PCA, robust PCA, etc., all generalized for heteroscedastic noise
- Supervised HePPCAT?

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)
- Nonnegative PCA, sparse PCA, robust PCA, etc., all generalized for heteroscedastic noise
- Supervised HePPCAT?
- missing data, heterogeneity across features

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)
- Nonnegative PCA, sparse PCA, robust PCA, etc., all generalized for heteroscedastic noise
- Supervised HePPCAT?
- missing data, heterogeneity across features
- bias correction for variance estimates

Extensions

- Incremental / online algorithms
- Acceleration via momentum
- convergence theory / performance guarantees
- Sketching to reduce computation (e.g., for $\boldsymbol{v}$)
- Nonnegative PCA, sparse PCA, robust PCA, etc., all generalized for heteroscedastic noise
- Supervised HePPCAT?
- missing data, heterogeneity across features
- bias correction for variance estimates
- probabalistic dictionary learning ($\boldsymbol{F}$ wide, $\boldsymbol{z}$ sparse, so non-normal)

Supervised HePPCAT (new in 2021)

Data $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$; Factors $\boldsymbol{F} \in \mathbb{R}^{d \times k}$; Clay's version:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \min_{\boldsymbol{F} \in \mathbb{R}^{d \times k}} L(\boldsymbol{Y}' \boldsymbol{F} \beta) + \lambda \| \boldsymbol{Y} - \boldsymbol{F}\boldsymbol{F}' \boldsymbol{Y} \|_{\mathrm{F}}^2. \tag{4}$$

(Any statistical model associated with norm? If so, then homoscedastic.)

Supervised HePPCAT (new in 2021)

Data $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$; Factors $\boldsymbol{F} \in \mathbb{R}^{d \times k}$; Clay's version:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \min_{\boldsymbol{F} \in \mathbb{R}^{d \times k}} L(\boldsymbol{Y}'\boldsymbol{F}\beta) + \lambda \|\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{F}'\boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{4}$$

(Any statistical model associated with norm? If so, then homoscedastic.)

Supervised HePPCAT for known noise variances:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \min_{\boldsymbol{F} \in \mathbb{R}^{d \times k}} L(\boldsymbol{Y}'\boldsymbol{F}\beta) + \underbrace{\frac{\lambda}{2} \sum_i - \ln \det(\boldsymbol{F}\boldsymbol{F}' + \sigma_i^2) + \boldsymbol{y}_i'(\boldsymbol{F}\boldsymbol{F}' + \sigma_i^2)^{-1} \boldsymbol{y}_i}_{\text{neg. log likelihood}}.$$

EM majorizer for $\boldsymbol{F}$ update is quadratic, not "quartic" like (4):

$$L(\boldsymbol{Y}'\boldsymbol{F}\beta) + \sum_i \frac{\lambda}{2\sigma_i^2} \left[-2\boldsymbol{y}_i'\boldsymbol{F}\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_i'\boldsymbol{F}'\boldsymbol{F}\bar{\boldsymbol{z}}_i + \operatorname{tr}(\boldsymbol{F}\bar{\boldsymbol{M}}\boldsymbol{F}') \right].$$