Dimensionality Reduction

(Part 2)

18-698 / 42-632

Neural Signal Processing
Prof. Byron Yu

C) Probabilistic PCA (PPCA)

- So far, we have formulated PCA as a linear projection of data into lower dimensional space.

- Here, we will show that PCA can be expressed as the maximum likelihood solution of a probabilistic latent variable model.

  In other words, we will construct a latent variable model for which, when we maximize $p(X|\theta)$ with respect to θ, the resulting θ will be the PC directions.

C.1) Advantages of PPCA over conventional PCA

- PPCA assigns probabilities to data, so we can select the dimensionality M of low-d space and compare to other models using cross-validated likelihoods.

- PPCA has an explicit noise model, so it is able to more effectively denoise data than PCA.

- If data dimensionality D is large, diagonalization in conventional PCA is costly $O(D^3)$. If we only need top eigenvectors, we can compute them more efficiently using PPCA.

- Because PPCA is a probabilistic model, it can deal with missing data. PCA cannot.

$$\underline{x}_1 = \begin{bmatrix} \otimes \\ \\ \otimes \\ \\ \\ \end{bmatrix} \quad \underline{x}_2 = \begin{bmatrix} \otimes \\ \\ \\ \\ \otimes \end{bmatrix} \quad \underline{x}_3 = \begin{bmatrix} \\ \\ \otimes \\ \\ \\ \end{bmatrix} \quad \cdots$$

$\otimes$ denotes missing data

- PPCA represents a constrained form of Gaussian distribution

$$\begin{bmatrix} \ddots & & 0 \\ & \ddots & \\ 0 & & \ddots \end{bmatrix} \qquad \begin{bmatrix} & & \\ & \Sigma & \\ & & \end{bmatrix}$$

Diagonal covariance is often too constrained

Full covariance is often too flexible (easy to overfit)

PPCA provides nice compromise between diagonal and full covariance.

- Because PPCA is a probabilistic model, we can easily propose extensions, such as mixtures of PPCA models (analogous to mixtures of Gaussians).

- PPCA is a generative model, so we can generate samples from the distribution.

<u>Note</u>: Many of these advantages of PPCA over conventional PCA are typical advantages of probabilistic over non-probabilistic models.

C.2) Generative model for PPCA

$\underline{x} \in \mathbb{R}^D$ is high-dimensional observed data

$\underline{z} \in \mathbb{R}^M$ is low-dimensional latent variable

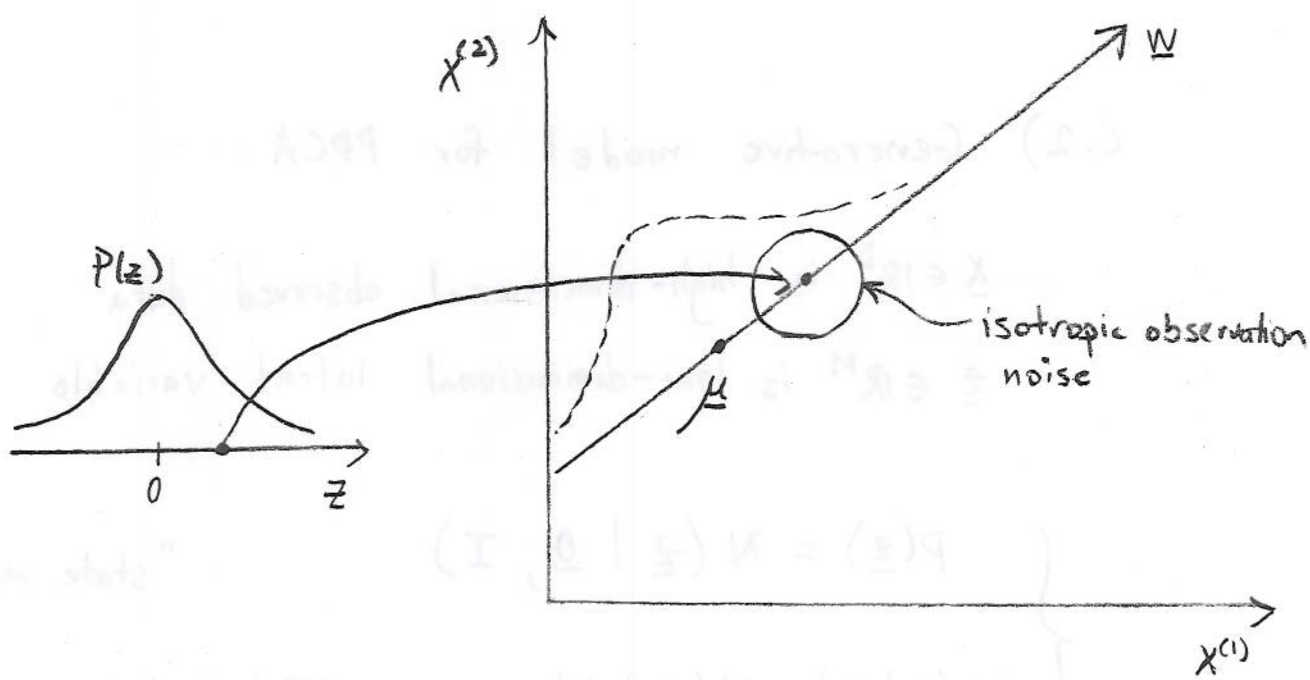
$$\begin{cases} p(\underline{z}) = N(\underline{z} \mid \underline{0}, I) & \text{"state model"} \\ p(\underline{x} \mid \underline{z}) = N(\underline{x} \mid \underbrace{W}_{D \times M \text{ matrix}} \underline{z} + \underline{\mu}, \underbrace{\sigma^2 I}_{\text{"observation noise"}}) & \text{"observation model"} \end{cases} \quad (7)$$

Relationship to PCA:

- If we fit this model to data $\underline{x}_1, \ldots, \underline{x}_N$ (i.e., we fit the model parameters $\theta = \{W, \underline{\mu}, \sigma^2\}$), the columns of W will span the principal component space (i.e., the space spanned by the columns of U_M in PCA)
- In the limit $\sigma^2 \to 0$, PPCA low-d projections approach PCA low-d projections.

Illustration of PPCA generative model:

Let $D = 2$ and $M = 1$.



To generate from PPCA model:

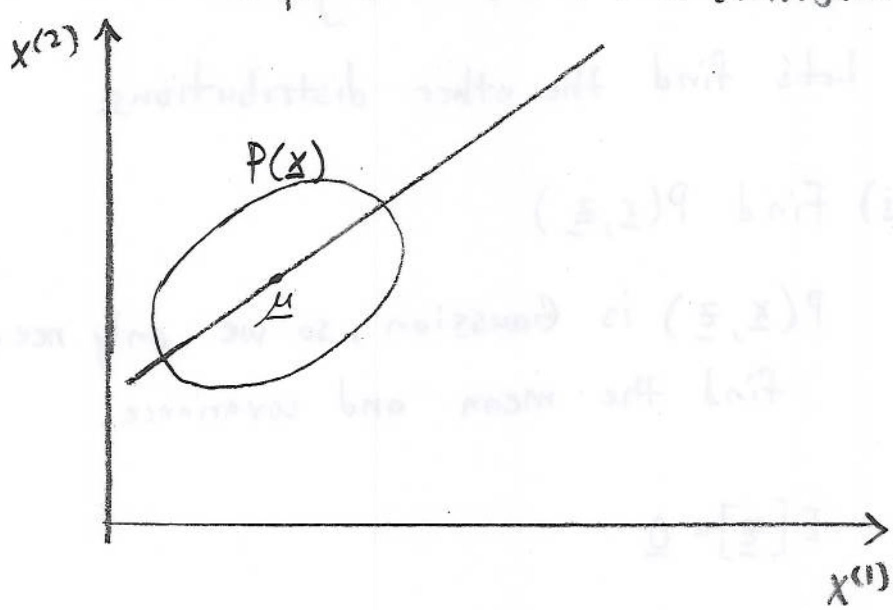
1) Draw latent variable $\underline{z} \sim N(\underline{0}, I)$.

2) Project $\underline{z}$ into high-dimensional data space

3) Add observation noise $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I)$, where $\underline{\epsilon} \in \mathbb{R}^D$.

$$\underline{x} = W\underline{z} + \underline{\mu} + \underline{\epsilon}. \tag{8}$$

If we generate many simulated data points $\underline{x}$, the data points will have a distribution $P(\underline{x})$



Notes:

More details later

- W for PPCA is closely related to U_M for PCA.
- $\underline{z}$ for PPCA is closely related to $\underline{z}$ for PCA.
- $\underline{\mu}$ for PPCA is identical to $\underline{\mu}$ for PCA.

C.3) PPCA is a linear-Gaussian model

- The variables $\underline{x}$ and $\underline{z}$ are <u>linearly</u> related (see (8))
- All marginal, conditional, and joint distributions are <u>Gaussian</u>.

$P(\underline{z})$, $P(\underline{x})$: Marginal distributions

$P(\underline{x}|\underline{z})$, $P(\underline{z}|\underline{x})$: Conditional distributions

$P(\underline{x}, \underline{z})$: Joint distribution

$P(\underline{z})$ and $P(\underline{x}|\underline{z})$ are given (see (7))

Let's find the other distributions.

i) Find $P(\underline{x}, \underline{z})$

$P(\underline{x}, \underline{z})$ is Gaussian, so we only need to find the mean and covariance.

$$E[\underline{z}] = \underline{0}$$

From (8),

$$\begin{aligned} E[\underline{x}] &= E[W\underline{z} + \underline{\mu} + \underline{\epsilon}] \\ &= W\cancel{E[\underline{z}]} + \underline{\mu} + \cancel{E[\underline{\epsilon}]} \\ &= \underline{\mu} \end{aligned}$$

$$\mathrm{cov}(\underline{z}) = I$$

$$\mathrm{cov}(\underline{x}) = E[\underline{x}\,\underline{x}^T] - E[\underline{x}]\,E[\underline{x}]^T$$

from (8)

$$= E\left[(W\underline{z} + \underline{\mu} + \underline{\epsilon})(W\underline{z} + \underline{\mu} + \underline{\epsilon})^T\right] - \underline{\mu}\,\underline{\mu}^T$$

$$= E\Big[W\underline{z}\,\underline{z}^T W^T + \cancel{\underline{\mu}\,\underline{z}^T W^T} + \cancel{\underline{\epsilon}\,\underline{z}^T W^T} +$$

$$\cancel{W\underline{z}\,\underline{\mu}^T} + \underline{\mu}\,\underline{\mu}^T + \cancel{\underline{\epsilon}\,\underline{\mu}^T} +$$

$$\cancel{W\underline{z}\,\underline{\epsilon}^T} + \cancel{\underline{\mu}\,\underline{\epsilon}^T} + \underline{\epsilon}\,\underline{\epsilon}^T \Big] - \underline{\mu}\,\underline{\mu}^T$$

$$= W\,E[\underline{z}\,\underline{z}^T]\,W^T + \cancel{\underline{\mu}\,\underline{\mu}^T} + E[\underline{\epsilon}\,\underline{\epsilon}^T] - \cancel{\underline{\mu}\,\underline{\mu}^T}$$

$$= WW^T + \sigma^2 I$$

$$\mathrm{cov}(\underline{x}, \underline{z}) = E[\underline{x}\,\underline{z}^T] - E[\underline{x}]\,\underbrace{\cancel{E[\underline{z}]}}_{0}{}^T$$

$$= E\left[(W\underline{z} + \underline{\mu} + \underline{\epsilon})\,\underline{z}^T\right]$$

$$= W\,E[\underline{z}\,\underline{z}^T] + \underline{\mu}\,\cancel{E[\underline{z}^T]} + \cancel{E[\underline{\epsilon}\,\underline{z}^T]}$$

$$= W$$

$$\begin{bmatrix} \underline{z} \\ \underline{x} \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{0} \\ \underline{\mu} \end{bmatrix}, \begin{bmatrix} I & W^T \\ W & WW^T + \sigma^2 I \end{bmatrix} \right) \qquad (9)$$

ii) Find $P(\underline{x})$

From (9),

$$\underline{x} \sim N(\underline{\mu}, WW^T + \sigma^2 I). \qquad (10)$$

iii) Find $P(\underline{z} | \underline{x})$

In general, we would use Bayes rule.

An easier way is to apply the results of conditioning for Gaussian random variables (see PRML Section 2.3.1)

If $\underline{x} = \begin{bmatrix} \underline{x}_a \\ \underline{x}_b \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{\mu}_a \\ \underline{\mu}_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \right)$,

$P(\underline{x}_a | \underline{x}_b)$ is Gaussian with the following mean and covariance.

$$E[\underline{x}_a | \underline{x}_b] = \underline{\mu}_a + \Sigma_{ab} \Sigma_{bb}^{-1} (\underline{x}_b - \underline{\mu}_b)$$

$$\mathrm{cov}(\underline{x}_a | \underline{x}_b) = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}$$

$$E[\underline{z} | \underline{x}] = \underline{0} + W^T C^{-1} (\underline{x} - \underline{\mu})$$

$$\mathrm{cov}(\underline{z} | \underline{x}) = I - W^T C^{-1} W,$$

where $C = WW^T + \sigma^2 I$

Thus,

$$\underline{z} \mid \underline{x} \sim N\left(W^T C^{-1} (\underline{x} - \underline{\mu}),\ \underbrace{I - W^T C^{-1} W}_{\text{note that covariance does not depend on } \underline{x}} \right) \quad (11)$$

C.4) EM algorithm for PPCA

Goal: Maximize $\log p(\{\underline{x}\} \mid \theta)$ w.r.t. θ,

where $\theta = \{ W, \underline{\mu}, \sigma^2 \}$.

EM will find the sample mean exactly for $\underline{\mu}$ and attempt to find W and σ^2 such that

$$\underbrace{S}_{\text{sample covariance}} \approx W W^T + \sigma^2 I.$$

This should make sense intuitively from (10). EM is trying to match the mean and covariance of the data.

For simplicity here, we will fix $\underline{\mu}$ to be the sample mean, which is the maximum likelihood solution.

E-step:

$P(\underline{z}_n \mid \underline{x}_n)$ is shown in (11).

M-step:

$$\log P(X, Z) = \sum_{n=1}^{N} \log P(\underline{x}_n, \underline{z}_n)$$

$$= \sum_{n=1}^{N} \Big(\log P(\underline{x}_n | \underline{z}_n) + \log P(\underline{z}_n) \Big)$$

$$= \sum_{n=1}^{N} \Big(-\frac{D}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2)^D - \frac{1}{2} (\underline{x}_n - W\underline{z}_n - \underline{\mu})^T (\sigma^2 I)^{-1} (\;\cdot\;)$$

$$- \frac{M}{2} \log(2\pi) - \frac{1}{2} \log 1 - \frac{1}{2} \underline{z}_n^T \underline{z}_n \Big)$$

$$Q = E_z \big[\log P(X, Z) \big]$$

$$= \sum_{n=1}^{N} \Big\{ -\frac{D}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} E\Big[\big((\underline{x}_n - \underline{\mu}) - W\underline{z}_n\big)^T \big((\underline{x}_n - \underline{\mu}) - W\underline{z}_n\big) \Big]$$

$$- \frac{M}{2} \log(2\pi) - \frac{1}{2} E[\underline{z}_n^T \underline{z}_n] \Big\}$$

$$= \sum_{n=1}^{N} \Big\{ -\frac{D}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \Big((\underline{x}_n - \underline{\mu})^T (\underline{x}_n - \underline{\mu}) - E[\underline{z}_n^T W^T (\underline{x}_n - \underline{\mu})]$$

$$- E[(\underline{x}_n - \underline{\mu})^T W \underline{z}_n] + E[\underline{z}_n^T W^T W \underline{z}_n] \Big)$$

$$- \frac{M}{2} \log(2\pi) - \frac{1}{2} E[\underline{z}_n^T \underline{z}_n] \Big\}$$

$$= \sum_{n=1}^{N} \Big\{ -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Big((\underline{x}_n-\underline{\mu})^T(\underline{x}_n-\underline{\mu}) - E[\underline{z}_n]^T W^T(\underline{x}_n-\underline{\mu})$$

$$- (\underline{x}_n-\underline{\mu})^T W E[\underline{z}_n] + Tr\big(W^T W E[\underline{z}_n \underline{z}_n^T]\big)\Big)$$

$$- \frac{M}{2}\log(2\pi) - \frac{1}{2}E[\underline{z}_n^T \underline{z}_n]\Big\}$$

$$\frac{\partial Q}{\partial W} = \sum_{n=1}^{N}\Big\{ -\frac{1}{2\sigma^2}\Big(-(\underline{x}_n-\underline{\mu})E[\underline{z}_n]^T - (\underline{x}_n-\underline{\mu})E[\underline{z}_n]^T$$

$$+ 2W\big(E[\underline{z}_n \underline{z}_n^T]\big)\Big)\Big\} = 0$$

$$W\Big(\sum_{n=1}^{N} E[\underline{z}_n \underline{z}_n^T]\Big) = \sum_{n=1}^{N}(\underline{x}_n-\underline{\mu})E[\underline{z}_n]^T$$

$$\boxed{W_{new} = \Big(\sum_{n=1}^{N}(\underline{x}_n-\underline{\mu})E[\underline{z}_n]^T\Big)\Big(\sum_{n=1}^{N}E[\underline{z}_n \underline{z}_n^T]\Big)^{-1}} \quad (12)$$

$$\frac{\partial Q}{\partial \sigma^2} = \sum_{n=1}^{N}\Big\{-\frac{D}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2}\cdot\frac{1}{(\sigma^2)^2}(\ \cdot\ )\Big\} = 0$$

$$-ND + \frac{1}{\sigma^2}\sum_{n=1}^{N}(\ \cdot\ ) = 0$$

$$\sigma^2 = \frac{1}{ND}\sum_{n=1}^{N}(\ \cdot\ )$$

We can simplify the update for σ^2 by plugging in W_{new}.

$$\sigma^2 = \frac{1}{ND}\left\{ Tr\left(\sum_{n=1}^{N}(\underline{x}_n-\underline{\mu})(\underline{x}_n-\underline{\mu})^T\right) - Tr\left(\sum_{n=1}^{N}(\underline{x}_n-\underline{\mu})E[\underline{z}_n]^T \cdot W^T\right)\right.$$

$$\left. - Tr\left(W\sum_{n=1}^{N}E[\underline{z}_n](\underline{x}_n-\underline{\mu})^T\right) + Tr\left(W\sum_{n=1}^{N}E[\underline{z}_n\underline{z}_n^T]W^T\right)\right\}$$

(13) $$\boxed{\sigma^2_{new} = \frac{1}{ND} Tr\left(\sum_{n=1}^{N}(\underline{x}_n-\underline{\mu})(\underline{x}_n-\underline{\mu})^T - W_{new}\sum_{n=1}^{N}E[\underline{z}_n](\underline{x}_n-\underline{\mu})^T\right)}$$

[See physical analogy of EM for PPCA, PRML Figure 12.12.]

[The expressions for W_{new} and σ^2_{new} should make sense intutively.

Consider the linear regression problem:

$$x_n = w z_n + \mu$$

Minimizing mean squared error with respect to w,

$$w^* = \frac{\sum_{n=1}^{N}(x_n-\mu)z_n}{\sum_{n=1}^{N} z_n^2} \qquad \text{(compare to (12))}$$

Using this w^*, minimum mean squared error is

$$\frac{1}{N}\left(\sum_{n=1}^{N}(x_n-\mu)^2 - w^*\sum_{n=1}^{N} z_n(\underline{x}_n-\underline{\mu})\right) \qquad \text{(compare to (13))}$$
]

C.5) Relating PPCA to PCA

- PC directions

The columns of W for PPCA span the <u>same</u> space as that spanned by the columns of U_M.

The difference between W and U_M is that the columns of U_M are orthonormal and ordered based on amount of variance explained. In general, the columns of W are neither orthonormal nor ordered.

To obtain U_M from W, apply the singular value decomposition (SVD) to W. The SVD is a generalization of diagonalization to non-square matrices.

$$W = \underbrace{\begin{bmatrix} | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \cdots & \underline{u}_M \\ | & | & & | \end{bmatrix}}_{\tilde{U}} \underbrace{\begin{bmatrix} d_1 & & & 0 \\ & d_2 & & \\ & & \ddots & \\ 0 & & & d_M \end{bmatrix}}_{\tilde{D}} \underbrace{\begin{bmatrix} | & | & & | \\ \underline{v}_1 & \underline{v}_2 & \cdots & \underline{v}_M \\ | & | & & | \end{bmatrix}^T}_{\tilde{V}^T}$$

$\tilde{U}$	$\tilde{D}$	$\tilde{V}^T$
$(D \times M)$	$(M \times M)$	$(M \times M)$
columns of $\tilde{U}$ orthonormal	diagonal matrix	columns of $\tilde{V}$ orthonormal
	$d_1 \geq d_2 \geq \ldots \geq d_M \geq 0$ "singular values"	

Now, $\hat{U}$ for PPCA is identical to U_M for PCA.

- Low-dimensional projections

In PPCA, the low-d projection corresponding to W is $E[\underline{z}_n \mid \underline{x}_n] = W^T C^{-1}(\underline{x}_n - \underline{\mu})$ from (11).

The same point has high-d coordinates:

$$W\, E[\underline{z}_n \mid \underline{x}_n] + \underline{\mu} \qquad (\text{from } (8))$$

$$= \hat{U} \underbrace{\hat{D} \hat{V}^T E[\underline{z}_n \mid \underline{x}_n]}_{\text{call this } \hat{\underline{z}}_n} + \underline{\mu}$$

There are several important reasons why $\hat{\underline{z}}_n$ is easier to interpret than $E[\underline{z}_n \mid \underline{x}_n]$. All of the reasons stem from the fact that the columns of $\hat{U}$ are orthonormal and ordered, while those of W are not:

i) $\hat{\underline{z}}_n$ has the same units as $\underline{x}_n$ (as in PCA)

ii) The dimensions of $\hat{\underline{z}}_n$ are ordered (as in PCA), whereas $E[\underline{z}_n \mid \underline{x}_n]$ is subject to arbitrary rotations and exchange-of-dimensions in latent space.

iii) $\hat{\underline{z}}_n$ can be easily compared to PCA low-d projection.

- How does the low-dimensional projection for PPCA ($\hat{z}_n$) relate to that for PCA $\left(U_M^T (\underline{x}_n - \underline{\mu}) \right)$?

$$\hat{\underline{z}}_n = \begin{bmatrix} \frac{\lambda_1 - \sigma^2}{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \frac{\lambda_M - \sigma^2}{\lambda_M} \end{bmatrix} U_M^T (\underline{x}_n - \underline{\mu}) \qquad (14).$$

(We won't show this here.)

In other words, PPCA projection is the PCA projection shrunk towards the origin, since

$$0 \le \frac{\lambda_i - \sigma^2}{\lambda_i} \le 1 \qquad i = 1, \ldots, M.$$

As $\sigma^2 \to 0$, PPCA becomes identical to PCA.

- Intuition for PPCA vs. PCA

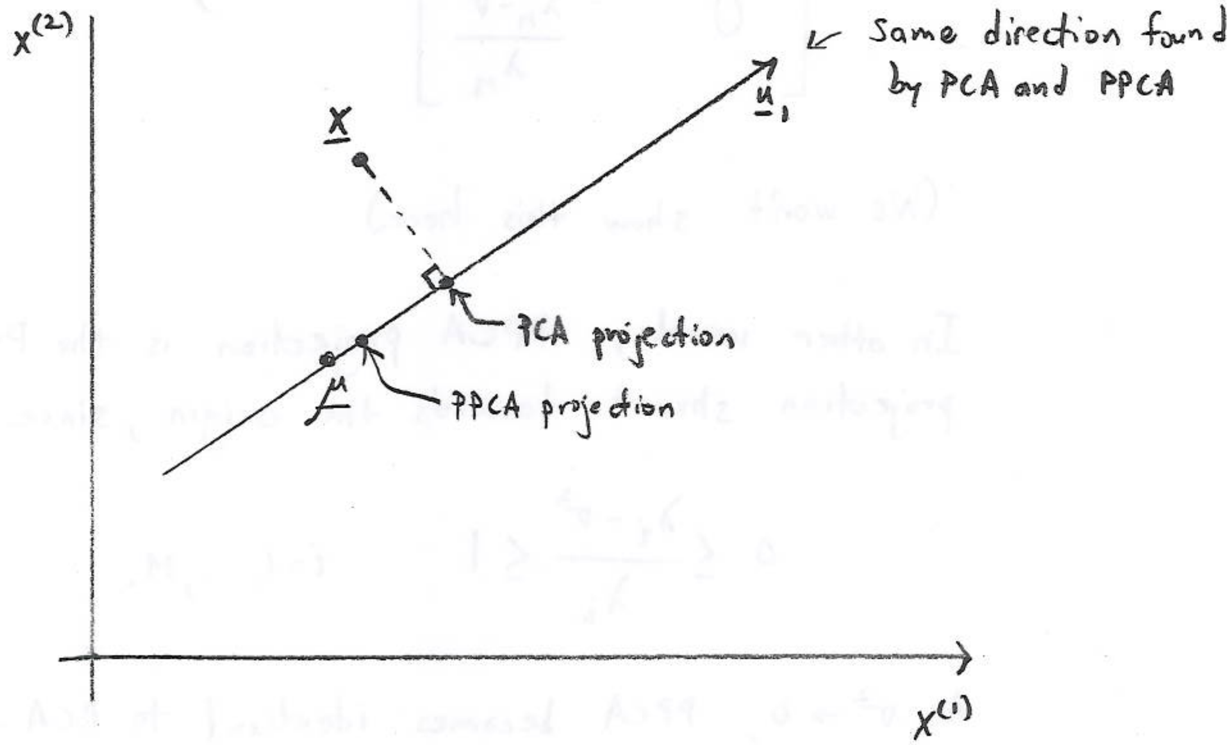
Each model tries to explain deviations of $\underline{x}$ from $\underline{\mu}$.

PPCA can explain this using a combination of the latent variable $\underline{z}$ and the observation noise $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I)$ (see (8)). But how much of each?

As σ^2 increases (more observation noise), the proportion of the deviation attributed to observation noise increases and $\underline{z}$ gets shrunk towards the mean.

PCA is the limit $\sigma^2 \to 0$, where there is no observation noise and all deviations of $\underline{x}$ from $\underline{\mu}$ must be explained using the latent variable $\underline{z}$.

In other words, PPCA is more effective at denoising the data than PCA.



C.6) Returning to advantages of PPCA over PCA

- Dimensionality M of latent space for PPCA can be selected using cross-validated likelihoods, where $P(\underline{x})$ given in (10).

- PPCA defines a constrained Gaussian in (10), that's a useful compromise between a Gaussian with diagonal covariance (overconstrained in some settings) and a Gaussian with full covariance (underconstrained in some settings).

$$\mathrm{cov}(\underline{x}) = WW^T + \sigma^2 I$$

D) Factor Analysis (FA)

D.1) Motivation: PPCA assumes isotropic observation noise. In some settings, we would like each dimension of $\underline{x}$ to have a different level of observation noise.

The only difference between FA and PPCA is that instead of $\sigma^2 I$ in (7), the observation noise covariance is a diagonal matrix Ψ.

For FA,

$$\underline{x} \sim N(\underline{\mu}, WW^T + \Psi),$$

which is similar to (10).

We can define an EM algorithm that is nearly identical to that for PPCA. After replacing all instances of $\sigma^2 I$ with Ψ, (13) becomes

$$\Psi_{new} = \frac{1}{N} \operatorname{diag}\left\{ \sum_{n=1}^{N} (\underline{x}_n - \underline{\mu})(\underline{x}_n - \underline{\mu})^T - W_{new} \sum_{n=1}^{N} E[\underline{z}_n](\underline{x}_n - \underline{\mu})^T \right\}$$

where $\operatorname{diag}\{\cdot\}$ zeroes all off-diagonal elements.

D.2) Comparing FA and PPCA

Because of the non-isotropic observation noise, FA will identify different directions in the data space compared to PCA / PPCA.

As with PPCA, we will want to orthonormalize the columns of W for interpretability.

- PCA / PPCA is invariant to rotations in the data space, whereas FA is not.

  The reason is the FA observation noise must be axis-aligned.

- FA is invariant to a component-wise rescaling of the data, whereas PCA / PPCA is not.

  The reason is that PCA / PPCA uses the same σ^2 for each component of $\underline{x}$.

Example of component-wise rescaling:

$$\underline{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \rightarrow \begin{bmatrix} 3X^{(1)} \\ \frac{1}{2}X^{(2)} \end{bmatrix}$$

Appendix

Matrix derivatives:

$$\frac{d}{dX}(\underline{a}^T X \underline{b}) = \underline{a}\,\underline{b}^T$$

$$\frac{d}{dX}(\underline{a}^T X^T \underline{b}) = \underline{b}\,\underline{a}^T$$

$$\frac{d}{dX}\mathrm{Tr}(X^T X A) = X(A + A^T)$$

Matrix inversion lemma:

Inverting $C = WW^T + \sigma^2 I$ directly in (11) is a costly $O(D^3)$ operation. Instead, can apply the matrix inversion lemma:

$$C^{-1} = \sigma^{-2} I - \sigma^{-2} W \underbrace{(\sigma^2 I + W^T W)}^{-1} W^T$$

now, matrix to be inverted is $M \times M$, which is $O(M^3)$ operation

Can use same trick for FA.