Lecture 5: Linear Gaussian Latent Variable Models

STATS305C: Applied Statistics III

Scott Linderman

April 11, 2022

Last Time...

- Directed Graphical Models
- Hierarchical Gaussian Model
- MCMC: MH and Gibbs

Today...

Outline:

- Principal Components Analysis (PCA)
- PCA as a linear autoencoder
- PCA as a linear Gaussian latent variable model
- Factor analysis
- Other continuous latent variable models

Reading:

- Required: Bishop, Ch 12

Motivating Example: More scores

Continuing on our academic theme, suppose we have not only SAT scores but entire transcripts, $\{\boldsymbol{x}_n\}_{n=1}^N$ where $\boldsymbol{x}_n \in \mathbb{R}^D$ is a vector of D grades from one student. For example, we might have grades for all their classes throughout the four years of high school. (Assume all students took the same D classes.)

We might have a few objectives in mind:

- **Dimensionality reduction:** are there a few dimensions along which the students primarily vary? E.g. do students vary along a "mathy" to "artsy" axis? Are there "late bloomers" and "senioritis sufferers"?
- **Visualization:** Like above, but how can we embed these points in 2 or 3 dimensions to best visualize them?
- **Compression:** Like above, but how can I best summarize the D scores if I am willing to sacrifice some reconstruction accuracy?

Principal Components Analysis (PCA)

Two classical definitions:

maximum variance formulation

1. An orthogonal projection of the data onto a lower dimensional linear space, known as the *principal subspace*, such that the variance of the projected data is maximized (Hotelling, 1933).

2. The linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections (Pearson, 1901).

minimum-error formulation

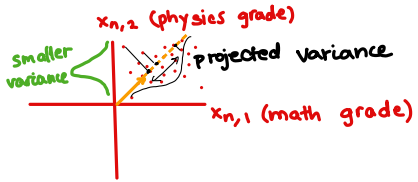
(Quoted from Bishop, Ch 12)

3. Probabilistic PCA
(Bishop, 1999)

PCA: Maximum Variance Formulation

Goal: Project data $\{\boldsymbol{x}_n\}_{n=1}^N$ onto a lower dimensional space of dimension $M < D$ while maximizing the variance of the projected data.

Illustration:



optimizing variance in best direction possible

→ start from 1D subspace & go to M-subspace

→ use constraint that new direction is orthogonal to previous directions

PCA: Maximum Variance Formulation II

To start, assume $M = 1$. The principal subspace is defined by a unit vector $\boldsymbol{u}_1 \in \mathbb{R}^D$. This is called the first **principal component**.

Projecting a data point $\boldsymbol{x}_n$ onto this subspace amounts to taking an inner product, $\boldsymbol{u}_1^\top \boldsymbol{x}_n$. These is variously called the **scores**, **embeddings**, or **signals**.

PCA: Maximum Variance Formulation III

The mean of the projected data is, 1st principal component.

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}_1^\top \boldsymbol{x}_n = \boldsymbol{u}_1^\top\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\right) = \boldsymbol{u}_1^\top \bar{\boldsymbol{x}}, \tag{1}$$

where $\bar{\boldsymbol{x}}$ is the sample mean.

The variance is

$$\frac{1}{N}\sum_{n=1}^{N}\left[\boldsymbol{u}_1^\top \boldsymbol{x}_n - \boldsymbol{u}_1^\top \bar{\boldsymbol{x}}\right]^2 = \frac{1}{N}\sum_{n=1}^{N}\left[\boldsymbol{u}_1^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})\right]^2 \tag{2}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u}_1^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top \boldsymbol{u}_1 \tag{3}$$

$$= \boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 \tag{4}$$

want to find u_1 to maximize this quantity

where $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top \in \mathbb{R}^{D\times D}$ is the sample covariance matrix.

PCA: Maximum Variance Formulation IV

Now maximize the projected variance wrt $\boldsymbol{u}_1$, subject to the constraint that $\boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$

$$\mathscr{L}(\boldsymbol{u}_1) = \boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1). \tag{5}$$

Taking the gradient wrt $\boldsymbol{u}_1$ and setting to zero,

$$\nabla\mathscr{L}(\boldsymbol{u}_1) = 2\boldsymbol{S}\boldsymbol{u}_1 - 2\lambda_1\boldsymbol{u}_1 = 0 \Rightarrow \boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1. \tag{6}$$

So $\boldsymbol{u}_1$ must be an eigenvector of $\boldsymbol{S}$. Left multiplying by $\boldsymbol{u}^\top$,

$$\boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 = \lambda_1, \tag{7}$$

so the projected variance $\boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1$ is maximized when we choose $\boldsymbol{u}_1$ to be the eigenvector with the largest eigenvalue λ_1.

eigenvalues ≥ 0

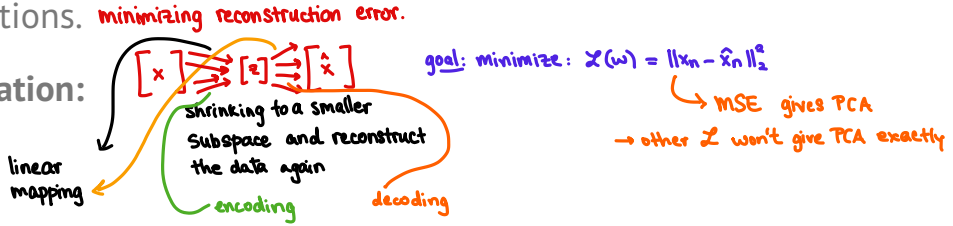
eigenvectors can be orthonormal

More generally, to find an M dimensional principal subspace, take the M eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ with the largest eigenvalues $\lambda_1, \ldots, \lambda_M$.

PCA: Linear Autoencoder Formulation

Now consider Pearson's formulation of PCA as the linear projection that minimizes the average projection cost, defined as the mean squared distance between the data points and their projections.

Illustration:



PCA: Linear Autoencoder Formulation II

To formalize this, let

$$\boldsymbol{W} = \begin{bmatrix} | & & | \\ \boldsymbol{w}_1 & \cdots & \boldsymbol{w}_M \\ | & & | \end{bmatrix} \in \mathbb{R}^{D \times M} \quad \text{with} \quad \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I} \tag{8}$$

be an orthogonal basis for the principal subspace.

We will **encode** each data point by subtracting the mean and projecting onto the principal subspace to obtain $\boldsymbol{z}_n = \boldsymbol{W}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})$.

Since $\boldsymbol{W}$ is an orthogonal matrix, all we need to do to **decode** the encoded data point is multiply $\boldsymbol{W}\boldsymbol{z}_n$ and add back the mean. That gives us,

$$\hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{z}_n + \bar{\boldsymbol{x}} \implies \hat{\boldsymbol{x}}_n = \boldsymbol{W}\boldsymbol{W}^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) + \bar{\boldsymbol{x}}. \tag{9}$$

PCA: Linear Autoencoder Formulation III

Goal: Find an orthogonal matrix $\boldsymbol{W}$ that minimizes the squared reconstruction error,

$$\mathscr{L}(\boldsymbol{W}) = \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n\|_2^2 \tag{10}$$

We can write this in matrix notation instead. Let $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ be the **centered data matrix** with rows $\boldsymbol{x}_n - \bar{\boldsymbol{x}}$. Then,

$$\mathscr{L}(\boldsymbol{W}) = \frac{1}{N}\mathrm{Tr}[(\boldsymbol{X} - \hat{\boldsymbol{X}})^\top(\boldsymbol{X} - \hat{\boldsymbol{X}})] \tag{11}$$

$$= \frac{1}{N}\mathrm{Tr}[(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^\top)^\top(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{W}^\top)] \tag{12}$$

$$= \frac{1}{N}\mathrm{Tr}[(\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top))^\top(\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top))] \tag{13}$$

$$= \mathrm{Tr}[(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)^\top\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)] \tag{14}$$

where $\boldsymbol{S} = \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}$ is the sample covariance matrix.

PCA: Linear Autoencoder Formulation IV

Now apply the circular trace property,

$$\mathscr{L}(\boldsymbol{W}) = \mathrm{Tr}[\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)^\top] \tag{15}$$

tr(ABC) = tr(BAC) = tr(BCA)...

Question: What does $(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)^\top$ equal?

$= \underbrace{I - WW^T}$
a projection matrix onto the orthogonal complement of the subspace spanned by W's columns.

PCA: Linear Autoencoder Formulation V

Thus, the objective simplifies to,

$$\mathscr{L}(\boldsymbol{W}) = \mathrm{Tr}[\boldsymbol{S}(\boldsymbol{I} - \boldsymbol{W}\boldsymbol{W}^\top)] = \mathrm{Tr}[\boldsymbol{S}] - \mathrm{Tr}[\boldsymbol{S}\boldsymbol{W}\boldsymbol{W}^\top] = \mathrm{const} - \mathrm{Tr}[\boldsymbol{W}^\top\boldsymbol{S}\boldsymbol{W}]. \tag{16}$$

Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the eigendecomposition of $\boldsymbol{S}$. (Since it is a covariance matrix, the eigenvectors are orthogonal.) Plugging in,

$$\mathscr{L}(\boldsymbol{W}) = \mathrm{const} - \mathrm{Tr}[\boldsymbol{W}^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\boldsymbol{W}] \tag{17}$$

$$= \mathrm{const} - \mathrm{Tr}\left[\boldsymbol{W}^\top\left(\sum_{d=1}^{D}\lambda_d\boldsymbol{u}_d\boldsymbol{u}_d^\top\right)\boldsymbol{W}\right] \tag{18}$$

$$= \mathrm{const} - \sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_d\boldsymbol{w}_m^\top\boldsymbol{u}_d\boldsymbol{u}_d^\top\boldsymbol{w}_m \tag{19}$$

$$= \mathrm{const} - \sum_{m=1}^{M}\sum_{d=1}^{D}\lambda_d(\boldsymbol{w}_m^\top\boldsymbol{u}_d)^2 \tag{20}$$

Question: We want to minimize $\mathscr{L}(\boldsymbol{W})$ subject to $\boldsymbol{W}$ being orthogonal. What is the solution?

PCA and the Singular Value Decomposition

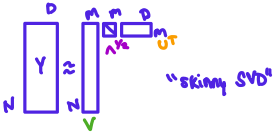
We've seen two formulations of PCA, both showing us that the first M **principal components** are the leading M **eigenvectors of the covariance matrix**.

Let

$$\boldsymbol{Y} = \frac{1}{\sqrt{N}}\boldsymbol{X} = \frac{1}{\sqrt{N}}\begin{bmatrix} - & (\boldsymbol{x}_1 - \bar{\boldsymbol{x}})^\top & - \\ & \vdots & \\ - & (\boldsymbol{x}_N^\top - \bar{\boldsymbol{x}})^\top & - \end{bmatrix} \tag{21}$$

be the **centered and scaled** data matrix. Then $\boldsymbol{Y}^\top\boldsymbol{Y} = \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{S}$ is the covariance matrix.

The **singular value decomposition (SVD)** of $\boldsymbol{Y}$ is,



$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^\top. \tag{22}$$

I.e. the **right singular vectors** of $\boldsymbol{Y}$ are the same (up to sign flips) as the eigenvectors of $\boldsymbol{S}$, and **singular values** of $\boldsymbol{Y}$ are the square root of the eigenvalues of $\boldsymbol{S}$.

PCA Explained Variance

How well do the M principal components explain the data?

Let $\boldsymbol{z}_n = \boldsymbol{U}_m^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \in \mathbb{R}^M$. Its covariance is,

$$\mathrm{Cov}[\boldsymbol{z}] = \mathrm{Cov}[\boldsymbol{U}_m^\top (\boldsymbol{x} - \bar{\boldsymbol{x}})] = \boldsymbol{U}_m^\top \mathrm{Cov}[\boldsymbol{x}] \boldsymbol{U}_m = \mathrm{diag}([\lambda_1, \ldots, \lambda_M]). \tag{23}$$

Of course, if we let $M = D$, then we have $\mathrm{Cov}(\boldsymbol{z}) = \mathrm{diag}([\lambda_1, \ldots, \lambda_D])$.

One way of assessing how well M components fits the data is via the **fraction of variance explained**,

$$\text{variance explained} = \frac{\mathrm{Tr}(\mathrm{Cov}[z; M \text{ components}])}{\mathrm{Tr}(\mathrm{Cov}[z; D \text{ components}])} = \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{D} \lambda_m} \in [0, 1]. \tag{24}$$

Scree Plots

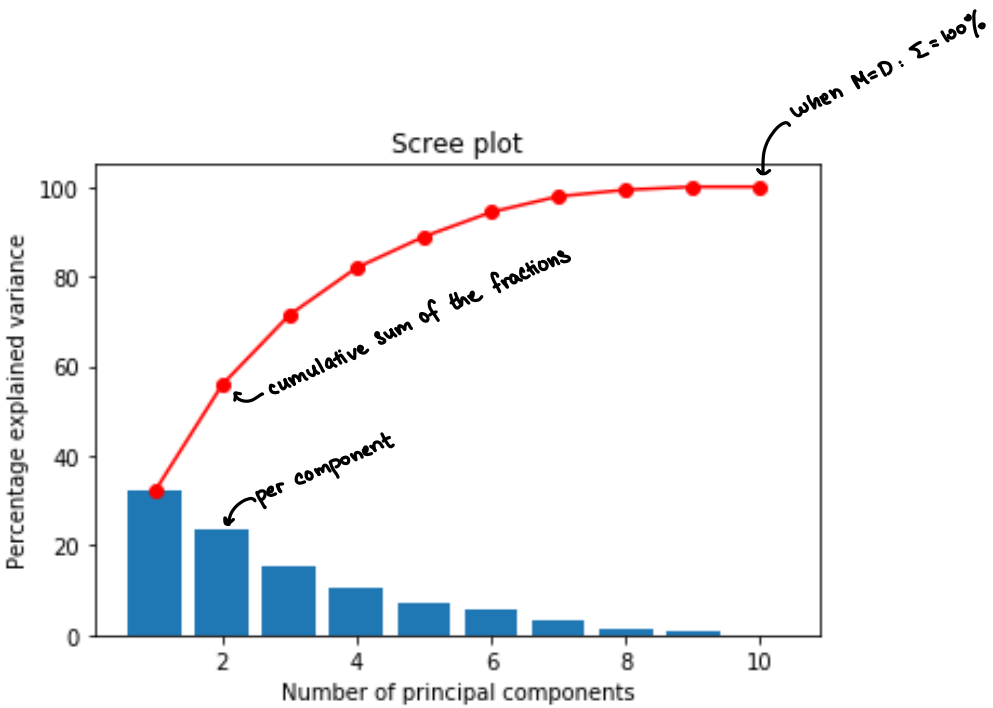


Figure: "Scree" plot showing percent variance explained per component and cumulatively.

Probabilistic PCA: A Continuous Latent Variable Model

The previous two formulations cast the principal components as the solutions to optimization problems: maximize the projected variance or minimize the reconstruction error.

A more modern view of PCA is as the **maximum likelihood estimate** of a **latent variable model**.

Probabilistic PCA (PPCA) has many advantages:

- It's a multivariate normal model with **low-rank plus diagonal covariance**, which takes only $O(MD)$ parameters.
- We can fit the model using a **host of inference algorithms**.
- It can handle **missing data**.
- We can obtain posterior distributions of the principal components and scores.
- It can be embedded in larger probabilistic models.

Probabilistic PCA: A Continuous Latent Variable Model

The PPCA model is quite simple,

$$\boldsymbol{z}_n \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{25}$$

$$\boldsymbol{x}_n \mid \boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}), \tag{26}$$

where $\boldsymbol{z}_n \in \mathbb{R}^M$ is a latent variable, $\boldsymbol{W} \in \mathbb{R}^{D \times M}$ are the weights, $\boldsymbol{\mu} \in \mathbb{R}^D$ is the bias parameter, and $\sigma^2 \in \mathbb{R}_+$ is a variance.

Equivalently, we can think of $\boldsymbol{x}_n$ as a linear function of $\boldsymbol{z}_n$ with additive noise,

$$\boldsymbol{x}_n = \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \tag{27}$$

where $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}) \in \mathbb{R}^D$.

$E(x_n) = W \cdot 0 + \mu + 0 = \mu$

$cov(x_n) = W \cdot cov(z) \cdot W^T + cov(\varepsilon_n)$; $z_n \perp\!\!\!\perp \varepsilon_n$

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

Maximum likelihood estimation of the parameters

Suppose we only need a **point estimate** of the parameters $\boldsymbol{W}$, $\boldsymbol{\mu}$, and σ^2.

A natural approach is the **maximum likelihood estimate** (MLE),

$$\boldsymbol{W}_{\text{ML}}, \boldsymbol{\mu}_{\text{ML}}, \sigma^2_{\text{ML}} = \arg\max \mathscr{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2), \tag{28}$$

where $\mathscr{L}$ is the marginal likelihood,

$$\mathscr{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \log p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \tag{29}$$

$$= \log \int \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2)\, p(\boldsymbol{z}_n)\, \mathrm{d}\boldsymbol{z}_n \tag{30}$$

likelihood, prior

$$= \log \int \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})\, \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{0}, \boldsymbol{I})\, \mathrm{d}\boldsymbol{z}_n \tag{31}$$

Exercise: Simplify this expression.

$$= \log \prod_{n=1}^{N} \int \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \cdot \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{0}, \boldsymbol{I})\, \mathrm{d}\boldsymbol{z}_n$$

$$= \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} + \sigma^2 \boldsymbol{I})$$

Maximum likelihood estimation of the parameters II

The log likelihood simplifies to,

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) - \frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{C}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{32}$$

where $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^\top + \sigma^2\boldsymbol{I}$.

Setting the derivative wrt $\boldsymbol{\mu}$ to zero and solving yields $\boldsymbol{\mu}_{\text{ML}} = \bar{\boldsymbol{x}}$, the sample mean.

Maximizing wrt $\boldsymbol{W}$ and σ^2 is more complex but still has a closed form solution, (Bishop + Tipping, 1999)

$$\boldsymbol{W}_{\text{ML}} = \boldsymbol{U}_M(\boldsymbol{\Lambda}_M - \sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{R}, \tag{33}$$

where $\boldsymbol{U}_M \in \mathbb{R}^{D\times M}$ has columns given by the leading eigenvectors of the sample covariance matrix $\boldsymbol{S}$, where $\boldsymbol{\Lambda}_M = \text{diag}([\lambda_1, \ldots, \lambda_M])$, and where $\boldsymbol{R} \in \mathbb{R}^{M\times M}$ is an arbitrary *orthogonal* matrix.

Put differently, the MLE weights are only identifiable up to orthogonal transformation. Or, only the subspace spanned by $\boldsymbol{U}_M$ is identifiable.

Maximum likelihood estimation of the parameters III

Finally, the MLE of the variance is,

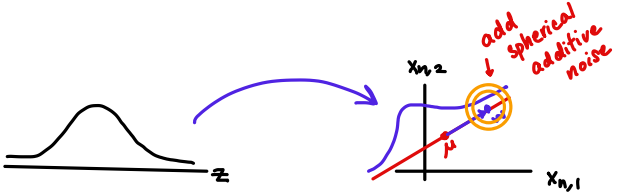
$$\sigma^2_{\text{ML}} = \frac{1}{D-M} \sum_{m=M+1}^{D} \lambda_m, \tag{34}$$

the average variance in the remaining dimensions. $WW^T + \sigma^2 I = U_m \Lambda_m U_m^T$

Question: What is the marginal covariance $\boldsymbol{C}$ using the MLE $\boldsymbol{W}_{\text{ML}}$ and σ^2_{ML}?

Question: Intuitively, why is the marginal covariance invariant to rotations of the weights?

→ RR^T cancels out.



The Posterior Distribution on the Latent Variables

Now fix $\boldsymbol{W}$, $\boldsymbol{\mu}$, and σ^2 (e.g. to their maximum likelihood values). What is the posterior of $\boldsymbol{z}_n$?

$$p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \propto \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{0}, \boldsymbol{I})\, \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \tag{35}$$

$$\propto \exp\left\{-\frac{1}{2}\boldsymbol{z}_n^\top \boldsymbol{z}_n - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu})^\top (\sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu})\right\} \tag{36}$$

$$\propto \exp\left\{-\frac{1}{2}\boldsymbol{z}_n^\top \boldsymbol{J}_n \boldsymbol{z}_n + \boldsymbol{h}_n^\top \boldsymbol{z}_n\right\} \tag{37}$$

$$\tag{38}$$

where $\boldsymbol{J}_n = \boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{W}^\top \boldsymbol{W}$ and $\boldsymbol{h}_n = \frac{1}{\sigma^2}\boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu})$

Completing the square,

$$p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{J}_n^{-1}\boldsymbol{h}_n, \boldsymbol{J}_n^{-1}). \tag{39}$$

The Posterior Distribution in the Zero Noise Limit

In the limit where $\sigma^2 \to 0$, the posterior mean of $\boldsymbol{z}_n$ is,

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2] = \lim_{\sigma^2 \to 0} (\boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{W}^\top \boldsymbol{W})^{-1} [\frac{1}{\sigma^2} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu})] \tag{40}$$

$$= \lim_{\sigma^2 \to 0} (\sigma^2 \boldsymbol{I} + \boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{41}$$

$$= (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{42}$$

Now suppose $\boldsymbol{W} = \boldsymbol{W}_{\text{ML}} = \boldsymbol{U}_M (\boldsymbol{\Lambda}_M - \sigma^2 \boldsymbol{I})^{\frac{1}{2}} \boldsymbol{R}$ and set $\boldsymbol{R} = \boldsymbol{I}$. This goes to $\boldsymbol{W} = \boldsymbol{U}_M \boldsymbol{\Lambda}_M^{\frac{1}{2}}$. Then,

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2] = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{43}$$

$$= \boldsymbol{\Lambda}_M^{-\frac{1}{2}} \boldsymbol{U}_M^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{44}$$

This is the same as $\boldsymbol{z}_n$ from the linear autoencoder formulation, except here the $\boldsymbol{z}_n$'s are scaled by $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ to be unit variance in all dimensions.

Gibbs Sampling for Probabilistic PCA I

For simplicity, assume the data is centered and fix $\mu = 0$. Now let's put a prior on the parameters,

$$p(\boldsymbol{W}, \sigma^2) = \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2) \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{w}_d \mid \boldsymbol{0}, \tfrac{\sigma^2}{\kappa_0} \boldsymbol{I}), \tag{45}$$

where $\boldsymbol{w}_d \in \mathbb{R}^M$ are the *rows* of $\boldsymbol{W}$.

Under this prior, the complete conditional distribution of the parameters is,

$$p(\boldsymbol{w}_d \mid \{\boldsymbol{x}_n, \boldsymbol{z}_n\}_{n=1}^N, \sigma^2, \kappa_0) \propto \mathcal{N}(\boldsymbol{w}_d \mid \boldsymbol{0}, \tfrac{\sigma^2}{\kappa_0} \boldsymbol{I}) \prod_{n=1}^{N} \mathcal{N}(x_{n,d} \mid \boldsymbol{w}_d^\top \boldsymbol{z}_n, \sigma^2) \tag{46}$$

$$= \mathcal{N}(\boldsymbol{w}_d \mid \boldsymbol{J}_d^{-1} \boldsymbol{h}_d, \boldsymbol{J}_d^{-1}) \tag{47}$$

d^{th} coordinate of the n^{th} data point

where $\boldsymbol{J}_d = \frac{\kappa_0}{\sigma^2} \boldsymbol{I} + \frac{1}{\sigma^2} \sum_{n=1}^N \boldsymbol{z}_n \boldsymbol{z}_n^\top$ and $\boldsymbol{h}_d = \frac{1}{\sigma^2} \sum_{n=1}^N x_{n,d} \boldsymbol{z}_n$.

Gibbs Sampling for Probabilistic PCA II

For the variance,

$$p(\sigma^2 \mid \{\boldsymbol{x}_n, \boldsymbol{z}_n\}_{n=1}^N \boldsymbol{W}, \kappa_0, \nu_0, \sigma_0^2) \propto \chi^{-2}(\sigma^2 \mid \nu_0, \sigma_0^2) \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{w}_d \mid \boldsymbol{0}, \tfrac{\sigma^2}{\kappa_0} \boldsymbol{I}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n, \sigma^2 \boldsymbol{I}) \tag{48}$$

$$\propto \chi^{-2}(\sigma^2 \mid \nu_N, \sigma_N^2) \tag{49}$$

where

$$\nu_N = \nu_0 + DM + DN \tag{50}$$

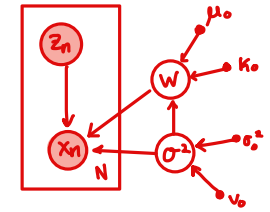
$$\sigma_N^2 = \nu_N^{-1} \left[\nu_0 \sigma_0^2 + \kappa_0 \sum_{d=1}^{D} \boldsymbol{w}_d^\top \boldsymbol{w}_d + \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n)^\top (\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n) \right] \tag{51}$$

outer products!

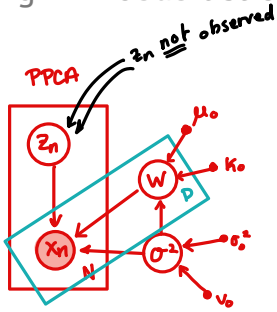
Note: It's a little strange to put a prior on the rows of $\boldsymbol{W}$; it's more natural to put a prior on the columns. We only did this for simplicity. It turns out you can put a conjugate prior on $(\boldsymbol{W}, \sigma^2)$ that specifies the conditional variance of the columns. It's called a **matrix normal inverse Wishart** prior.

Connection to Bayesian Linear Regression

Question: How does Gibbs sampling in Probabilistic PCA relate to Bayesian linear regression from HW1?



Bayesian linear regression



Factor Analysis

Factor analysis is another continuous latent variable model. In fact, it's almost the same as probabilistic PCA!

The difference is that FA allows σ^2 to vary across output dimensions. The generative model is,

$$\boldsymbol{z}_n \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{52}$$

$$\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}^2)) \tag{53}$$

where $\boldsymbol{\sigma}^2 = [\sigma_1^2, \ldots, \sigma_D^2]^\top$.

Put a similar prior on the parameters as before,

$$p(\boldsymbol{W}, \boldsymbol{\sigma}^2) = \prod_{d=1}^{D} \Big[\chi^{-2}(\sigma_d^2 \mid \nu_0, \sigma_0^2)\, \mathcal{N}(\boldsymbol{w}_d \mid \mathbf{0}, \tfrac{\sigma_d^2}{\kappa_0} \boldsymbol{I}) \Big]. \tag{54}$$

Exercise: without doing any math, derive a Gibbs sampler for this factor analysis model.

Independent Components Analysis (PCCA)

Independent Components Analysis (ICA) is yet another linear latent variable model. It aims to find *independent* factors of variation in the data. One probabilistic formulation is,

$$\boldsymbol{z}_n \overset{\text{iid}}{\sim} p(\boldsymbol{z}) \quad \text{(non Gaussian prior)} \tag{55}$$

$$\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \tag{56}$$

where the prior distribution on $\boldsymbol{z}$ assumes the coordinates are independent,

$$p(\boldsymbol{z}) = \prod_{m=1}^{M} p(z_m). \tag{57}$$

The success of this approach requires that $p(\boldsymbol{z})$ be non-Gaussian; otherwise, we could always reduce it to factor analysis by moving any correlation in $p(\boldsymbol{z})$ into the weights.

Instead, we typically choose priors that have **heavy tails**, like a **Laplace distribution**.

Probabilistic Canonical Correlations Analysis

Now consider a slightly different setting in which we have two types of observations, $\boldsymbol{x} \in \mathbb{R}^{D_x}$ and $\boldsymbol{y} \in \mathbb{R}^{D_y}$.

Goal: find *shared* latent variables $\boldsymbol{z}_n^{(s)}$ that capture common factors of variation across domains, as well as *private* latent variables $\boldsymbol{z}_n^{(x)}$ and $\boldsymbol{z}_n^{(y)}$ that capture domain-specific variation.

Model:

$$\boldsymbol{z}_n^{(s)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{M_s}) \qquad \boldsymbol{z}_n^{(x)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{M_x}) \qquad \boldsymbol{z}_n^{(y)} \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{M_y}) \tag{58}$$

and

$$\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{W}_{xx}\boldsymbol{z}_n^{(x)} + \boldsymbol{W}_{xs}\boldsymbol{z}_n^{(s)} + \boldsymbol{\mu}_x, \sigma^2 \boldsymbol{I}_{D_x}) \tag{59}$$

$$\boldsymbol{y}_n \sim \mathcal{N}(\boldsymbol{W}_{yy}\boldsymbol{z}_n^{(y)} + \boldsymbol{W}_{ys}\boldsymbol{z}_n^{(s)} + \boldsymbol{\mu}_y, \sigma^2 \boldsymbol{I}_{D_y}) \tag{60}$$

where M_s, M_x, and M_y are the dimensions of $\boldsymbol{z}_n^{(s)}$, $\boldsymbol{z}_n^{(x)}$, and $\boldsymbol{z}_n^{(y)}$, respectively, and the parameters consist of $\boldsymbol{W}_{xx}$, $\boldsymbol{W}_{xs}$, $\boldsymbol{\mu}_x$, $\boldsymbol{W}_{yy}$, $\boldsymbol{W}_{ys}$, $\boldsymbol{\mu}_y$, and σ^2.

Probabilistic Canonical Correlations Analysis II

Exercise: Draw the graphical model for probabilistic CCA.

Probabilistic Canonical Correlations Analysis III

Just as the MLE for probabilistic PCA yields the classical PCA solution, Bach and Jordan [2005] showed that the MLE for probabilistic CCA yields the classical CCA solution,

$$\boldsymbol{w}_{xs,1}, \boldsymbol{w}_{ys,1} = \arg\max \operatorname{corr}(\boldsymbol{w}_{xs,1}^\top \boldsymbol{x}, \boldsymbol{w}_{ys,1}^\top \boldsymbol{y}), \tag{61}$$

where $\boldsymbol{w}_{xs,1}$ and $\boldsymbol{w}_{ys,1}$ are the first columns of $\boldsymbol{W}_{xs}$ and $\boldsymbol{W}_{ys}$ respectively; aka the first **pair of canonical variables**.

Subsequent pairs of canonical variables are found by maximizing the same objective, subject to being orthogonal to previous pairs.

As with PCA, the classical CCA solution can be found with an SVD. Here, $\boldsymbol{W}_{xs}$ and $\boldsymbol{W}_{ys}$ are the left and right singular vectors of the sample correlation matrix $\operatorname{diag}(\boldsymbol{S}_{xx})^{-\frac{1}{2}} \boldsymbol{S}_{xy} \operatorname{diag}(\boldsymbol{S}_{yy})^{-\frac{1}{2}}$.

See Witten et al. [2009] for (non-Bayesian) sparse extensions to PCA and CCA.

References I

Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3): 515–534, 2009.