Probabilistic PCA

Yingzhen Li

Department of Computing
Imperial College London

@liyzhen2
`yingzhen.li@imperial.ac.uk`

Nov 25, 2022

PCA: Recap

Motivation: real-world data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \in \mathbb{R}^{D \times 1}$ often lies in a lower-dimensional space

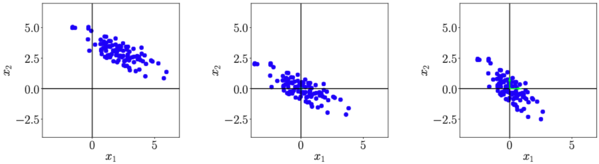
PCA's idea to "save memory":

- Project $\boldsymbol{x}_n$ onto a lower-dim space $span(\{\mathbf{b}_1, ..., \mathbf{b}_M\})$ to get

$$\boldsymbol{z}_n := (z_{n1}, ..., z_{nM}), \quad z_{nm} = \mathbf{b}_m, \quad M < D,$$

  then store $\boldsymbol{z}_n$ instead of $\boldsymbol{x}_n$;
- When needed, get reconstruction $\tilde{\boldsymbol{x}}_n = \sum_{m=1}^M z_{nm} \mathbf{b}_m$
- To get orthonormal basis $\{\mathbf{b}_1, ..., \mathbf{b}_M\}$: PCA
  - maximum variance view
  - minimum reconstruction error view

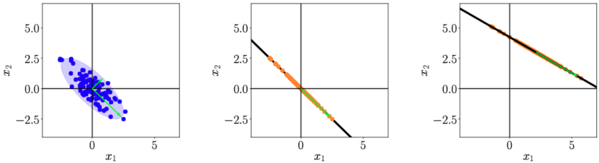
PCA: Recap



(a) Original dataset.

(b) Step 1: Centering by subtracting the mean from each data point.

(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).

(e) Step 4: Project data onto the principal subspace.

(f) Undo the standardization and move projected data back into the original data space from (a).

Fig from the MML book.

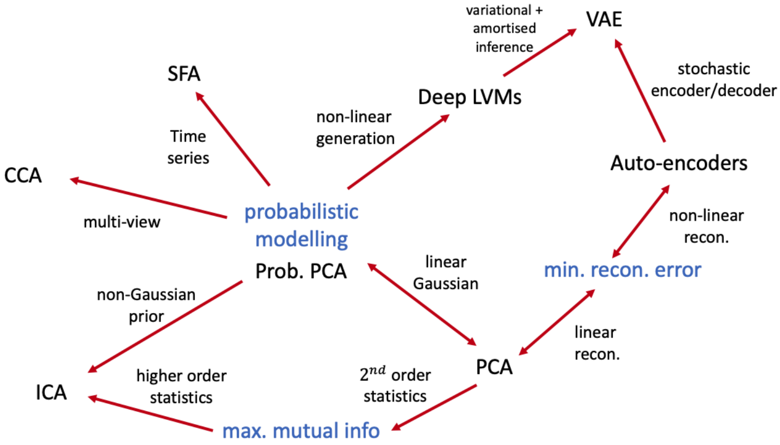
PCA: Recap

An issue with PCA in test time:

- Given an $\boldsymbol{x}$, we can find the low-dim projection $\boldsymbol{z}$ of it using trained PCA
- However, PCA alone cannot generate new $\boldsymbol{x}$ (unless we do something further)

Generative models

To name a few dimensionality reduction methods:



Generative models

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Google's Imagen text-to-image generation model

Latent Variable Models

Data distribution: $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \sim \pi(\boldsymbol{x})$
Make a generative model that generates $\boldsymbol{x}$ as follows:

$$\boldsymbol{z} \sim p_{\boldsymbol{\theta}}(\boldsymbol{z}), \quad \boldsymbol{x} \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$$

- $\boldsymbol{z}$: latent variable
- $\boldsymbol{x}$: data
- $\boldsymbol{\theta}$: model parameter to be fitted
- if $p_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx \pi(\boldsymbol{x})$, then the model can generate realistic data

Probabilistic PCA

Data distribution: $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \sim \pi(\boldsymbol{x})$
Probabilistic PCA: make a latent variable model as follows:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Sampling from this generative model:

$$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{x} = \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu} + \sigma\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

- model parameter: $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}\}$

Probabilistic PCA

Data distribution: $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \sim \pi(\boldsymbol{x})$
Probabilistic PCA: make a latent variable model as follows:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Marginal distribution:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$$

Probabilistic PCA

Data distribution: $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \sim \pi(\boldsymbol{x})$
Probabilistic PCA: make a latent variable model as follows:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Marginal distribution:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\ &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}). \end{aligned}$$

Probabilistic PCA

Data distribution: $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^N, \boldsymbol{x}_n \sim \pi(\boldsymbol{x})$
Probabilistic PCA: make a latent variable model as follows:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Fitting $\boldsymbol{\theta}$ with Maximum Likelihood Estimation (MLE):

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^N \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$$
$$= \frac{1}{N}\sum_{n=1}^N \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})$$

Probabilistic PCA

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}), \quad \boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}\}$$

Derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{\mu}$: denoting $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$

$$\frac{\partial}{\partial \boldsymbol{\mu}} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) \right)$$

Probabilistic PCA

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}), \quad \boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}\}$$

Derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{\mu}$: denoting $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) &= \frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right) \\ &= (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}\end{aligned}$$

$$\Rightarrow \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}$$

Setting $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\mu}^* = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$

Probabilistic PCA

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}), \quad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}), \quad \boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}, \sigma\}$$

Derivative of $\mathcal{L}$ w.r.t. $\mathbf{W}$: denoting $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{W}} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}) \\ =&\frac{\partial}{\partial \mathbf{W}} \left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) - \frac{1}{2} \log |\mathbf{C}| \right) \end{aligned}$$

- $\mathbf{C}$ depends on $\mathbf{W}$, so “chain rule” applies
- However, so far we’ve only learned about chain rule applied to scalars and vectors.

Probabilistic PCA

Applying chain rule: let $\mathcal{L}_n := \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$

- Chain rule for individual elements W_{kl} of $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_{i,j} \frac{\partial \mathcal{L}_n}{\partial C_{ij}} \frac{\partial C_{ij}}{\partial W_{kl}}, \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$$

- Calculate $\frac{\partial C_{ij}}{\partial W_{kl}}$:

Probabilistic PCA

Applying chain rule: let $\mathcal{L}_n := \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$

- Chain rule for individual elements W_{kl} of $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_{i,j} \frac{\partial \mathcal{L}_n}{\partial C_{ij}} \frac{\partial C_{ij}}{\partial W_{kl}}, \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$$

- Calculate $\frac{\partial C_{ij}}{\partial W_{kl}}$: notice $C_{ij} = \sum_l W_{il} W_{jl} + \sigma^2 \delta(i = j)$

$$\frac{\partial C_{ij}}{\partial W_{kl}} = \begin{cases} 0, & k \notin \{i, j\} \\ W_{jl}, & k = i \neq j \\ W_{il}, & k = j \neq i \\ 2W_{il}, & k = i = j \end{cases}$$

Probabilistic PCA

Applying chain rule: let $\mathcal{L}_n := \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$

- Chain rule for individual elements W_{kl} of $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_{i,j} \frac{\partial \mathcal{L}_n}{\partial C_{ij}} \frac{\partial C_{ij}}{\partial W_{kl}}, \quad \mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$$

- Calculate $\frac{\partial C_{ij}}{\partial W_{kl}}$: notice $C_{ij} = \sum_l W_{il} W_{jl} + \sigma^2 \delta(i = j)$

$$\frac{\partial C_{ij}}{\partial W_{kl}} = \begin{cases} 0, & k \notin \{i, j\} \\ W_{jl}, & k = i \neq j \\ W_{il}, & k = j \neq i \\ 2W_{il}, & k = i = j \end{cases}$$

- This means for fixed k, l:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_j \frac{\partial \mathcal{L}_n}{\partial C_{kj}} W_{jl} + \sum_i \frac{\partial \mathcal{L}_n}{\partial C_{ik}} W_{il}$$

Probabilistic PCA

Applying chain rule: let $\mathcal{L}_n := \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$

- Chain rule for individual elements W_{kl} of $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_j \frac{\partial \mathcal{L}_n}{\partial C_{kj}} W_{jl} + \sum_i \frac{\partial \mathcal{L}_n}{\partial C_{ik}} W_{il}$$

- Writing the derivatives of $\mathcal{L}_n$ in matrix forms:

$$\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}} = \begin{bmatrix} \frac{\partial \mathcal{L}_n}{\partial C_{11}} & \frac{\partial \mathcal{L}_n}{\partial C_{21}} & \cdots \\ \vdots & \ddots & \\ \frac{\partial \mathcal{L}_n}{\partial C_{1D}} & & \frac{\partial \mathcal{L}_n}{\partial C_{DD}} \end{bmatrix}, \quad \frac{\partial \mathcal{L}_n}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial \mathcal{L}_n}{\partial W_{11}} & \frac{\partial \mathcal{L}_n}{\partial W_{21}} & \cdots \\ \vdots & \ddots & \\ \frac{\partial \mathcal{L}_n}{\partial W_{1M}} & & \frac{\partial \mathcal{L}_n}{\partial W_{DM}} \end{bmatrix}$$

$$\Rightarrow \quad \sum_j \frac{\partial \mathcal{L}_n}{\partial C_{kj}} W_{jl} = (\mathbf{W}^\top)_{l\cdot} \left(\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}}\right)_{\cdot k}, \quad \sum_i \frac{\partial \mathcal{L}_n}{\partial C_{ik}} W_{il} = \left(\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}}\right)_{k\cdot} \mathbf{W}_{\cdot l}$$

Probabilistic PCA

Applying chain rule: let $\mathcal{L}_n := \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I})$

- Chain rule for individual elements W_{kl} of $\mathbf{W}$:

$$\frac{\partial \mathcal{L}_n}{\partial W_{kl}} = \sum_j \frac{\partial \mathcal{L}_n}{\partial C_{kj}} W_{jl} + \sum_i \frac{\partial \mathcal{L}_n}{\partial C_{ik}} W_{il}$$

- Writing the derivatives of $\mathcal{L}_n$ in matrix forms:

$$\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}} = \begin{bmatrix} \frac{\partial \mathcal{L}_n}{\partial C_{11}} & \frac{\partial \mathcal{L}_n}{\partial C_{21}} & \cdots \\ \vdots & \ddots & \\ \frac{\partial \mathcal{L}_n}{\partial C_{1D}} & & \frac{\partial \mathcal{L}_n}{\partial C_{DD}} \end{bmatrix}, \quad \frac{\partial \mathcal{L}_n}{\partial \mathbf{W}} = \begin{bmatrix} \frac{\partial \mathcal{L}_n}{\partial W_{11}} & \frac{\partial \mathcal{L}_n}{\partial W_{21}} & \cdots \\ \vdots & \ddots & \\ \frac{\partial \mathcal{L}_n}{\partial W_{1M}} & & \frac{\partial \mathcal{L}_n}{\partial W_{DM}} \end{bmatrix}$$

$$\Rightarrow \quad \frac{\partial \mathcal{L}_n}{\partial \mathbf{W}} = \mathbf{W}^\top \left(\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}} + \frac{\partial \mathcal{L}_n}{\partial \mathbf{C}}^\top \right)$$

Probabilistic PCA

$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$ is symmetric, the matrix form of the derivatives are:

$$\frac{\partial}{\partial \mathbf{C}}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) = -\mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}$$

$$\frac{\partial}{\partial \mathbf{C}} \log |\mathbf{C}| = \mathbf{C}^{-1}$$

Notice that both derivatives are symmetric matrices:

$$\frac{\partial \mathcal{L}_n}{\partial \mathbf{W}} = \mathbf{W}^\top \left(\frac{\partial \mathcal{L}_n}{\partial \mathbf{C}} + \frac{\partial \mathcal{L}_n}{\partial \mathbf{C}}^\top \right) = 2\mathbf{W}^\top \frac{\partial \mathcal{L}_n}{\partial \mathbf{C}}$$

Derivative of $\mathcal{L}$ w.r.t. $\mathbf{W}$:

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) \\
=& 2\mathbf{W}^\top \frac{\partial}{\partial \mathbf{C}} \left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) - \frac{1}{2} \log |\mathbf{C}| \right) \\
=& \mathbf{W}^\top \left(\mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} - \mathbf{C}^{-1} \right).
\end{aligned}$$

Probabilistic PCA

Derivative of $\mathcal{L}$ w.r.t. $\mathbf{W}$ with $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$:

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}} \log \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) \\
=&2\mathbf{W}^\top \frac{\partial}{\partial \mathbf{C}} \left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}) - \frac{1}{2}\log|\mathbf{C}| \right) \\
=&\mathbf{W}^\top \left(\mathbf{C}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \mathbf{C}^{-1} - \mathbf{C}^{-1} \right).
\end{aligned}
$$

$$
\Rightarrow \quad \left(\frac{\partial \mathcal{L}}{\partial \mathbf{W}}\right)^\top = \mathbf{C}^{-1}(\underbrace{\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^\top}_{:=\mathbf{S}\text{ , covariance when } \boldsymbol{\mu}=\boldsymbol{\mu}^*}\mathbf{C}^{-1} - \mathbf{I})\mathbf{W}
$$

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Probabilistic PCA

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2 \mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2 \mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{D\times D}, \Sigma \in \mathbb{R}^{D\times M}, \mathbf{V} \in \mathbb{R}^{M\times M}$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \sigma_M \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix} \Rightarrow \Sigma\Sigma^\top + \sigma^2 \mathbf{I} = \begin{bmatrix} \sigma_1^2 + \sigma^2 & 0 & \cdots & & & \\ 0 & \ddots & & & & \\ \vdots & & \sigma_M^2 + \sigma^2 & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ & & & & & \sigma^2 \end{bmatrix}$$

Probabilistic PCA

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2\mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{D\times D}, \Sigma \in \mathbb{R}^{D\times M}, \mathbf{V} \in \mathbb{R}^{M\times M}$

$$\begin{aligned}
\mathbf{S}(\mathbf{U}\Sigma\Sigma^\top\mathbf{U}^\top + \sigma^2\mathbf{I})^{-1}\mathbf{U}\Sigma\mathbf{V}^\top &= \mathbf{U}\Sigma\mathbf{V}^\top \\
\Rightarrow \quad \mathbf{S}(\mathbf{U}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})\mathbf{U}^\top)^{-1}\mathbf{U}\Sigma\mathbf{V}^\top &= \mathbf{U}\Sigma\mathbf{V}^\top \\
\Rightarrow \quad \mathbf{S}\mathbf{U}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\mathbf{U}^\top\mathbf{U}\Sigma\mathbf{V}^\top &= \mathbf{U}\Sigma\mathbf{V}^\top \\
\Rightarrow \quad \mathbf{S}\mathbf{U}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\Sigma\mathbf{V}^\top &= \mathbf{U}\Sigma\mathbf{V}^\top \\
\Rightarrow \quad \mathbf{S}\mathbf{U}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\Sigma &= \mathbf{U}\Sigma
\end{aligned}$$

Probabilistic PCA

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2\mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2\mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{D\times D}, \Sigma \in \mathbb{R}^{D\times M}, \mathbf{V} \in \mathbb{R}^{M\times M}$

$$\mathbf{SU}\begin{bmatrix} (\sigma_1^2+\sigma^2)^{-1}\sigma_1 & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & (\sigma_M^2+\sigma^2)^{-1}\sigma_M \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix} = \mathbf{U}\begin{bmatrix} \sigma_1 & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \sigma_M \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix}$$

Probabilistic PCA

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2 \mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2 \mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{D\times D}, \Sigma \in \mathbb{R}^{D\times M}, \mathbf{V} \in \mathbb{R}^{M\times M}$

$$\mathbf{S}\mathbf{U}\begin{bmatrix} 1 & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & 1 \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix} = \mathbf{U}\begin{bmatrix} \sigma_1^2 + \sigma^2 & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \sigma_M^2 + \sigma^2 \\ & & 0 \\ & & \vdots \\ & & 0 \end{bmatrix}$$

Write $\mathbf{U} := (\boldsymbol{u}_1, ..., \boldsymbol{u}_D)$:

$$(\mathbf{S}\boldsymbol{u}_1, ..., \mathbf{S}\boldsymbol{u}_M) = ((\sigma_1^2 + \sigma^2)\boldsymbol{u}_1, ..., (\sigma_M^2 + \sigma^2)\boldsymbol{u}_M)$$

$\Rightarrow \quad$ the first M columns of $\mathbf{U}$ contain eigenvectors of $\mathbf{S}$!

Probabilistic PCA

Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2 \mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2\mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$,
   $\mathbf{U} \in \mathbb{R}^{D \times D}, \Sigma \in \mathbb{R}^{D \times M}, \mathbf{V} \in \mathbb{R}^{M \times M}$
   $\Rightarrow \quad \mathbf{S}\mathbf{U}(\Sigma\Sigma^\top + \sigma^2\mathbf{I})^{-1}\Sigma = \mathbf{U}\Sigma$
   Then given $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$, $\mathbf{Q} = (\boldsymbol{q}_1, ..., \boldsymbol{q}_D)$, $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$,

   $$\mathbf{U} := (\boldsymbol{u}_1, ..., \boldsymbol{u}_D), \boldsymbol{u}_m = \boldsymbol{q}_{i_m}, 1 \leqslant i_m \leqslant D, m = 1, ..., M$$

   ($\mathbf{U}$ can contain any other columns for $\boldsymbol{u}_{M+1}$ to $\boldsymbol{u}_D$)

Probabilistic PCA

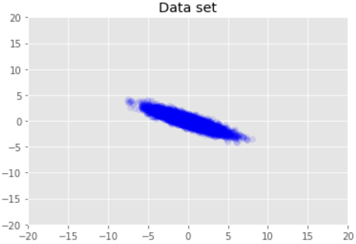
Setting $\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{0} \quad \Rightarrow \quad \mathbf{W}^*$ satisfies $\mathbf{S}(\mathbf{W}^*(\mathbf{W}^*)^\top + \sigma^2 \mathbf{I})^{-1}\mathbf{W}^* = \mathbf{W}^*$

Possible solutions for the fixed points:

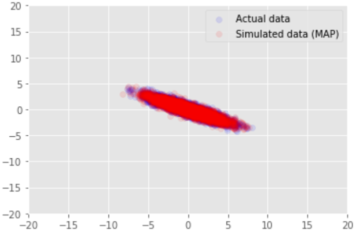
1. $\mathbf{W}^* = \mathbf{0}$ (then $p_{\boldsymbol{\theta}^*}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^*, \sigma^2 \mathbf{I})$, not interesting)
2. Lets write down the SVD of $\mathbf{W}^*$ and assume $\mathbf{W}^* = \mathbf{U}\Sigma\mathbf{V}^\top$,
   $\mathbf{U} \in \mathbb{R}^{D \times D}, \Sigma \in \mathbb{R}^{D \times M}, \mathbf{V} \in \mathbb{R}^{M \times M}$
   Then given $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$, $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$
   - **Exercise:** For $m = 1, ..., M$, $\Sigma_{mm} = \sqrt{\lambda_{i_m} - \sigma^2}$ if $\boldsymbol{u}_m = \boldsymbol{q}_{i_m}$
   - **Exercise:** Global maximum: $\boldsymbol{u}_m = \boldsymbol{q}_m$ for $mi = 1, ..., M$

   $\Rightarrow$ picking the M principal components (like PCA)

Probabilistic PCA



Dataset



Generate data with Prob. PCA

https://www.tensorflow.org/probability/examples/Probabilistic_PCA

Extensions of Probabilistic PCA

Probabilistic PCA: make a latent variable model as follows:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

From Probabilistic PCA to other interesting generative models:

- **Factor analysis**: change conditional output covariance from $\sigma^2\mathbf{I}$ to Ψ (a learnable diagonal matrix)
- **Generator for a VAE**: change conditional output mean from $\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}$ to $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{z})$ (See Deep Learning course next term)
- **Training**: (variational) expectation maximisation (See Probabilistic Inference course next term)

Summary

Probabilistic PCA

- One of the simplest generative model (linear generator)
- Optimal solution closely related to PCA

One more exercise for you if you have time:
Derive the posterior $p_{\boldsymbol{\theta}^*}(\boldsymbol{z}|\boldsymbol{x})$ using the optimal $\boldsymbol{\theta}^* = \{\boldsymbol{\mu}^*, \mathbf{W}^*\}$