Linear Gaussian Latent Variable Models

STATS305B: Applied Statistics II

Scott Linderman

February 19, 2025

Last Time...

- Mixture Models & the EM Algorithm
- HMMs & the Forward-Backward Algorithm

Today...

Outline:

- Principal Components Analysis (PCA)
- PCA as a linear Gaussian latent variable model
- Factor analysis
- Linear Dynamical Systems & the Kalman Filter/Smoother

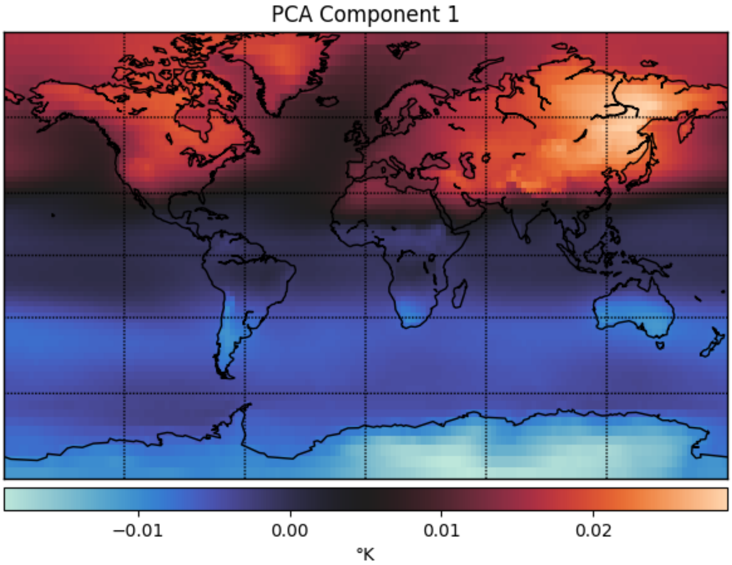
Motivating Example

Take HW3 as an example: you have temperature measurements at 9504 locations across the globe. Are the measurements really 9504 dimensional?

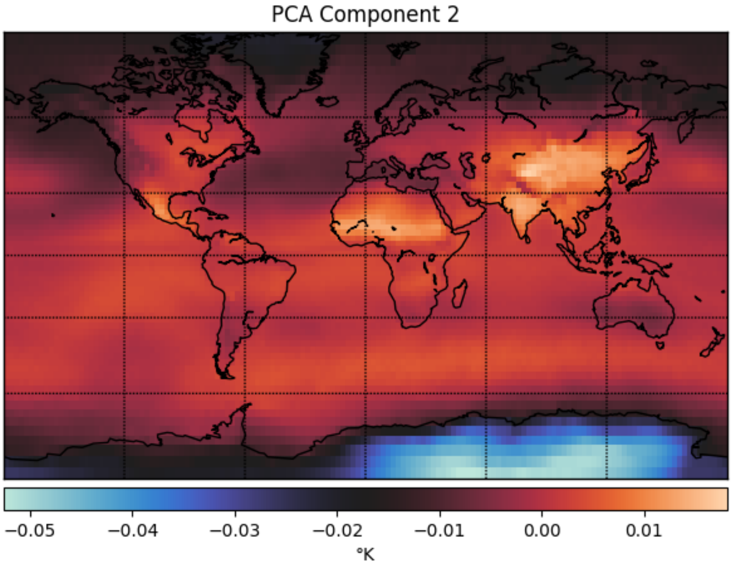
We might have a few objectives in mind:

- **Dimensionality reduction:** are there a few dimensions along which the temperatures primarily vary? Maybe northern and southern hemisphere, or land and sea?
- **Visualization:** Sometimes, we want to embed high-dimensional points in 2 or 3 dimensions for visualization.
- **Compression:** How can I best summarize the data if I am willing to sacrifice some reconstruction accuracy?

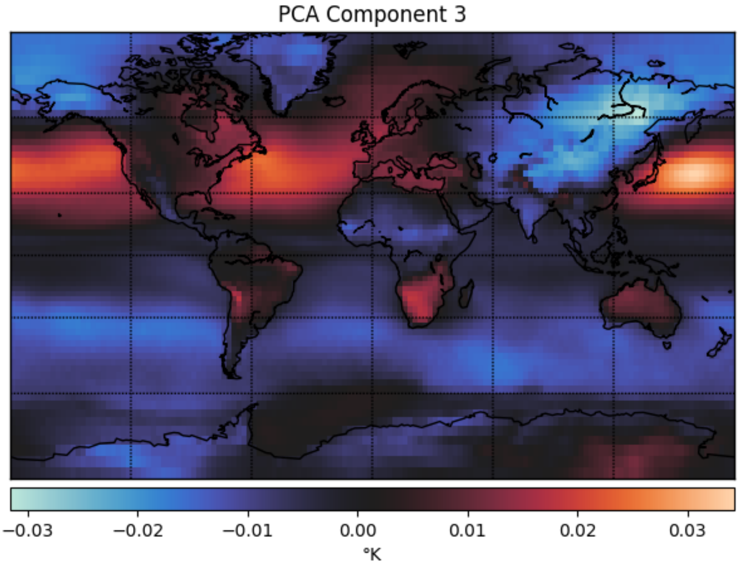
Principal Components of Global Temperature



Principal Components of Global Temperature



Principal Components of Global Temperature



Principal Components Analysis (PCA)

Two classical definitions:

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).

(Quoted from Bishop, Ch 12)

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:

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,

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

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

The variance is

$$\begin{aligned} \frac{1}{N}\sum_{n=1}^{N} \left[\boldsymbol{u}_1^\top \boldsymbol{x}_n - \boldsymbol{u}_1^\top \bar{\boldsymbol{x}} \right]^2 &= \frac{1}{N}\sum_{n=1}^{N} \left[\boldsymbol{u}_1^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \right]^2 && (2) \\ &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_1^\top (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^\top \boldsymbol{u}_1 && (3) \\ &= \boldsymbol{u}_1^\top \boldsymbol{S} \boldsymbol{u}_1 && (4) \end{aligned}$$

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 $\boldsymbol{u}_1$,

$$\boldsymbol{u}_1 = \arg\max_{\boldsymbol{u} \in \mathbb{S}_D} \boldsymbol{u}^\top \boldsymbol{S} \boldsymbol{u}. \tag{5}$$

This is the variational definition of the eigenvector with maximal eigenvalue!

I.e., $\boldsymbol{u}_1$ is the eigenvector of $\boldsymbol{S}$ with the largest eigenvalue, λ_1.

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

Since $\boldsymbol{S}$ is real and symmetric positive definite, the eigenvectors are orthogonal.

PCA and the Singular Value Decomposition

The first M **principal components** are the leading M **eigenvectors of the covariance matrix**. Equivalently, they are the first M **right singular vectors of the data matrix**.

Let

$$\boldsymbol{Y} = \frac{1}{\sqrt{N}}\boldsymbol{X} = \frac{1}{\sqrt{N}}\begin{bmatrix} - & (\boldsymbol{x}_1 - \bar{\boldsymbol{x}})^\top & - \\ & \vdots & \\ - & (\boldsymbol{x}_N^\top - \bar{\boldsymbol{x}})^\top & - \end{bmatrix} \tag{6}$$

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,

$$\boldsymbol{Y} = \boldsymbol{V}\Lambda^{\frac{1}{2}}\boldsymbol{U}^\top \Rightarrow \boldsymbol{Y}^\top \boldsymbol{Y} = \frac{1}{N}\boldsymbol{U}\Lambda^{\frac{1}{2}}\boldsymbol{V}^\top \boldsymbol{V}\Lambda^{\frac{1}{2}}\boldsymbol{U}^\top = \frac{1}{N}\boldsymbol{U}\Lambda\boldsymbol{U}^\top \tag{7}$$

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,

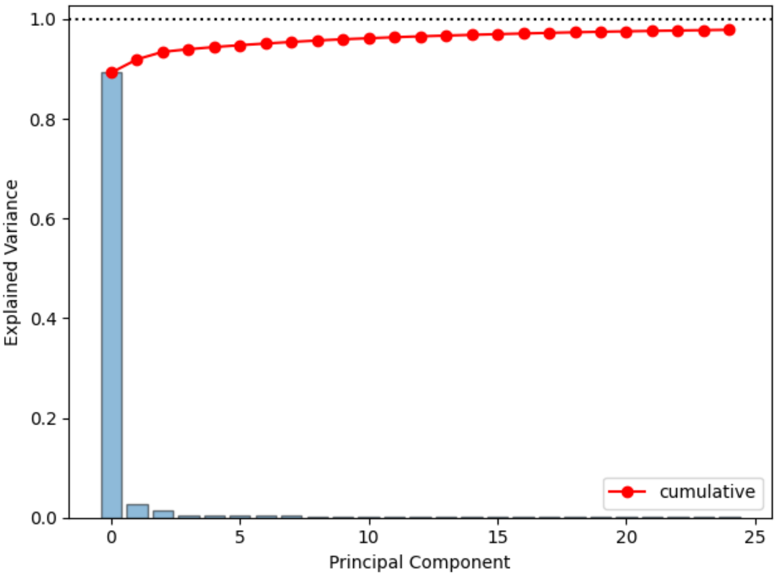
$$\mathrm{Cov}[\boldsymbol{z}] = \mathrm{Cov}[\boldsymbol{U}_M^\top(\boldsymbol{x} - \bar{\boldsymbol{x}})] = \boldsymbol{U}_M^\top \mathrm{Cov}[\boldsymbol{x}]\boldsymbol{U}_M = \mathrm{diag}([\lambda_1, \ldots, \lambda_M]). \tag{8}$$

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**,

$$\text{variance explained} = \frac{\mathrm{Tr}(\mathrm{Cov}[z; M \text{ components}])}{\mathrm{Tr}(\mathrm{Cov}[z; D \text{ components}])} = \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{D} \lambda_m} \in [0, 1]. \tag{9}$$

Scree Plots


1.0
0.8
0.6
0.4
0.2
0.0
Explained Variance
0
5
10
15
20
25
Principal Component
cumulative

Outline

- Principal Components Analysis (PCA)
- **PCA as a linear Gaussian latent variable model**
- Factor analysis
- Linear Dynamical Systems & the Kalman Filter/Smoother

Probabilistic PCA: A Continuous Latent Variable Model

We cast the principal components as the solutions to an optimization problem: maximize the projected variance.

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**, including EM.
- 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,

$$\boldsymbol{z}_n \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{10}$$

$$\boldsymbol{x}_n \mid \boldsymbol{z}_n \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}), \tag{11}$$

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,

$$\boldsymbol{x}_n = \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \tag{12}$$

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

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),

$$\boldsymbol{W}_{\text{ML}}, \boldsymbol{\mu}_{\text{ML}}, \sigma^2_{\text{ML}} = \arg\max \mathscr{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2), \tag{13}$$

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

$$\begin{aligned} \mathscr{L}(\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) &= \log p(\boldsymbol{X} \mid \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) & (14) \\ &= \log \int \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{z}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2)\, p(\boldsymbol{z}_n)\, \mathrm{d}\boldsymbol{z}_n & (15) \\ &= \log \int \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I})\, \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{0}, \boldsymbol{I})\, \mathrm{d}\boldsymbol{z}_n & (16) \end{aligned}$$

Exercise: Simplify this expression.

Maximum likelihood estimation of the parameters II

The log likelihood simplifies to,

$$\mathscr{L}(\boldsymbol{W}, \mu, \sigma^2) - \frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\boldsymbol{C}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \mu)^\top \boldsymbol{C}^{-1}(\boldsymbol{x} - \mu) \tag{17}$$

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,

$$\boldsymbol{W}_{\text{ML}} = \boldsymbol{U}_M(\Lambda_M - \sigma^2\boldsymbol{I})^{\frac{1}{2}}\boldsymbol{R}, \tag{18}$$

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 $\Lambda_M = \mathrm{diag}([\lambda_1, \ldots, \lambda_M])$, 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,

$$\sigma^2_{\text{ML}} = \frac{1}{D - M} \sum_{m=M+1}^{D} \lambda_m, \tag{19}$$

the average variance in the remaining dimensions.

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?

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 $\boldsymbol{z}_n$?

$$p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) \propto \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{0}, \boldsymbol{I})\, \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}) \tag{20}$$

$$\propto \exp\left\{-\frac{1}{2}\boldsymbol{z}_n^\top \boldsymbol{z}_n - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu})^\top (\sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n - \boldsymbol{\mu})\right\} \tag{21}$$

$$\propto \exp\left\{-\frac{1}{2}\boldsymbol{z}_n^\top \boldsymbol{J}_n \boldsymbol{z}_n + \boldsymbol{h}_n^\top \boldsymbol{z}_n\right\} \tag{22}$$

$$\tag{23}$$

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,

$$p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2) = \mathcal{N}(\boldsymbol{z}_n \mid \boldsymbol{J}_n^{-1}\boldsymbol{h}_n, \boldsymbol{J}_n^{-1}). \tag{24}$$

The Posterior Distribution in the Zero Noise Limit

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

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2] = \lim_{\sigma^2 \to 0} (\boldsymbol{I} + \frac{1}{\sigma^2} \boldsymbol{W}^\top \boldsymbol{W})^{-1} [\frac{1}{\sigma^2} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu})] \tag{25}$$

$$= \lim_{\sigma^2 \to 0} (\sigma^2 \boldsymbol{I} + \boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{26}$$

$$= (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{27}$$

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

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{W}, \boldsymbol{\mu}, \sigma^2] = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{28}$$

$$= \Lambda_M^{-\frac{1}{2}} \boldsymbol{U}_M^\top (\boldsymbol{x}_n - \boldsymbol{\mu}) \tag{29}$$

EM for Probabilistic PCA

The **E-step** is to compute the posterior $q(\boldsymbol{z}_n) = p(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{W}, \boldsymbol{\mu}, \sigma^2)$ are the current parameters. For simplicity, assume the data is centered so that $\boldsymbol{\mu}^\star = 0$.

The **M-step** is to maximize the expected complete data log likelihood,

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\log p(\boldsymbol{x}_n \mid \boldsymbol{z}_n; \boldsymbol{\theta})\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{W}\boldsymbol{z}_n, \sigma^2 \boldsymbol{I})\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[-\frac{D}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n)^\top (\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{z}_n)\right].
\end{aligned}$$

EM for Probabilistic PCA

As a function of $\boldsymbol{W}$,

$$\begin{aligned}\mathscr{L}(\boldsymbol{W}) &= \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[-\frac{1}{2\sigma^2}\left\langle \boldsymbol{W}^\top \boldsymbol{W}, \boldsymbol{z}_n\boldsymbol{z}_n^\top \right\rangle + \frac{1}{\sigma^2}\left\langle \boldsymbol{W}, \boldsymbol{x}_n\boldsymbol{z}_n^\top \right\rangle\right] \\ &= -\frac{1}{2\sigma^2}\left\langle \boldsymbol{W}^\top \boldsymbol{W}, \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\boldsymbol{z}_n\boldsymbol{z}_n^\top\right] \right\rangle + \frac{1}{\sigma^2}\left\langle \boldsymbol{W}, \sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\boldsymbol{x}_n\boldsymbol{z}_n^\top\right] \right\rangle.\end{aligned}$$

where $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$ is the Frobenius inner product for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

Taking derivatives wrt $\boldsymbol{W}$ and setting to zero yields,

$$\boldsymbol{W}^\star = \left(\sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\boldsymbol{x}_n\boldsymbol{z}_n^\top\right]\right)\left(\sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z}_n)}\left[\boldsymbol{z}_n\boldsymbol{z}_n^\top\right]\right)^{-1}.$$

It depends on sums of expected sufficient statistics!

Exercise: Derive the expected sufficient statistics and the M-step update for σ^2.

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,

$$\boldsymbol{z}_n \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{30}$$

$$\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2)) \tag{31}$$

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

Exercise: without doing any math, derive EM for this factor analysis model.

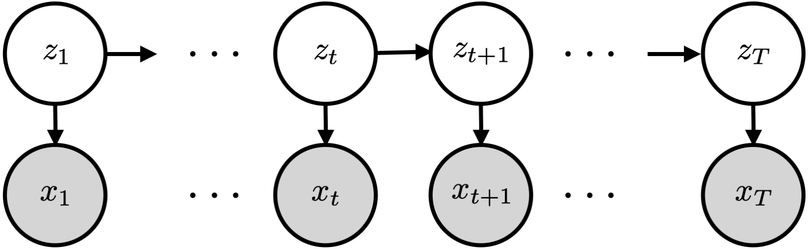
Outline

- Principal Components Analysis (PCA)
- PCA as a linear Gaussian latent variable model
- Factor analysis
- **Linear Dynamical Systems & the Kalman Filter/Smoother**

Recap: Hidden Markov Models (HMMs)

We generalized from mixture models to HMMs by assuming that the latent states were dependent.

HMMs assume a particular factorization of the joint distribution on latent states (z_t) and observations ($\boldsymbol{x}_t$). The graphical model looks like this:



This graphical model says that the joint distribution factors as,

$$p(z_{1:T}, \boldsymbol{x}_{1:T}) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid z_t). \tag{32}$$

State space models (SSMs)

Note that nothing above assumes that z_t is a discrete random variable!

HMM's are a special case of more general **state space models** with discrete states.

State space models assume the same graphical model but allow for arbitrary types of latent states.

For example, suppose that $\boldsymbol{z}_t \in \mathbb{R}^D$ are continuous valued latent states and that,

$$p(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \tag{33}$$

$$= \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{b}_1, \boldsymbol{Q}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{b}, \boldsymbol{Q}) \tag{34}$$

This is called a Gaussian **linear dynamical system** (LDS).

Stability of Gaussian linear dynamical systems

Question: What is the asymptotic mean of a Gaussian LDS, $\lim_{t\to\infty} \mathbb{E}[\boldsymbol{z}_t]$?

Question: When is a Gaussian LDS stable? I.e. when is the asymptotic mean finite?

Message passing in HMMs

In the HMM with discrete states, we showed how to compute posterior marginal distributions using message passing,

$$p(z_t \mid \boldsymbol{x}_{1:T}) \propto \sum_{z_1} \cdots \sum_{z_{t-1}} \sum_{z_{t+1}} \cdots \sum_{z_T} p(z_{1:T}, \boldsymbol{x}_{1:T}) \tag{35}$$

$$= \alpha_t(z_t)\, p(\boldsymbol{x}_t \mid z_t)\, \beta_t(z_t) \tag{36}$$

where the *forward and backward messages* are defined recursively

$$\alpha_t(z_t) = \sum_{z_{t-1}} p(z_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid z_{t-1})\, \alpha_{t-1}(z_{t-1}) \tag{37}$$

$$\beta_t(z_t) = \sum_{z_{t+1}} p(z_{t+1} \mid z_t)\, p(\boldsymbol{x}_{t+1} \mid z_{t+1})\, \beta_{t+1}(z_{t+1}) \tag{38}$$

The initial conditions are $\alpha_1(z_1) = p(z_1)$ and $\beta_T(z_T) = 1$.

What do the forward messages tell us?

The forward messages are equivalent to,

$$\alpha_t(z_t) = \sum_{z_1} \cdots \sum_{z_{t-1}} p(z_{1:t}, \boldsymbol{x}_{1:t-1}) \tag{39}$$

$$p(z_t, \boldsymbol{x}_{1:t-1}). \tag{40}$$

The normalized message is the *predictive distribution*,

$$\frac{\alpha_t(z_t)}{\sum_{z_t'} \alpha_t(z_t')} = \frac{p(z_t, \boldsymbol{x}_{1:t-1})}{\sum_{z_t'} p(z_t', \boldsymbol{x}_{1:t-1})} = \frac{p(z_t, \boldsymbol{x}_{1:t-1})}{p(\boldsymbol{x}_{1:t-1})} = p(z_t \mid \boldsymbol{x}_{1:t-1}), \tag{41}$$

The final normalizing constant yields the marginal likelihood, $\sum_{z_T} \alpha_T(z_T) = p(\boldsymbol{x}_{1:T})$.

Message passing in state space models

The same recursive algorithm applies (in theory) to any state space model with the same factorization, but the sums are replaced with integrals,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) \propto \int \mathrm{d}\boldsymbol{z}_1 \cdots \int \mathrm{d}\boldsymbol{z}_{t-1} \int \mathrm{d}\boldsymbol{z}_{t+1} \cdots \int \mathrm{d}\boldsymbol{z}_T \, p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}) \tag{42}$$

$$= \alpha_t(\boldsymbol{z}_t) \, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t) \, \beta_t(\boldsymbol{z}_t) \tag{43}$$

where the *forward and backward messages* are defined recursively

$$\alpha_t(\boldsymbol{z}_t) = \int p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1}) \, \alpha_{t-1}(\boldsymbol{z}_{t-1}) \, \mathrm{d}\boldsymbol{z}_{t-1} \tag{44}$$

$$\beta_t(\boldsymbol{z}_t) = \int p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t) \, p(\boldsymbol{x}_{t+1} \mid \boldsymbol{z}_{t+1}) \, \beta_{t+1}(\boldsymbol{z}_{t+1}) \, \mathrm{d}\boldsymbol{z}_{t+1} \tag{45}$$

The initial conditions are $\alpha_1(\boldsymbol{z}_1) = p(\boldsymbol{z}_1)$ and $\beta_T(\boldsymbol{z}_T) \propto 1$.

Forward pass in a linear dynamical system

Consider an linear dynamical system (LDS) with Gaussian emissions,

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \tag{46}$$

$$= \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{b}_1, \boldsymbol{Q}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{b}, \boldsymbol{Q}) \prod_{t=1}^{\mathcal{N}} (\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R}) \tag{47}$$

Let's derive the forward message $\alpha_{t+1}(\boldsymbol{z}_{t+1})$. Assume $\alpha_t(\boldsymbol{z}_t) \propto \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1})$.

$$\alpha_{t+1}(\boldsymbol{z}_{t+1}) = \int p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)\, \alpha_t(\boldsymbol{z}_t)\, \mathrm{d}\boldsymbol{z}_t \tag{48}$$

$$= \int \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{b}, \boldsymbol{Q})\, \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1})\, \mathrm{d}\boldsymbol{z}_t \tag{49}$$

The update step

The first step is the **update step**, where we **condition on** the emission $\boldsymbol{x}_t$,

Exercise: Expand the densities, collect terms, and complete the square to compute the mean $\boldsymbol{\mu}_{t|t}$ and covariance $\boldsymbol{\Sigma}_{t|t}$ after the update step,

$$\mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \propto \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}). \tag{50}$$

The update step II

Write the joint distribution,

$$p(\boldsymbol{z}_t, \boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1}) \tag{51}$$

$$= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{z}_t \\ \boldsymbol{x}_t \end{bmatrix} \,\middle|\, \begin{bmatrix} \boldsymbol{\mu}_{t|t-1} \\ \boldsymbol{C}\boldsymbol{\mu}_{t|t-1} + \boldsymbol{d} \end{bmatrix}, \begin{bmatrix} \Sigma_{t|t-1} & \Sigma_{t|t-1}\boldsymbol{C}^\top \\ \boldsymbol{C}\Sigma_{t|t-1} & \boldsymbol{C}\Sigma_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R} \end{bmatrix}\right) \tag{52}$$

Since $(\boldsymbol{z}_t, \boldsymbol{x}_t)$ are jointly Gaussian, $\boldsymbol{z}_t$ must be conditionally Gaussian as well,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_{t|t}, \Sigma_{t|t}). \tag{53}$$

Exercise: Now use the **Schur complement** from Week 1 to solve for $\boldsymbol{\mu}_{t|t}$ and $\Sigma_{t|t}$

The update step III

Exercise: Write $\mu_{t|t}$ and $\Sigma_{t|t}$ in terms of the **Kalman gain**,

$$\boldsymbol{K}_t = \Sigma_{t|t-1}\boldsymbol{C}^\top(\boldsymbol{C}\Sigma_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R})^{-1} \tag{54}$$

What is the Kalman gain doing?

The predict step

The predict step yields $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) = \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \Sigma_{t|t})$. To complete the forward pass, we need the **predict step**,

$$\alpha_{t+1}(\boldsymbol{z}_{t+1}) = \int p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)\, \alpha_t(\boldsymbol{z}_t)\, \mathrm{d}\boldsymbol{z}_t \tag{55}$$

$$= \int \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{b}, \boldsymbol{Q})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \Sigma_{t|t})\, \mathrm{d}\boldsymbol{z}_t \tag{56}$$

$$= \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{\mu}_{t+1|t}, \Sigma_{t+1|t}) \tag{57}$$

Exercise: Solve for the mean $\boldsymbol{\mu}_{t+1|t}$ and covariance $\Sigma_{t+1|t}$ after the predict step.

Completing the recursions

That wraps up the recursions! All that's left is the base case, which comes from the initial state distribution,

$$\boldsymbol{\mu}_{1|0} = \boldsymbol{b}_1 \qquad \text{and} \quad \boldsymbol{\Sigma}_{1|0} = \boldsymbol{Q}_1. \tag{58}$$

Computing the marginal likelihood

Like in the discrete HMM, we can compute the marginal likelihood along the forward pass.

$$p(\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}) \tag{59}$$

$$= \prod_{t=1}^{T} \int p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t-1})\, \mathrm{d}\boldsymbol{z}_t \tag{60}$$

$$= \prod_{t=1}^{T} \int \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1})\, \mathrm{d}\boldsymbol{z}_t \tag{61}$$

Exercise: Obtain a closed form expression for the integrals.

Computing the smoothing distributions

- The forward pass gives us the filtering distributions $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})$. How can we compute the smoothing distributions, $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T})$?
- In the discrete HMM we essentially ran the *same algorithm in reverse* to get the backward messages, starting from $\beta_T(\boldsymbol{z}_T) \propto 1$.
- We can do the same sort of thing here, but it's a bit funky because we need to start with an improper Gaussian distribution $\beta_T(\boldsymbol{z}_T) \propto \mathcal{N}(\boldsymbol{0}, \infty \boldsymbol{I})$.
- Instead, we'll derive an alternative way of computing the smoothing distributions.

Bayesian Smoothing

Note: $\boldsymbol{z}_t$ is conditionally independent of $\boldsymbol{x}_{t+1:T}$ given $\boldsymbol{z}_{t+1}$, so

$$p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:T}) = p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:t}) \tag{62}$$

$$= \frac{p(\boldsymbol{z}_t, \boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})} \tag{63}$$

$$= \frac{p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})} \tag{64}$$

Question: what rules did we apply in each of these steps?

Bayesian Smoothing II

Now we can write the joint distribution as,

$$
\begin{aligned}
p(\boldsymbol{z}_t, \boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T}) &= p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T}) p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T}) && (65) \\
&= \frac{p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T})}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})}. && (66)
\end{aligned}
$$

Marginalizing over $\boldsymbol{z}_{t+1}$ gives us,

$$
p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) = p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) \int \frac{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T})}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})} \, \mathrm{d}\boldsymbol{z}_{t+1} \tag{67}
$$

Question: Can we compute each of these terms?

The Rauch-Tung-Striebel Smoother, aka Kalman Smoother

These recursions apply to any Markovian state space model. For the special case of a Gaussian linear dynamical system, the smoothing distributions are again Gaussians,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|T}, \boldsymbol{\Sigma}_{t|T}) \tag{68}$$

where

$$\boldsymbol{\mu}_{t|T} = \boldsymbol{\mu}_{t|t} + \boldsymbol{G}_t(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}) \tag{69}$$

$$\boldsymbol{\Sigma}_{t|T} = \boldsymbol{\Sigma}_{t|t} + \boldsymbol{G}_t(\boldsymbol{\Sigma}_{t+1|T} - \boldsymbol{\Sigma}_{t+1|t})\boldsymbol{G}_t^\top \tag{70}$$

$$\boldsymbol{G}_t \triangleq \boldsymbol{\Sigma}_{t|t}\boldsymbol{A}^\top \boldsymbol{\Sigma}_{t+1|t}^{-1}. \tag{71}$$

This is called the **Rauch-Tung-Striebel (RTS) smoother** or the **Kalman smoother**.

References