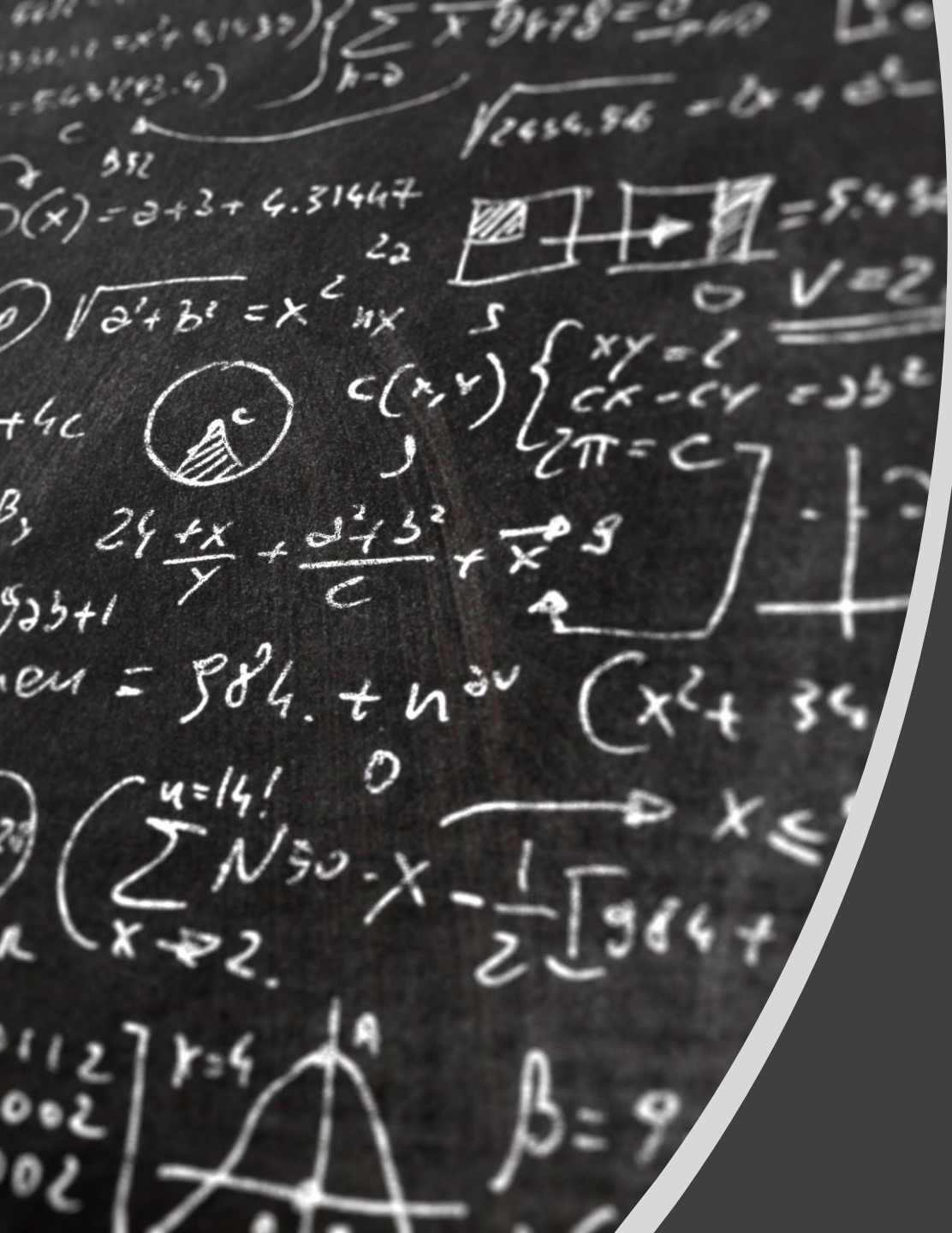
Questions?

https://pollev.com/haroldsohsoo986



CS5340: Tutorial 6

Asst. Prof. Harold Soh

TAs: Chen Kaiqi

Course Schedule

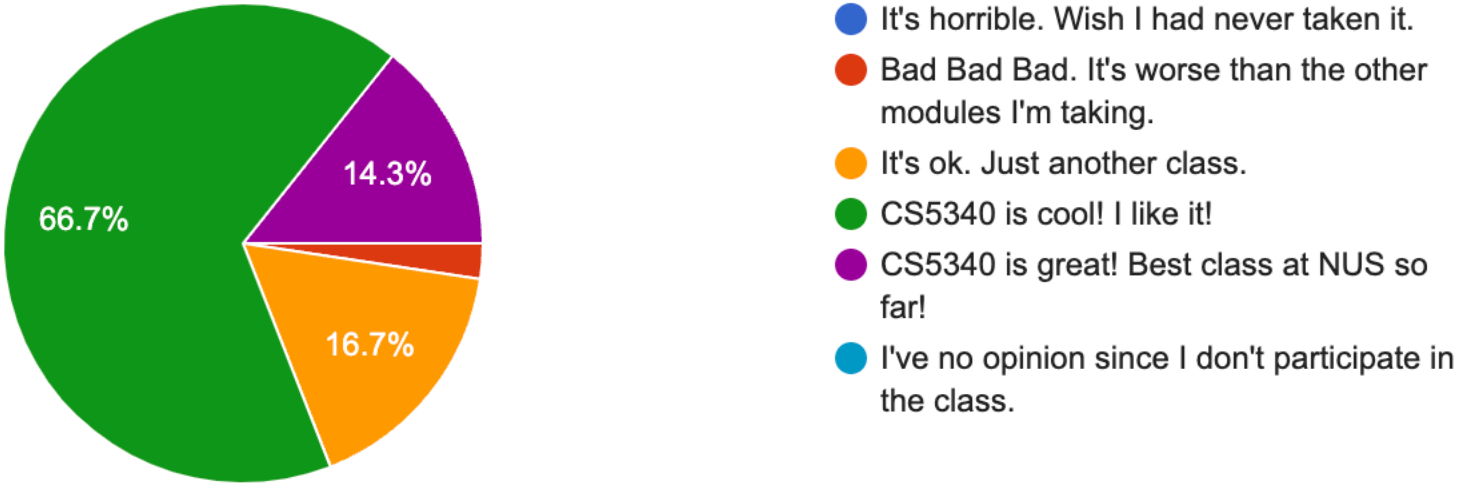
Week	Date	Lecture Topic	Tutorial Topic
1	12 Jan	Introduction to Uncertainty Modeling + Probability Basics	Introduction
2	19 Jan	Simple Probabilistic Models	Probability Basics
3	26 Jan	Bayesian networks (Directed graphical models)	More Basic Probability
4	2 Feb	Markov random Fields (Undirected graphical models)	DGM modelling and d-separation
5	9 Feb	Variable elimination and belief propagation	MRF + Sum/Max Product
6	16 Feb	Factor graph and the junction tree algorithm	**Quiz 1**
-	-	RECESS WEEK	
7	2 Mar	Mixture Models and Expectation Maximization (EM)	Linear Gaussian Models
8	9 Mar	Hidden Markov Models (HMM)	Probabilistic PCA
9	**16** Mar	Monte-Carlo Inference (Sampling)	Linear Gaussian Dynamical System
10	23 Mar	Variational Inference	MCMC + Sequential VAE
11	30 Mar	Inference and Decision-Making (Special Topic)	**Quiz 2**
12	6 Apr	Gaussian Processes (Special Topic)	Wellness Day
13	13 Apr	**Project Presentations**	Closing

Announcements

- Harold is away next week.
  - Video tutorial instead.
  - Tutorial Quiz to be released on Thursday 6:30pm. Open until Friday 11:59pm.
  - Extra class poll
- Quiz results are out.
  - Please check Canvas
  - Negative grading removed.
- Project abstracts graded
  - 5% to participation.
- Survey results are available.

Survey Feedback

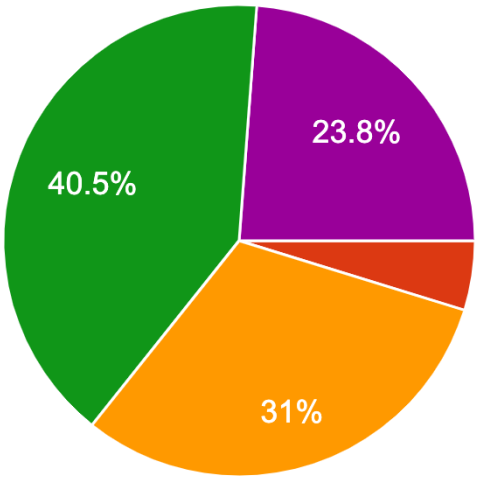
- Number of Respondents: 42 / 74 (56.7%)
- **Bottom line:** Going ok but lots of room for improvement!



Lectures

How do you find the lectures?

42 responses



- I have no clue what the lecturer is taking about most of the time
- I'm confused half the time.
- The lectures are ok. Not the worst, not the best.
- Lectures are clear and I understand the material quite well.
- Lectures are great! Understandable and interesting!
- No comment. I don't attend lectures.

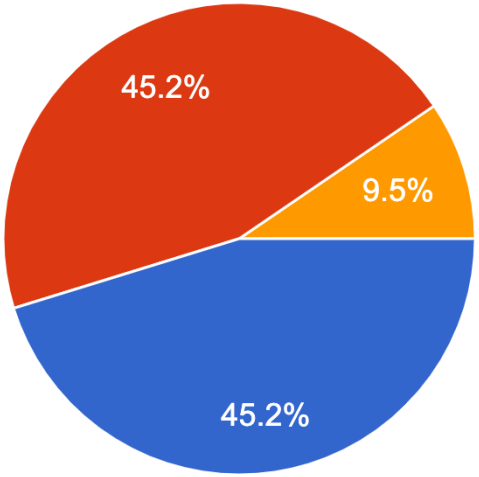
Lecture Suggestions / Comments

- More examples
  - more exercises or sample questions
  - unsure how to apply them cos the were not well explained in the lecture examples
  - show how these knowledge can be applied in some simple CV/NLP tasks, like image classification
  - topics with coding implementations
- More interaction:
  - “I like the parts where the vids are paused and we have to think through some questions, I think would be helpful to incorporate more of that in!”
  - “[opportunities to] ask clarifying questions”
- More structure / comparisons between methods
  - E.g., pros/cons/use cases table
- Content/Material:
  - Too much material in one lecture.
  - Too little explanation and material.
  - Lecture is too long
  - Pace is a bit too fast and condensed.

Tutorials

Do you think the tutorials are a useful learning experience?

42 responses



- Yes, I enjoy them very much.
- The tutorials are above average
- So so. Pretty ambivalent about it.
- The tutorials are below average. They don't help me much.
- Tutorials suck! They are a waste of time.
- I have not attended the lectures/tutorials so, no comment.

Suggestions / Comments

- Changes to structure:
  - Spend less time on "low-level" derivations
    - More time on motivating "higher-level" approaches.
  - Easier if you could explain the answers without spending time on rewriting them out. We could also refer to pre-printed answers line by line.
    - Past Year: Use whiteboard/notepad instead of slides.
- More problems:
  - Extra optional problems can be provided
  - Give more applications
    - Combine with deep learning
- After tutorials:
  - Recording for tutorial

Other Suggestions/Comments

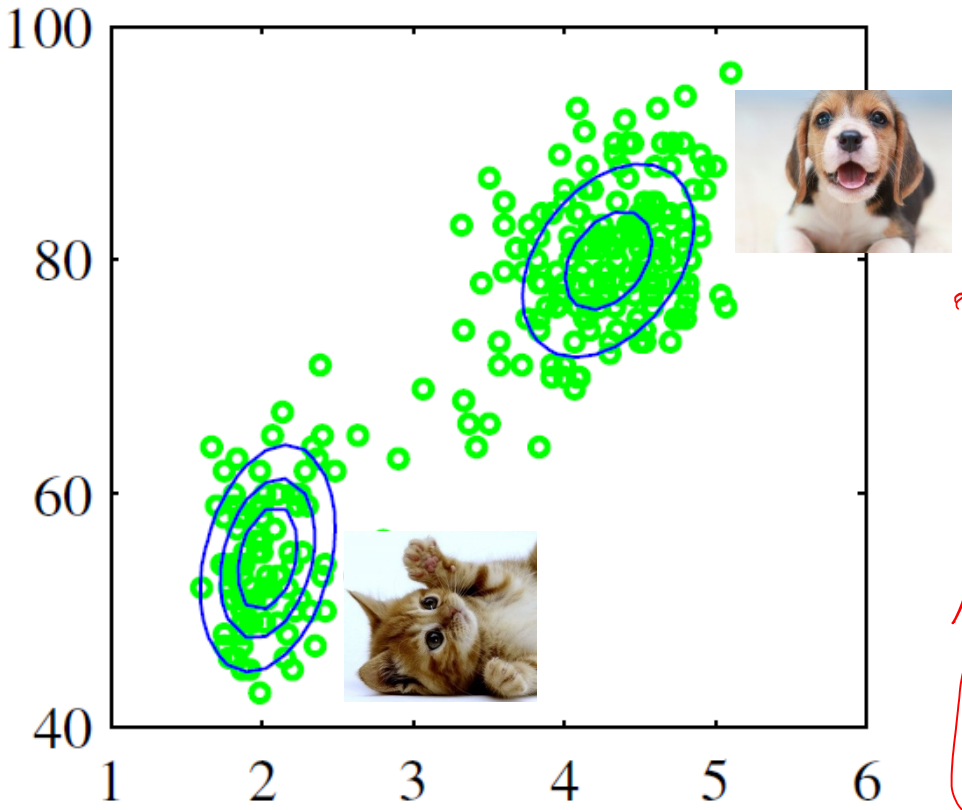
- Workload heavy! ☹
  - Takes long to understand material.
- Not enough materials
  - Don’t assume ”everyone knows everything”
  - Provide more materials
  - Mini-assignments to apply each concept for each lecture
  - More application-based questions/coding assignments
- Add TA or Office Hours.
  - Right now, by appointment.
  - We can hold office hours as well.
- Quiz was really hard. ☹
  - breakup the weightage of each quiz into 4 parts of 10% each to allow one to slowly improve
  - very hard assignment which takes more than a week to solve will help understand the concepts more than a speed quiz.
- Reduce forum participation score
  - Now is 5% since everyone got 5% for the abstract.

Questions?

https://pollev.com/haroldsohsoo986

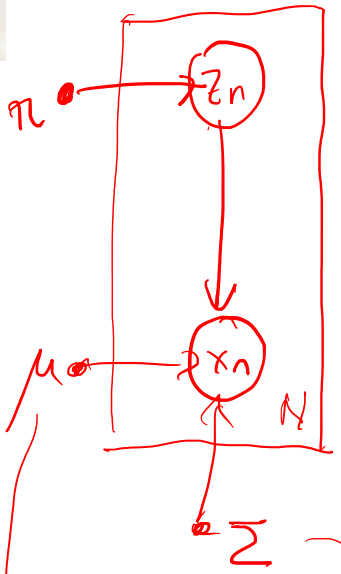
Recap of GMM

Intuition: Generative Story



$z_n \sim P(z)$

$x_n \sim P(x_n | z_n)$



$p(z)$?

$p(z) = Cat(\pi)$

$p(x|z)$?

$p(x|z) = N(\mu, \Sigma)$

$\mu = \{\mu_1, \mu_2\}$ $\Sigma = \{\Sigma_1, \Sigma_2\}$

Multivariate Normal Distribution

- Multivariate normal distribution describes a *D*-dimensional continuous variable $\boldsymbol{X}$, i.e. $\boldsymbol{x} \in \mathbb{R}^D$.
- *D*-dimensional mean $\boldsymbol{\mu} \in \mathbb{R}^D$, and $D \times D$ symmetrical positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}_+^{D \times D}$.

$$p(\boldsymbol{X} = \boldsymbol{a} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{a} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a} - \boldsymbol{\mu}) \}, \quad \boldsymbol{a} \in \mathbb{R}^D$$

Or

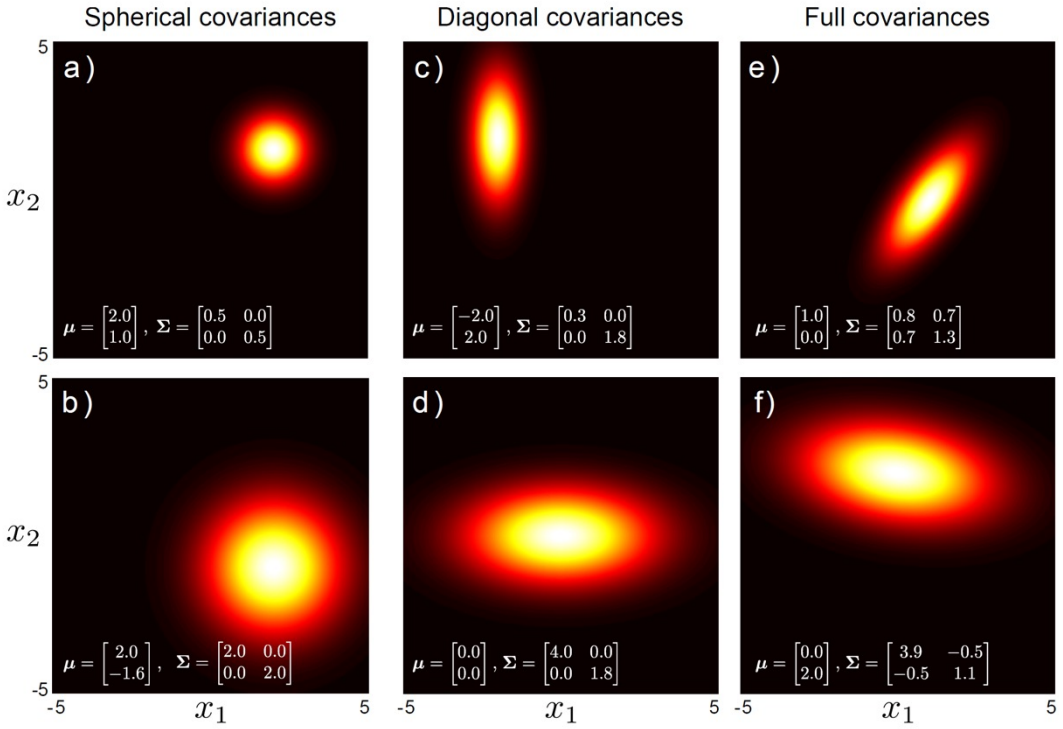
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \}$$

$$p(\boldsymbol{x}) = \mathrm{Norm}_{\boldsymbol{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

Types of Covariance

- Covariance matrix has three forms: spherical, diagonal and full.

$$\Sigma_{spher} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \Sigma_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \Sigma_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(\mathrm{Z}|\mathrm{X}, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\mathrm{old}} \leftarrow \boldsymbol{\theta}^{\mathrm{new}}$$

Quiz + Recess

"models, models, models"

Simple → Complex

Discrete/linear Cont/nonlinear

mixture model
- clustering
"latent variable"

↓

Hidden Markov model
- sequence modeling

↓

Liner Dynamical System
- Kalman Filtering

→

prob PCA "linear"
- dim red.

↓

"nolinear"

Variational Autoencoder (VAE)
- rep. learning / generative modeling

↓

Sequential VAE "world model"

Principal Components Analysis (PCA)

- Invented in 1901 by Karl Pearson
  - Independently by Hoteling in 1930s.
- Unsupervised Learning method
- Useful for:
  - Representation learning
  - Dimensionality reduction
  - Compression
  - Data-preprocessing
  - Visualization

Karl Pearson, 1912
(image credit: Wikipedia)

Image Credit: https://www.geeksforgeeks.org/ml-face-recognition-using-eigenfaces-pca-algorithm/

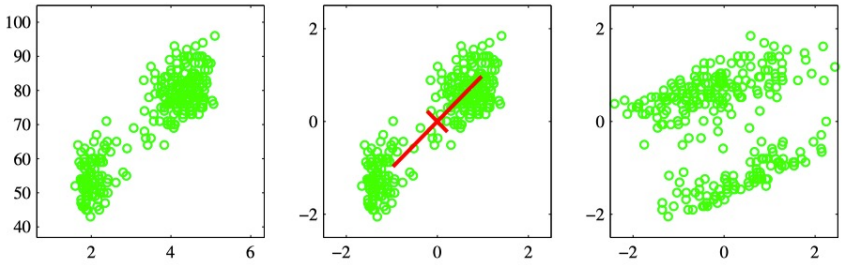
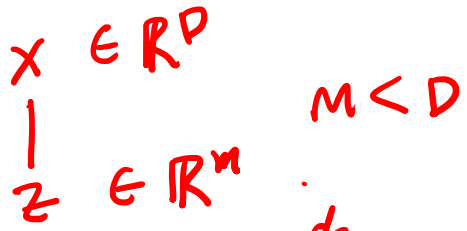
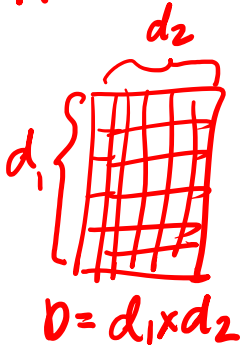


Image Credit: PRML Chp 12

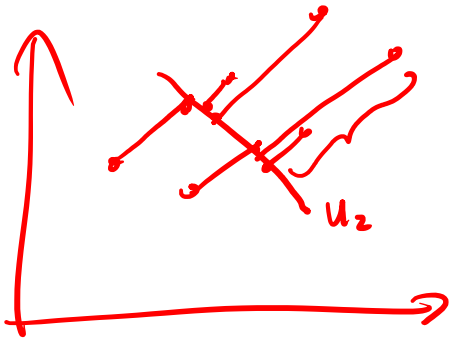
PCA Setup and Intuition

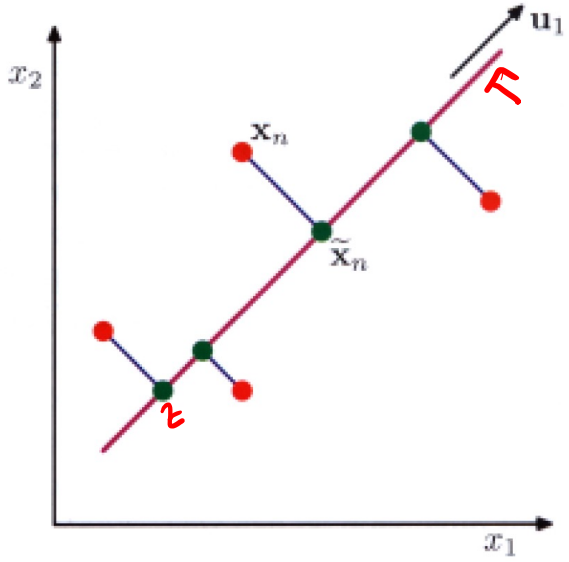




- Dataset of D-dimensional data points $\mathbf{x}_i$
- Want to associate each data point $\mathbf{x}_i$ with a corresponding M-dimensional point $\mathbf{z}_i$
  - where $M < D$
- 2 approaches to derivation. Project to:
  - Maximize variance
  - Minimize distortion
- In practice, we compute $\mathbf{X}\mathbf{X}^\top$ and find the M largest eigenvectors and eigenvalues

Maximizing Variance





direction u_1, $u_1^T u_1 = 1$

each x_n, project. $u_1^T x_n$

mean $u_1^T \bar{x}$, $\bar{x} = \frac{1}{N}\sum_n x_n$

var $\frac{1}{N}\sum (u_1^T x_n - u_1^T \bar{x})^2 = u_1^T S u_1$

$S = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$

maximize $u_1^T S u_1$

s.t $u_1^T u_1 = 1$

$\mathcal{L} = u_1^T S u_1 + \lambda(1 - u_1^T u_1)$

$\frac{\partial \mathcal{L}}{\partial u_1} = 0 \Rightarrow \boxed{S u_1 = \lambda u_1}$

Image Credit: PRML Chp 12

Probabilistic PCA (PPCA)

- Derive Probabilistic variant
- Learn via EM
- Advantages:
  - Efficient EM algorithm (avoids computing $\mathbf{X}\mathbf{X}^{\top}$)
  - Naturally deal with missing data
  - Can be extended to include class labels, factor analysis, kernel variants ...

PPCA – Generative View

$z_n \sim N(\,)$

$x_n = f(z_n).$

$\Downarrow$

$Az_n + b.$

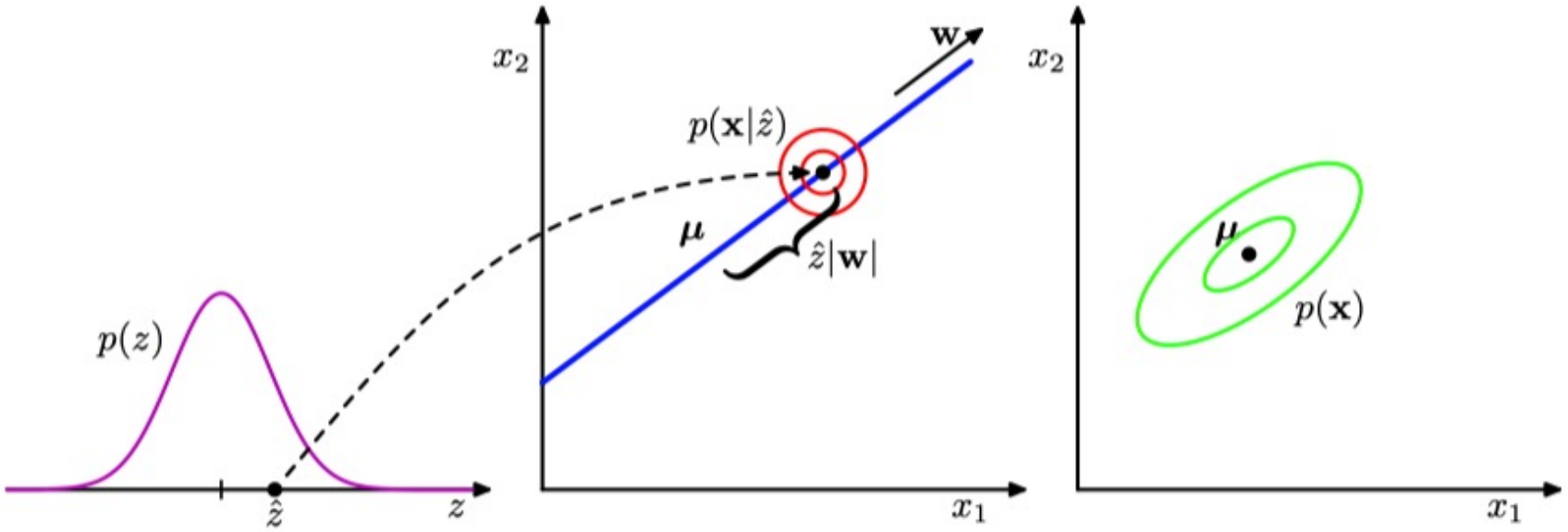
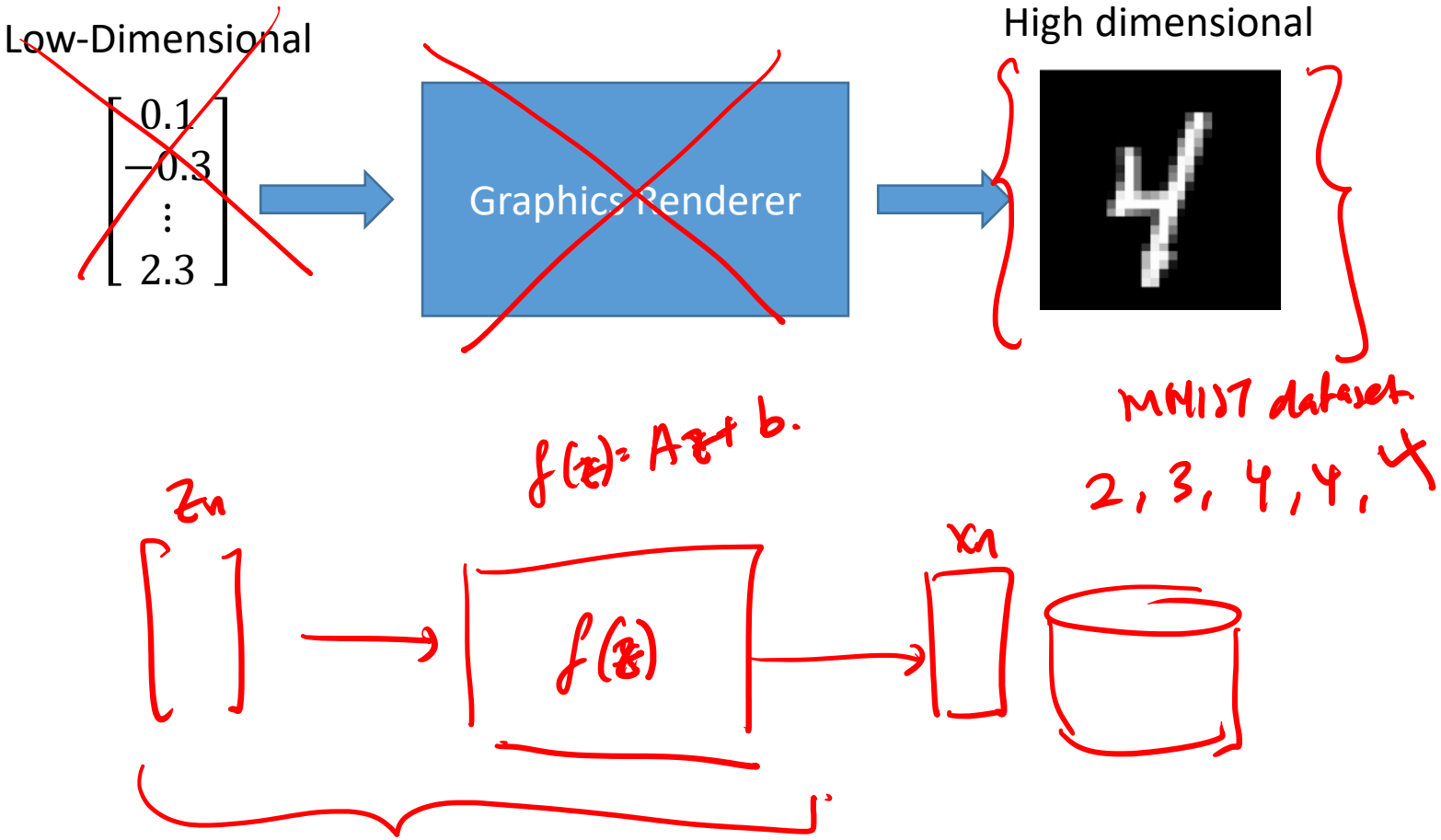


Figure 12.9 An illustration of the generative view of the probabilistic PCA model for a two-dimensional data space and a one-dimensional latent space. An observed data point $\mathbf{x}$ is generated by first drawing a value $\widehat{z}$ for the latent variable from its prior distribution $p(z)$ and then drawing a value for $\mathbf{x}$ from an isotropic Gaussian distribution (illustrated by the red circles) having mean $\mathbf{w}\widehat{z} + \boldsymbol{\mu}$ and covariance $\sigma^2\mathbf{I}$. The green ellipses show the density contours for the marginal distribution $p(\mathbf{x})$.

Image Credit: PRML Chp 12

Rendering Analogy



Probabilistic PCA

For the probabilistic PCA model, we have D-dimensional data points $\mathbf{x}_i$ for $i = 1, 2, \ldots, N$ and we aim to find some reduced structure for the data. For each data point, we associate a M-dimensional latent variable (where often $M < D$) $\mathbf{z}_i$ that has prior distribution,

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

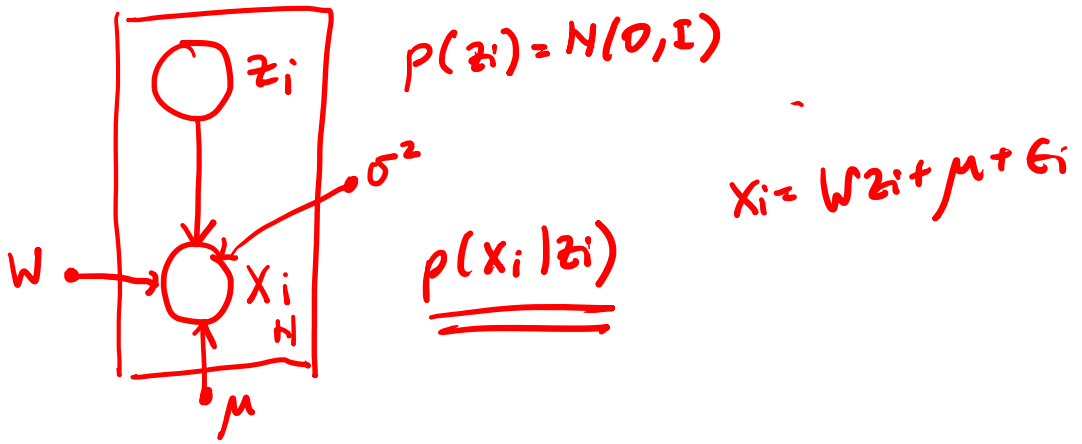
We define each observed variable $\mathbf{x}$ as,

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We can imagine that each data point is obtained by first sampling from the prior $p(\mathbf{z}_i)$ followed by an affine transformation and additive Gaussian noise.

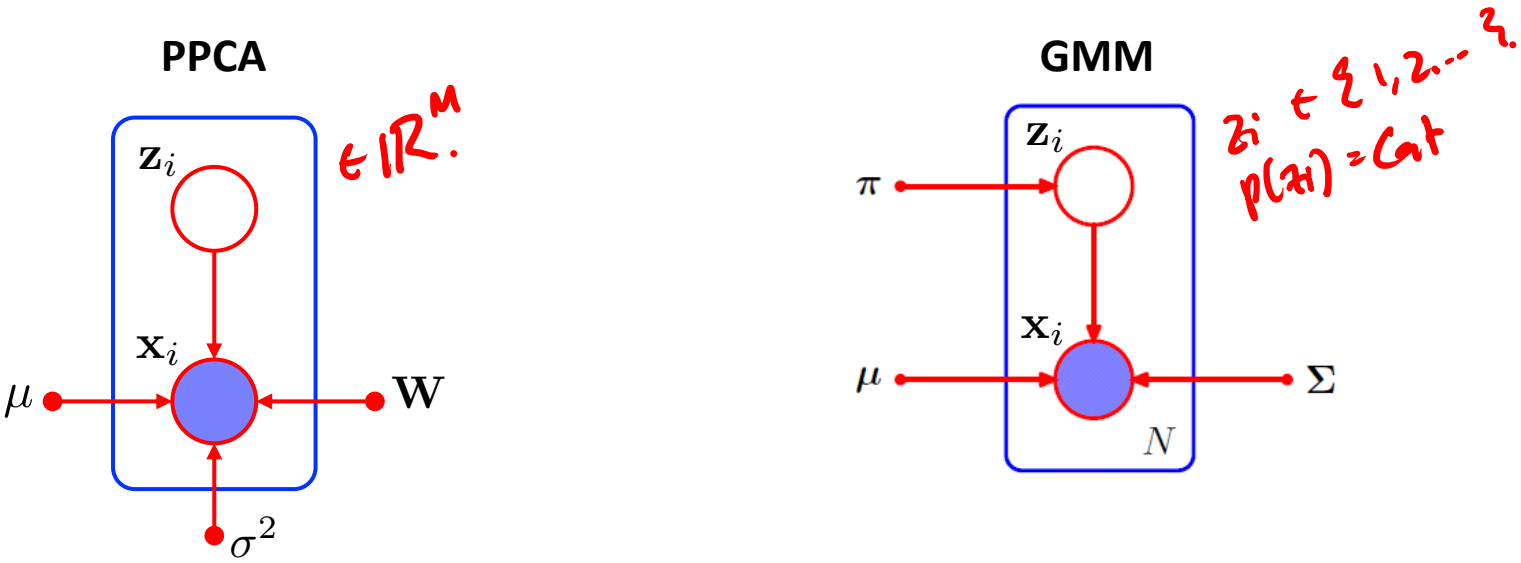
Problem 1.a. Draw the DGM corresponding to the model above. *Hint:* use plate notation for the different data points.

Solution:



DGM for PPCA

Relationship to GMMs?



Problem 1.b. Show that the conditional distribution for each observed variable $\mathbf{x}_i$ is given by:

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

Solution:

We can apply the affine property of Gaussian distribution,

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \Leftrightarrow (\mathbf{A}\mathbf{x} + \mathbf{b}) \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}^T) \tag{1}$$

Since we are computing $p(\mathbf{x}_i|\mathbf{z}_i)$, we consider $\mathbf{z}_i$ is observed ("deterministic", i.e., zero uncertainty)
Given $\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$,

$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \mathbf{0}, \sigma^2\mathbf{I}) \tag{2}$$

$$= \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) \tag{3}$$

$x_i = Wz_i + \mu + \epsilon_i$

$p(x_i|z_i)$

$\neq$

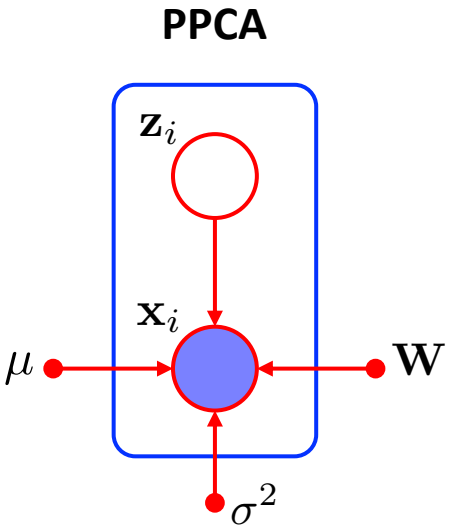
$p(x_i)$

$\epsilon_i \sim N(0, \sigma^2 I)$

$x_i|z_i \sim N(Wz_i + \mu, \sigma^2 I)$

How to Train?

- MLE (Direct)
  - Closed form solution but need covariance matrix
  - Cannot handle missing data.
  - Complexity $O(ND^2 + MD^2)$
- EM
  - Iterative method
  - Don't need full covariance matrix.
  - Can handle missing data
  - Each iteration $O(NDM)$



The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(\mathrm{Z}|\mathrm{X}, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

$= \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})}\left[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right]$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\boldsymbol{\theta}^{\mathrm{old}} \leftarrow \boldsymbol{\theta}^{\mathrm{new}}$$

Problem 1.c. To find the MLE values for the model parameters $\mathbf{W}, \boldsymbol{\mu}$, and σ^2, we would need the marginal distribution $p(\mathbf{X}) = \prod_i^N p(\mathbf{x}_i)$ (assuming i.i.d. data). Due to the latent variables, we will use the EM algorithm[2]. This requires us to marginalize out the latent $\mathbf{z}$'s. To help us along,

1. First, show that the marginal distribution of each data point is again a Gaussian given by
$$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$
where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
2. Then, show that the posterior distribution is also normally distributed,
$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$$
where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Hint: Given random variables $\mathbf{x}$ and variable $\mathbf{y}$ where:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x) \tag{4}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x}) \tag{5}$$

The marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T) \tag{6}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right) \tag{7}$$

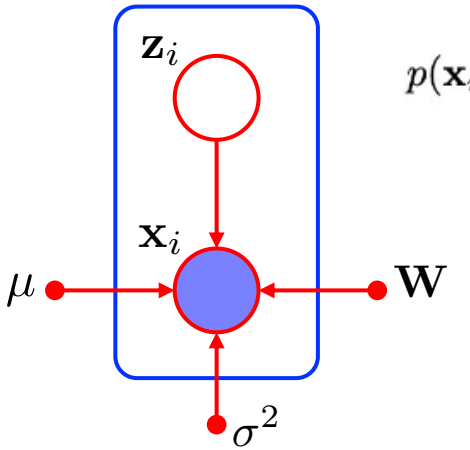
where
$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

1. First, show that the marginal distribution of each data point is again a Gaussian given by

$$p(\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$$

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

Solution:



$$p(\mathbf{x}_i) = \mathcal{N}(\mathbf{0} + \boldsymbol{\mu} + \mathbf{0}, \mathbf{W}\mathbf{I}\mathbf{W}^\top + \sigma^2\mathbf{I})$$
$$= \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})$$

General

If:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_x)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x})$$

then:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

PPCA

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

$p(y) = N(y|A\mu+b, \Sigma_{y|x} + A\Sigma_x A^T)$

$p(x_i) = N(x_i|W0+\mu, \sigma^2 I + WIW^T)$

$= N(x_i|\mu, C)$

C

$x = z_i$, $\hat{\mu} = 0$, $\Sigma_x = I$

$y = x_i$, $A = W$, $b = \mu$,

$\Sigma_{y|x} = \sigma^2 I$

2. Then, show that the posterior distribution is also normally distributed,

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})$$

where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

Solution:

$$p(\mathbf{z}_i|\mathbf{x}_i) = \mathcal{N}\left(\mathbf{z}_i|\boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i}\left(\mathbf{W}^T\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_i|\mathbf{z}_i}(\mathbf{x}_i - \boldsymbol{\mu}) + \boldsymbol{\Sigma}^{-1}_{\mathbf{z}_i}\mathbf{0}\right), \boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i}\right)$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{x}_i|\mathbf{z}_i} &= \sigma^2\mathbf{I} \\
\boldsymbol{\Sigma}_{\mathbf{z}_i|\mathbf{x}_i} &= \left(\boldsymbol{\Sigma}^{-1}_{\mathbf{x}_i} + \mathbf{W}^T\boldsymbol{\Sigma}^{-1}_{\mathbf{z}_i|\mathbf{x}_i}\mathbf{W}\right)^{-1} \\
&= \left(\mathbf{I}^{-1} + \mathbf{W}^T(\sigma^2\mathbf{I})^{-1}\mathbf{W}\right)^{-1} \\
&= \sigma^2(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1} = \sigma^2\mathbf{M}^{-1}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{z}_i|\mathbf{x}_i) &= \mathcal{N}\left(\mathbf{z}_i|\sigma^2\mathbf{M}^{-1}\left(\mathbf{W}^T(\sigma^2\mathbf{I})^{-1}\right)(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}\right) \\
&= \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1})
\end{aligned}$$

If:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x})$$

then:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}_{y|x}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}^{-1}_x\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}^{-1}_x + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}_{y|x}\mathbf{A}\right)^{-1}$$

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu} + \mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$$

Problem 1.d. Finally, derive the E-step and the M-step for the EM algorithm applied to probabilistic PCA. *Hint:* If you are really stuck, refer to Chapter 12.2.2. of Bishop's Pattern Recognition and Machine Learning. This portion is not especially difficult but is notationally heavy and requires algebraic manipulation.

$p(x_i, z_i) = p(z_i)p(x_i|z_i)$

$\prod p(x_i, z_i)$

The General EM Algorithm

1. Choose an initial setting for the parameters θ^{old}.
2. Expectation step: Evaluate $p(Z|X, \theta^{old})$.
3. Maximization step: Evaluate θ^{new} given by:

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

where

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

4. Check for convergence of either the log likelihood or the parameter values, if not converged:

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

1. Parameters: $\boldsymbol{\theta} = \{\mathbf{W}, \boldsymbol{\mu}, \sigma^2\}$
2. Expectation: Evaluate $p(\mathbf{z}_i|\mathbf{x}_i, \theta^{old})$
3. Maximization:
   - Obtain the Q function
   - We need:

$$\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right) = \sum_{n=1}^{N} \{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\}$$

   - Which will lead to the expectation over the posterior
   - That we then maximize in the usual way.

M-Step Summary

- First compute:

$$\mathbb{E}_{q(\mathbf{z}_n)}\left[\ln p\left(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2\right)\right] = \mathbb{E}_{q}\left[\sum_{n=1}^{N}\{\ln p(\mathbf{x}_n|\mathbf{z}_n) + \ln p(\mathbf{z}_n)\}\right]$$

$$\ln\ p(\mathbf{x}_n|\mathbf{z}_n, \theta) = -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_n)^T(\sigma^2\mathbf{I})^{-1}(\mathbf{x}_n - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_n)$$

$$= -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{x}_n - \boldsymbol{\mu})^T(\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{2\sigma^2}(\mathbf{W}\mathbf{z}_n)^T(\mathbf{W}\mathbf{z}_n) + \frac{1}{\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu})$$

$$= -\frac{D}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\mathbf{x}_n - \boldsymbol{\mu}||^2 - \frac{1}{2\sigma^2}\text{trace}(\mathbf{z}_n\mathbf{z}_n^T\mathbf{W}^T\mathbf{W}) + \frac{1}{\sigma^2}\mathbf{z}_n^T\mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu})$$

$$\ln\ p(\mathbf{z}_n|\theta) = -\frac{1}{2}\mathbf{z}_n^T\mathbf{z}_n$$

$$Q(q, \theta) = -\sum_{n=1}^{N}\left[\frac{D}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}||\mathbf{x}_n - \boldsymbol{\mu}||^2\right] + \frac{1}{2\sigma^2}\text{trace}\left[\mathbb{E}_{q(\mathbf{z}_n)}[\mathbf{z}_n\mathbf{z}_n^T]\mathbf{W}^T\mathbf{W}\right] +$$
$$\frac{1}{\sigma^2}\mathbb{E}_{q(\mathbf{z}_n)}\left[[\mathbf{z}_n]^T\mathbf{W}^T(\mathbf{x}_n - \boldsymbol{\mu})\right] + \frac{1}{2}[\text{trace}(\mathbb{E}_{q(\mathbf{z}_n)}[\mathbf{z}_n\mathbf{z}_n^T])\Big]$$

Code Demonstration

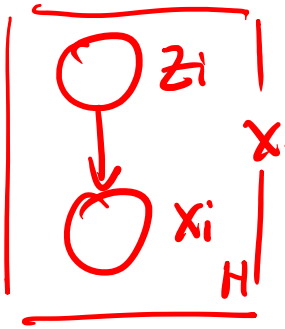
- Dimensionality reduction via PPCA and visualization.
- Simple Circle Dataset
- Oil Flow Dataset
  - 12 dimensional dataset
  - But only 2 intrinsic dimensions.

PPCA Recap

Data $D = \{x_i\}$ $x_i \in \mathbb{R}^D$

z_i $z_i \in \mathbb{R}^2$

model



$x_i = Wz_i + \mu + \epsilon$

Train/ learn E M.

Code & Execute

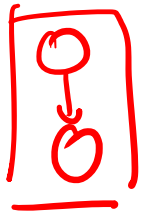
$p(z_i | x_i)$

$p(x_i | z_i)$

Linear Gaussian Model

- PPCA is actually a special case of a Linear Gaussian Model.
- Let’s examine this more general model.

Linear Gaussian Model

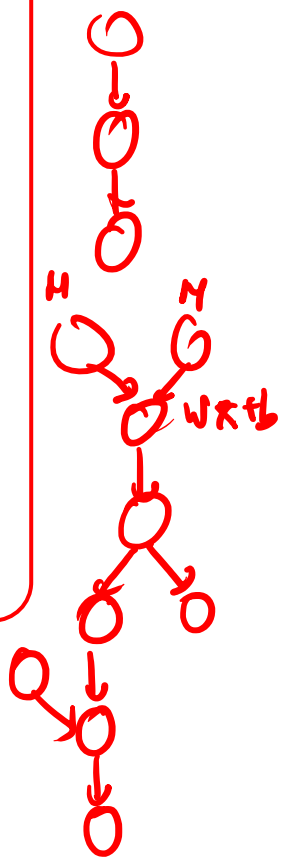


KEY IDEA

Any DGM where:

- nodes without parents are Gaussian *and*
- nodes with parents are linear combinations of their parents

is a multivariate Gaussian.



Remark: lets us leverage the nice properties of Gaussians and simplifies analysis/derivations.

Multivariate Normal Distribution

- Multivariate normal distribution describes a *D*-dimensional continuous variable $\boldsymbol{X}$, i.e. $\boldsymbol{x} \in \mathbb{R}^D$.
- *D*-dimensional mean $\boldsymbol{\mu} \in \mathbb{R}^D$, and $D \times D$ symmetrical positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}_+^{D \times D}$.

$$p(\boldsymbol{X} = \boldsymbol{a} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{a} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{a} - \boldsymbol{\mu}) \}, \quad \boldsymbol{a} \in \mathbb{R}^D$$

Or

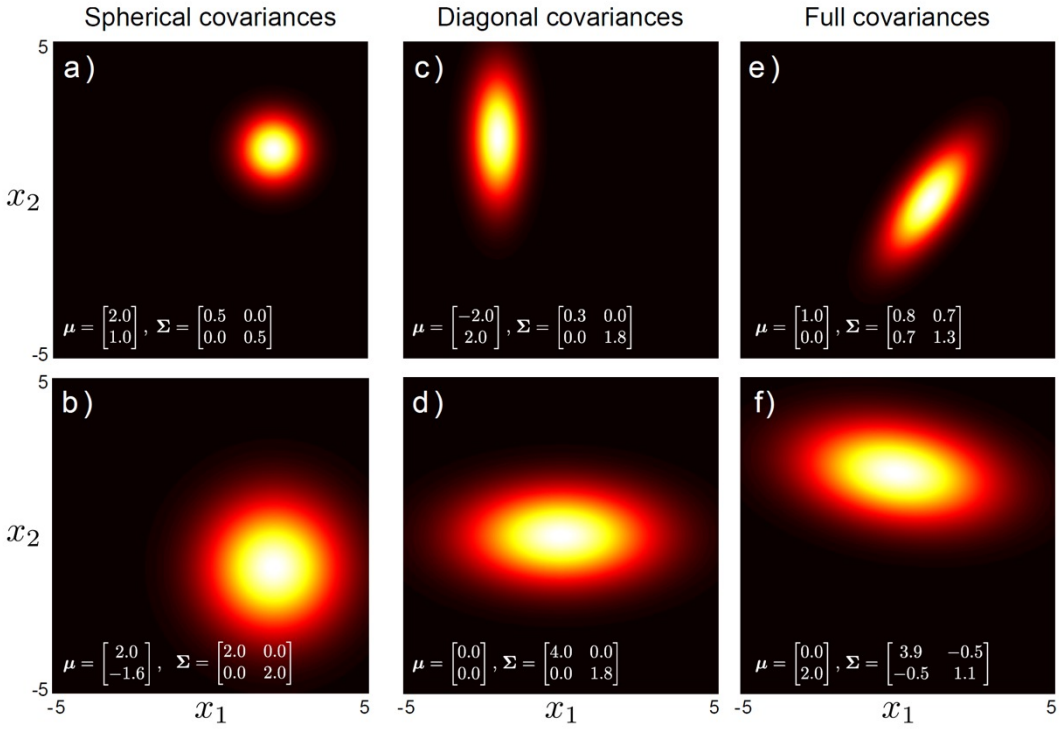
$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{ -0.5(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \}$$

$$p(\boldsymbol{x}) = \mathrm{Norm}_{\boldsymbol{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

Types of Covariance

- Covariance matrix has three forms: spherical, diagonal and full.

$$\Sigma_{spher} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \quad \Sigma_{diag} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \Sigma_{full} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$



Images Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Problem 2. (Linear Gaussian)

For this tutorial problem, we will consider a specific DGM that is the basis for more sophisticated models such as Probabilistic PCA and Linear Dynamical Systems. This model is called the Linear-Gaussian Model. *Note:* for this problem, we will be denoting random variables with lower case letters, and bolded lowercase letters to represent vectors, and bolded uppercase letters to represent matrices.

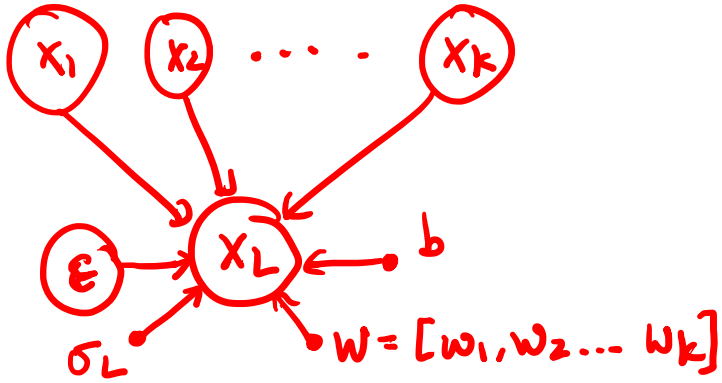
Problem 2.a. We will build our way up towards this model. As a prelude, consider K *independent* univariate Gaussian random variables $x_1, x_2, \ldots, x_K$,

$$p(x_k) = \mathcal{N}(\mu_k, \sigma_k^2)$$

for $k = 1, 2, \ldots, K$. Define the random variable x_L,

$$x_L = b + \sigma_L \epsilon + \sum_{k=1}^{K} w_k x_k$$

where $\epsilon \sim \mathcal{N}(0, 1)$.



1. Draw out the DGM for the model described above.
2. Show that $p(x_L|x_1, \ldots, x_K) = \mathcal{N}\left(b + \sum_{k=1}^{K} w_k x_k, \sigma_L^2\right)$. In other words, x_L is Gaussian distributed with mean $b + \sum_{k=1}^{K} w_k x_k$ and variance σ_L^2.
3. Define the random variable $\mathbf{x} = (x_1, x_2, \ldots, x_K, x_L)$. Show that $\mathbf{x}$ is a *multivariate* Gaussian random variable. *Hint: Consider the definition of the multivariate Gaussian and the properties of Gaussians.*

Problem 2.a. We will build our way up towards this model. As a prelude, consider K *independent* univariate Gaussian random variables $x_1, x_2, \ldots, x_K$,

$$p(x_k) = \mathcal{N}(\mu_k, \sigma_k^2)$$

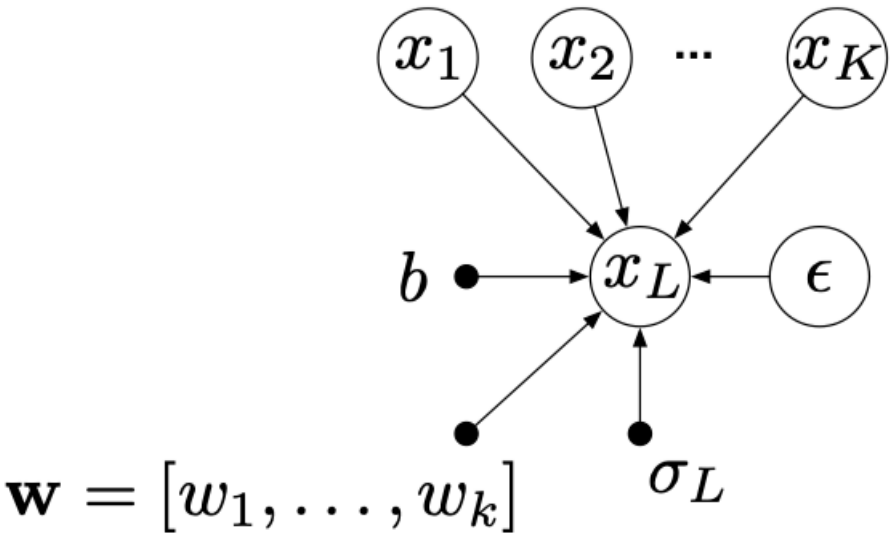
for $k = 1, 2, \ldots, K$. Define the random variable x_L,

$$x_L = b + \sigma_L \epsilon + \sum_{k=1}^{K} w_k x_k$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

1. Draw out the DGM for the model described above.

Solution:



$$p(x_L, x_1, x_2, x_3, \ldots, x_K) = p(x_L | x_1, x_2, \ldots, x_k) \prod_k p(x_k)$$

2. Show that $p(x_L|x_1,\ldots,x_K) = \mathcal{N}\left(b+\sum_{k=1}^{K} w_k x_k, \sigma_L^2\right)$. In other words, x_L is Gaussian distributed with mean $b+\sum_{k=1}^{K} w_k x_k$ and variance σ_L^2.

Solution:

Since ϵ is Gaussian, a linear function of it is still Gaussian.

$$\mathbb{E}\left[x_L|x_1,\cdots,x_K\right] = \mathbb{E}\left[b+\sigma_L\epsilon+\sum_{k=1}^{K} w_k x_k\right]$$

$$= \mathbb{E}[b] + \sigma_L\mathbb{E}[\epsilon] + \sum_{k=1}^{K} w_k \mathbb{E}[x_k]$$

$$\mathbb{E}\left[x_L|x_1,\cdots,x_K\right] = b+\sum_{k=1}^{K} w_k x_k$$

2. Show that $p(x_L|x_1,\ldots,x_K) = \mathcal{N}\left(b+\sum_{k=1}^{K} w_k x_k, \sigma_L^2\right)$. In other words, x_L is Gaussian distributed with mean $b+\sum_{k=1}^{K} w_k x_k$ and variance σ_L^2.

Solution:

$$\begin{aligned}
\mathrm{Var}\left[x_L|x_1,\cdots,x_K\right] &= \mathrm{Var}\left[b+\sigma_L\epsilon+\sum_{k=1}^{K} w_k x_k\right] \\
&= \mathrm{Var}[b]+\sigma_L^2\mathrm{Var}[\epsilon]+\sum_{k=1}^{K} w_k^2\mathrm{Var}[x_k] \\
&= \sigma_L^2\mathrm{Var}[\epsilon] = \sigma_L^2
\end{aligned}$$

3. Define the random variable $\mathbf{x} = (x_1, x_2, \ldots, x_K, x_L)$. Show that $\mathbf{x}$ is a *multivariate* Gaussian random variable. *Hint: Consider the definition of the multivariate Gaussian and the properties of Gaussians.*

Solution:

Strategy: we show that:

$$p(\mathbf{x}) \propto \exp(-(\mathbf{x}-\mathbf{b})^\top \mathbf{A}(\mathbf{x}-\mathbf{b}))$$

We will rearrange terms around to show this.

Denote $\mathbf{x}_{\pi_i} = (x_1, \cdots, x_K)$, $\mathbf{w} = [w_1, \cdots, w_K]^T$

$$p(\mathbf{x}_{\pi_i}) \propto \exp\left\{-\frac{1}{2}\left[\sum_{k \in \mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k - \mu_k)^2\right]\right\}$$

$$p(x_L|\mathbf{x}_{\pi_i}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma_L^2}(x_L - (b + \sum_{k \in \mathbf{x}_{\pi_i}} w_k x_k))^2\right]\right\}$$

Recall: $p(x_L, x_1, x_2, x_3, \ldots, x_K) = p(x_L|x_1, x_2, \ldots, x_k) \prod_k p(x_k)$

$$p(\mathbf{x}_{\pi_i})p(x_L|\mathbf{x}_{\pi_i}) \propto \exp\left\{-\frac{1}{2}\left[\sum_{k \in \mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k - \mu_k)^2 + \frac{1}{\sigma_L^2}(x_L - (b + \sum_{k \in \mathbf{x}_{\pi_i}} w_k x_k))^2\right]\right\}$$

We look at the terms in exponential,

$$\sum_{k \in \mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k - \mu_k)^2 + \frac{1}{\sigma_L^2}(x_L - (b + \sum_{k \in \mathbf{x}_{\pi_i}} w_k x_k))^2$$

$$\mu_L = b + \mathbf{w}^T \mathbf{x}_{\pi_i},\ \mathbf{\Sigma}_L = [\sigma_L^2] \text{ and } \mathbf{\Sigma}_{\pi_i} = \text{diagonal}(\sigma_1^2, \cdots, \sigma_K^2).$$

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \mathbf{\Sigma}_{\pi_i}^{-1} (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i}) + \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)$$

We look at the terms in exponential,

$$\sum_{k \in \mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k - \mu_k)^2 + \frac{1}{\sigma_L^2}(x_L - (b + \sum_{k \in \mathbf{x}_{\pi_i}} w_k x_k))^2$$

$$\mu_L = b + \mathbf{w}^T \mathbf{x}_{\pi_i},\ \mathbf{\Sigma}_L = [\sigma_L^2] \text{ and } \mathbf{\Sigma}_{\pi_i} = \text{diagonal}(\sigma_1^2, \cdots, \sigma_K^2).$$

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \mathbf{\Sigma}_{\pi_i}^{-1} (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i}) + \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)$$

(Hint: add in $+\mathbf{w}^T \boldsymbol{\mu}_{\pi_i} - \mathbf{w}^T \boldsymbol{\mu}_{\pi_i}$ and rearrange)

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \mathbf{\Sigma}_{\pi_i}^{-1} (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})$$
$$+ \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i}) - \mathbf{w}^T (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i}) - \mathbf{w}^T (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})\right)$$

We look at the terms in exponential,

$$\sum_{k \in \mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k - \mu_k)^2 + \frac{1}{\sigma_L^2}(x_L - (b + \sum_{k \in \mathbf{x}_{\pi_i}} w_k x_k))^2$$

$$\mu_L = b + \mathbf{w}^T \mathbf{x}_{\pi_i},\ \mathbf{\Sigma}_L = [\sigma_L^2] \text{ and } \mathbf{\Sigma}_{\pi_i} = \text{diagonal}(\sigma_1^2, \cdots, \sigma_K^2).$$

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \mathbf{\Sigma}_{\pi_i}^{-1} (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i}) + \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \mathbf{x}_{\pi_i})\right)$$

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \mathbf{\Sigma}_{\pi_i}^{-1} (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})$$
$$+ \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i}) - \mathbf{w}^T (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i}) - \mathbf{w}^T (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})\right)$$

$$= (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \left[\mathbf{\Sigma}_{\pi_i}^{-1} + \boldsymbol{w} \mathbf{\Sigma}_L^{-1} \boldsymbol{w}^T\right] (\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})$$
$$+ \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i})\right)^T \mathbf{\Sigma}_L^{-1} \left(x_L - (b + \mathbf{w}^T \boldsymbol{\mu}_{\pi_i})\right) - 2(\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i})^T \boldsymbol{w} \mathbf{\Sigma}_L^{-1} (x_L - \mu_L)$$

(Hint: Expand the square. Then group the $\mathbf{x}_{\pi_i} - \boldsymbol{\mu}_{\pi_i}$ along with the $\boldsymbol{w}\Sigma_L^{-1}\boldsymbol{w}^\top$ into the first term)

We look at the terms in exponential,

$$\sum_{k\in\mathbf{x}_{\pi_i}} \frac{1}{\sigma_k^2}(x_k-\mu_k)^2+\frac{1}{\sigma_L^2}(x_L-(b+\sum_{k\in\mathbf{x}_{\pi_i}} w_k x_k))^2$$

$$\mu_L = b+\mathbf{w}^T\mathbf{x}_{\pi_i},\ \mathbf{\Sigma}_L=[\sigma_L^2] \text{ and } \mathbf{\Sigma}_{\pi_i}=\text{diagonal}(\sigma_1^2,\cdots,\sigma_K^2).$$

$$=(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})^T\mathbf{\Sigma}_{\pi_i}^{-1}(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})+\left(x_L-(b+\mathbf{w}^T\mathbf{x}_{\pi_i})\right)^T\mathbf{\Sigma}_L^{-1}\left(x_L-(b+\mathbf{w}^T\mathbf{x}_{\pi_i})\right)$$

$$=(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})^T\mathbf{\Sigma}_{\pi_i}^{-1}(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})$$
$$+\left(x_L-(b+\mathbf{w}^T\boldsymbol{\mu}_{\pi_i})-\mathbf{w}^T(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})\right)^T\mathbf{\Sigma}_L^{-1}\left(x_L-(b+\mathbf{w}^T\boldsymbol{\mu}_{\pi_i})-\mathbf{w}^T(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})\right)$$

$$=(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})^T\left[\mathbf{\Sigma}_{\pi_i}^{-1}+\boldsymbol{w}\mathbf{\Sigma}_L^{-1}\boldsymbol{w}^T\right](\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})$$
$$+\left(x_L-(b+\mathbf{w}^T\boldsymbol{\mu}_{\pi_i})\right)^T\mathbf{\Sigma}_L^{-1}\left(x_L-(b+\mathbf{w}^T\boldsymbol{\mu}_{\pi_i})\right)-2(\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i})^T\boldsymbol{w}\mathbf{\Sigma}_L^{-1}(x_L-\mu_L)$$

$$=\begin{bmatrix}\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i}\\ x_L-\mu_L\end{bmatrix}^T\begin{bmatrix}\mathbf{\Sigma}_{\pi_i}^{-1}+\boldsymbol{w}\mathbf{\Sigma}_L^{-1}\boldsymbol{w}^T & \mathbf{\Sigma}_L^{-1}\boldsymbol{w}^T\\ \boldsymbol{w}\mathbf{\Sigma}_L^{-1} & \mathbf{\Sigma}_L^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{\pi_i}-\boldsymbol{\mu}_{\pi_i}\\ x_L-\mu_L\end{bmatrix}$$

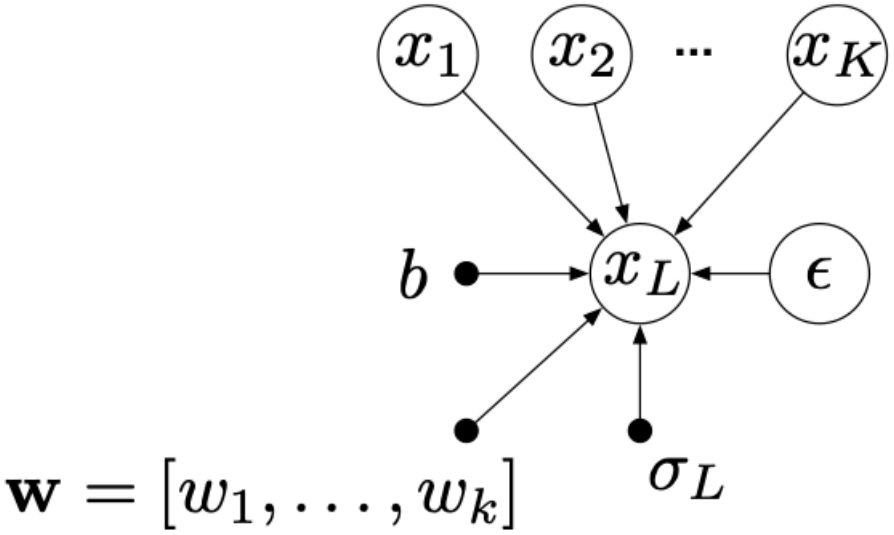
$(\tilde{x}-\tilde{b})$ A $(x-b)$

Rearrange into matrix form

The solution sheet additionally shows that the matrix is positive definite.

Recap: What have we shown?

- Linear combinations of Gaussian random variables are Gaussian.
- The vector of variables is multivariate Gaussian.



For this structure

- All marginals are Gaussian.

If:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_x)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x})$$

then:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\Sigma}_{x|y}\left(\mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}\right), \boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_{y|x}^{-1}\mathbf{A}\right)^{-1}$$

For multivariate Gaussians

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}.$$

Conditional distribution:

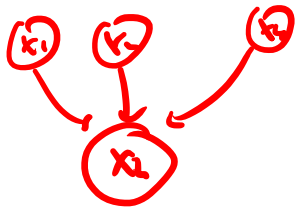
$$\begin{aligned} p(\mathbf{x}_a|\mathbf{x}_b) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \end{aligned}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}).$$

From Bishop PRML Chap 2.3.2

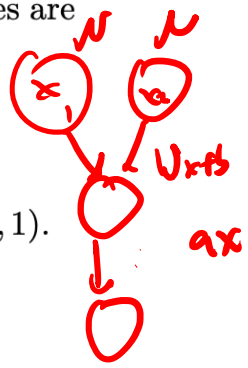
Extension to General Case



Problem 2.b. Let's now move to the more complex case. Consider an *arbitrary* DGM G where each node j without any parents is Gaussian distributed with mean μ_j and variance σ_j^2. The remaining nodes are defined as

$$x_i = b_i + \left(\sum_{j \in x_{\pi_i}} w_{i,j} x_j \right) + \sigma_i \epsilon_i$$

where x_{π_i} denotes the set of node i's parents and ϵ_i is the standard normal random variable $\epsilon_i \sim \mathcal{N}(0, 1)$.



1. Show that each node x_i has the conditional distribution: $p(x_i | x_{\pi_i}) = \mathcal{N}\left(b_i + \sum_{j \in x_{\pi_i}} w_{i,j} x_j, \sigma_i^2\right)$
2. Define the random variable $\mathbf{x} = (x_1, x_2, \ldots, x_D)$. Show that $\mathbf{x}$ is a *multivariate Gaussian*.

2. Define the random variable $\mathbf{x} = (x_1, x_2, \ldots, x_D)$. Show that $\mathbf{x}$ is a *multivariate Gaussian.*

Proof Sketch:

We prove by *induction* using the previous results in Problem 2.a. as a building block. We proceed in topological order, adding nodes into the graph one by one.

Base Case: the first node is Gaussian (by definition).

Inductive Hypothesis: $\mathbf{x}_{1:K}$ is multivariate Gaussian.

Inductive Step: (show for K+1 nodes). 2 cases:

- *Case 1*: x_{K+1} has no parents. Then the $\mathbf{x}_{1:K+1}$ is multivariate Gaussian (verify this).
- *Case 2:* x_{K+1} has parents. Since we proceed in topological order, the parents must be among $x_1, x_2, \ldots, x_K$ and are Gaussian. Using our previous result in Problem 2.a., $x_{1:K+1}$ is multivariate Gaussian.

What did we learn?

KEY IDEA

Any DGM where:

- nodes without parents are Gaussian *and*
- nodes with parents are linear combinations of their parents

is a multivariate Gaussian.

Remark: lets us leverage the nice properties of Gaussians and simplifies analysis/derivations.

Next...

We can even evaluate the exact parameters of this multivariate Gaussian.

Problem 2.c. We can determine the mean of $\mathbf{x}$ using a recursive method. Note that $\mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_1], \ldots, \mathbb{E}[x_D])^\top$. Show that the expectation of each component $\mathbb{E}[x_i]$ is given by:

$$\mathbb{E}[x_i] = b_i + \sum_{j \in x_{\pi_i}} w_{i,j} \mathbb{E}[x_j]$$

Solution: Since $x_i = b_i + \left(\sum_{j \in x_{\pi_i}} w_{i,j} x_j\right) + \sigma_i \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, 1)$

$$\begin{aligned}
\mathbb{E}[x_i] &= \mathbb{E}\left[b_i + \left(\sum_{j \in x_{\pi_i}} w_{i,j} x_j\right) + \sigma_i \epsilon_i\right] \\
&= \mathbb{E}[b_i] + \sum_{j \in x_{\pi_i}} w_{i,j} \mathbb{E}[(x_j)] + \sigma_i \mathbb{E}[\epsilon_i] \\
&= b_i + \sum_{j \in x_{\pi_i}} w_{i,j} \mathbb{E}[(x_j)]
\end{aligned}$$

Problem 2.d. Likewise, we can determine the covariance matrix of $\mathbf{x}$. Note that

$$\boldsymbol{\Sigma}_{ij} = \text{Cov}[x_i, x_j] = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])]$$

1. Show that $\text{Cov}[x_i, x_j] = I_{ij}\sigma_j^2 + \sum_{k \in x_{\pi_j}} w_{j,k}\text{Cov}[x_i, x_k]$
2. If the DGM G has no edges, is the covariance matrix $\boldsymbol{\Sigma}$ a spherical, diagonal, or general symmetric covariance matrix? How many parameters does it have?

Solution:

$$\begin{aligned}
&\text{Cov}[x_i, x_j] \\
&= E[(x_i - E[x_i])(x_j - E[x_j])] \\
&= E[(x_i - E[x_i])(b_j + \sigma_j\epsilon_j + \textstyle\sum_k w_{jk}x_k - b_j - \sum_k w_{jk}E[x_k])] \\
&= E[(x_i - E[x_i])(\sigma_j\epsilon_j + \textstyle\sum_k w_{jk}x_k - \sum_k w_{jk}E[x_k])] \\
&= E[(x_i - E[x_i])\sigma_j\epsilon_j + (\textstyle\sum_k w_{jk}(x_i - E[x_i])(x_k - E[x_k])] \\
&= E[(x_i - E[x_i])\sigma_j\epsilon_j] + E[\textstyle\sum_k w_{jk}(x_i - E[x_i])(x_k - E[x_k])] \\
&= E[(x_i - E[x_i])\sigma_j\epsilon_j] + \textstyle\sum_k w_{jk}E[(x_i - E[x_i])(x_k - E[x_k])] \\
&= E[(x_i - E[x_i])\sigma_j\epsilon_j] + \textstyle\sum_k w_{jk}\text{Cov}[x_i, x_k]
\end{aligned}$$

Problem 2.d. Likewise, we can determine the covariance matrix of $\mathbf{x}$. Note that

$$\boldsymbol{\Sigma}_{ij} = \text{Cov}[x_i, x_j] = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])]$$

1. Show that $\text{Cov}[x_i, x_j] = I_{ij}\sigma_j^2 + \sum_{k \in x_{\pi_j}} w_{j,k}\text{Cov}[x_i, x_k]$
2. If the DGM G has no edges, is the covariance matrix $\boldsymbol{\Sigma}$ a spherical, diagonal, or general symmetric covariance matrix? How many parameters does it have?

Solution:

$$E\left[(x_i - E[x_i])\sigma_j\epsilon_j\right]$$

$$= E\left[x_i\sigma_j\epsilon_j\right] - E\left[E[x_i]\sigma_j\epsilon_j\right]$$

$$= E\left[x_i\sigma_j\epsilon_j\right]$$

$$= E\left[\left(b_j + \textstyle\sum_k w_{ik}x_k + \sigma_i\epsilon_i\right)\sigma_j\epsilon_j)\right]$$

$$= E\left[b_j\sigma_j\epsilon_j\right] + E\left[\sigma_j\epsilon_j \textstyle\sum_k w_{ik}x_k\right] + E\left[\sigma_i\sigma_j\epsilon_i\epsilon_j\right]$$

$$= E[\sigma_i\sigma_j\epsilon_i\epsilon_j]$$

If $i = j$

$$E\left[\sigma_i\sigma_j\epsilon_i\epsilon_j\right] = \sigma_i^2 E\left[\epsilon_i^2\right] = \sigma_i^2$$

Else If $i \neq j$

$$E\left[\sigma_i\sigma_j\epsilon_i\epsilon_j\right] = \sigma_i\sigma_j E\left[\epsilon_i\epsilon_j\right] = 0$$

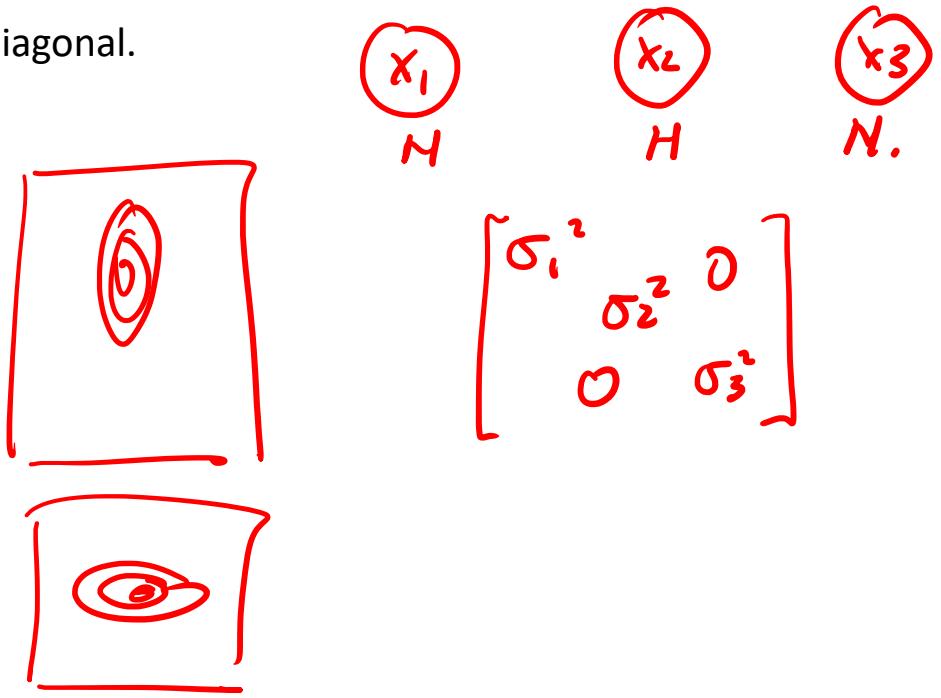
Problem 2.d. Likewise, we can determine the covariance matrix of $\mathbf{x}$. Note that

$$\mathbf{\Sigma}_{ij} = \text{Cov}[x_i, x_j] = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])]$$

1. Show that $\text{Cov}[x_i, x_j] = I_{ij}\sigma_j^2 + \sum_{k \in x_{\pi_j}} w_{j,k}\text{Cov}[x_i, x_k]$
2. If the DGM G has no edges, is the covariance matrix $\mathbf{\Sigma}$ a spherical, diagonal, or general symmetric covariance matrix? How many parameters does it have?
3. If the DGM G is fully-connected, what kind of matrix is the covariance matrix $\mathbf{\Sigma}$? Is it spherical, diagonal, or a general symmetric covariance matrix? How many parameters does it have?

Solution:

If no edges, the covariance is diagonal.



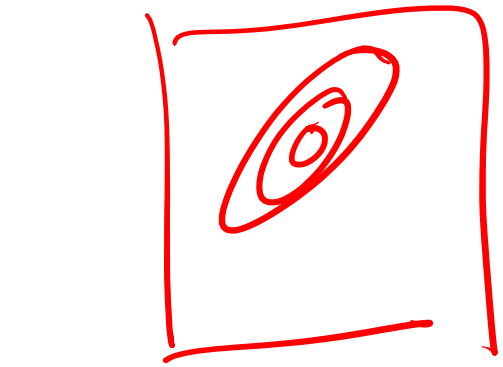
Problem 2.d. Likewise, we can determine the covariance matrix of $\mathbf{x}$. Note that

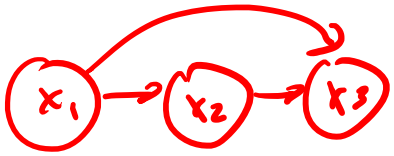
$$\mathbf{\Sigma}_{ij} = \text{Cov}[x_i, x_j] = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])]$$

1. Show that $\text{Cov}[x_i, x_j] = I_{ij}\sigma_j^2 + \sum_{k \in x_{\pi_j}} w_{j,k}\text{Cov}[x_i, x_k]$
2. If the DGM G has no edges, is the covariance matrix $\mathbf{\Sigma}$ a spherical, diagonal, or general symmetric covariance matrix? How many parameters does it have?
3. If the DGM G is fully-connected, what kind of matrix is the covariance matrix $\mathbf{\Sigma}$? Is it spherical, diagonal, or a general symmetric covariance matrix? How many parameters does it have?

Solution:

If fully-connected, then you get a general symmetric covariance matrix. It has D(D+1)/2 parameters





$$\Sigma = \begin{bmatrix} \text{cov}[x_1,x_1] & \text{cov}[x_1,x_2] & \text{cov}[x_1,x_3] \\ \text{cov}[x_2,x_1] & \text{cov}[x_2,x_2] & \text{cov}[x_2,x_3 \\ \cdot & \cdot & \cdot \end{bmatrix}$$

$$\begin{bmatrix} \sigma_1^2 & w_2\sigma_1^2 & \cdot \\ \cdot & \sigma_2^2 + w_2\sigma_1^2 & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$$

Questions?

https://pollev.com/haroldsohsoo986

Homework!

- Week 8
  - Week 8 Summary
  - Slides
    - L8-HMM.pdf
  - Video Lectures
    - L8 - Part 1 (Intro)
    - L8 - Part 2 (Intro to HMMs)
    - L8 - Part 3 (EM for HMMs)
    - L8 - Part 4 (Forward Backward Alg)
    - L8 - Part 5 (Code and Numerical Issues)
    - L8 - Part 6 (More HMM Inference)
    - L8 - Part 7 (Recap and Extensions)
  - Extra Readings
    - Raibiner's Tutorial on HMMs.pdf
    - HMMs (from Speech and Language Processing).pdf
  - Tutorial (for next week)
    - Tut7_23.pdf