Introduction to Machine Learning – Generative Approach

The 8th KIAS CAC Summer School
2017. 7. 30 (Fri.)

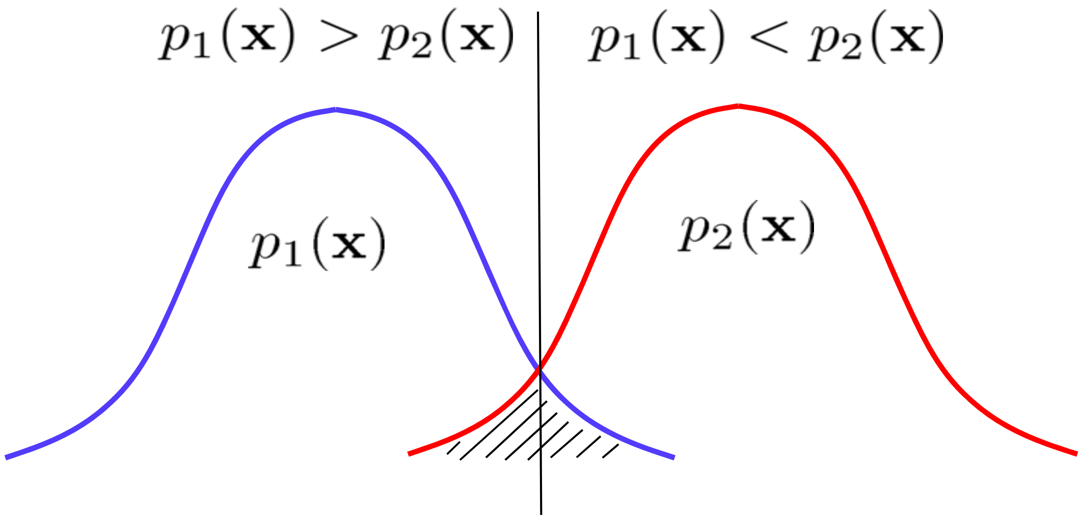
Yung-Kyun Noh
Seoul National University

Overview

- Motivation from the classification perspective
- Independence and model complexity
- How to incorporate independency while the model structure is kept intact as much as possible
- Directed graphical model and undirected graphical model

Probabilistic Assumption and Bayes Classification

- Bayes classification produces theoretical minimum error

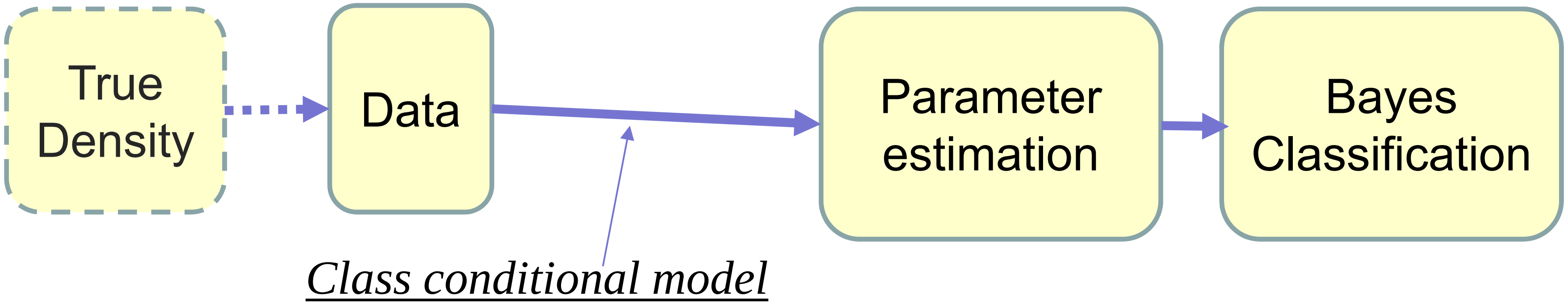


➔ Error:

$$\frac{1}{2}\int \min[p_1, p_2]\,d\mathbf{x}$$

Generative Model

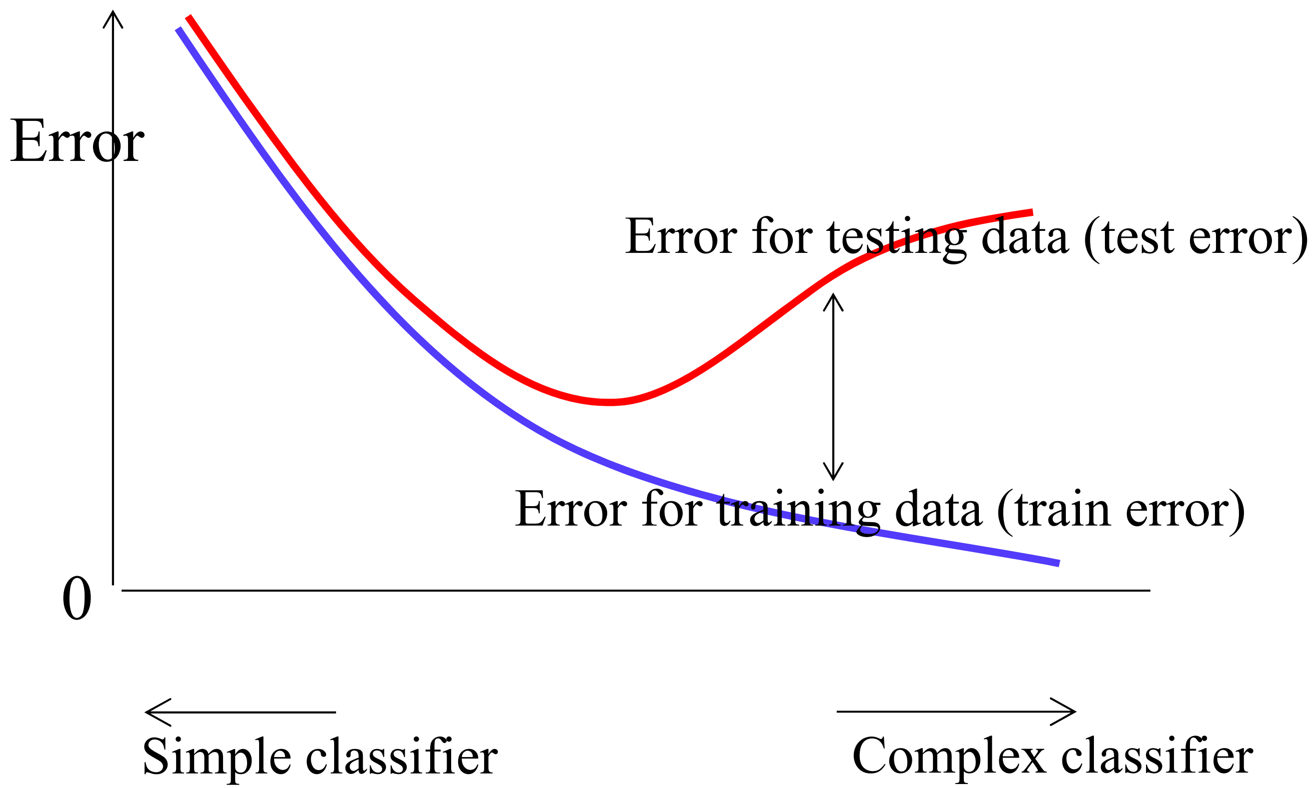
- Generative method for classification
  - Perform the Bayes classification
  - But we now use model instead of true underlying distribution



  - Bayes classification now uses model with estimated parameters, which is not true density, but is now considered as true density.

When do Algorithms Malfunction?

- Generalization and Overfitting



Complexity of Algorithms

- Statistical learning theory for classification
  - with probability at least 1 – δ
  - h: VC-dimension, n: number of data
  - *R(g)*: true risk of function *g*, $R_n(g)$: empirical risk of function *g*

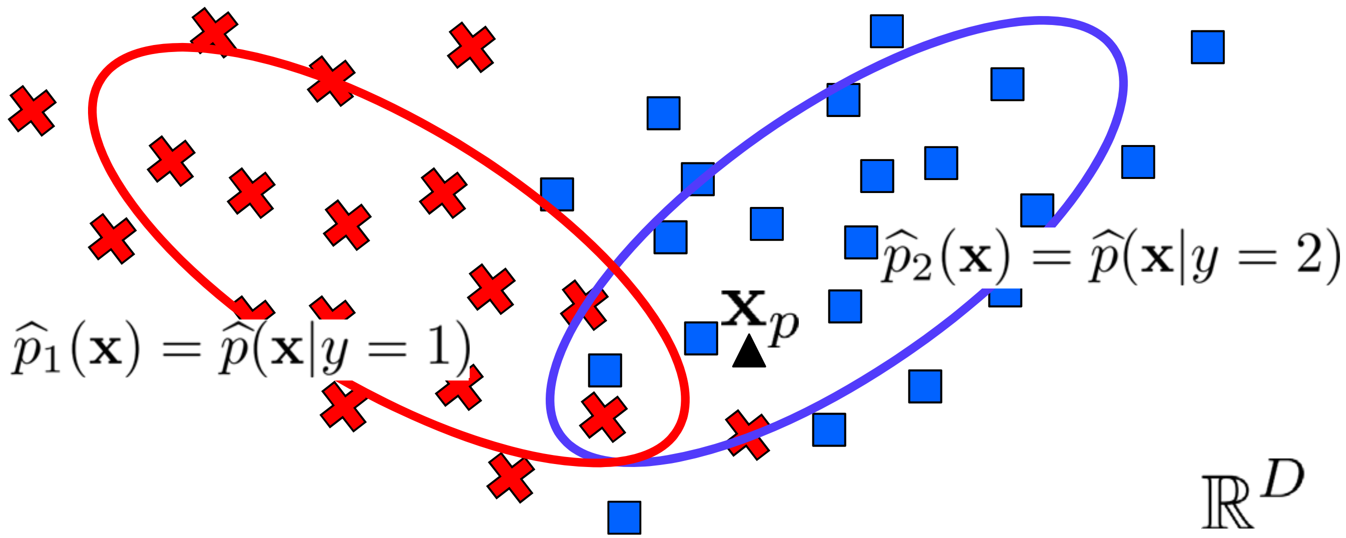
$$R(g) \leq R_n(g) + 2\sqrt{2\frac{h\log(2en/h) + \log(2/\delta)}{n}}$$

- Linear classifier
  - VC-dim = dimensionality + 1

- Generative model
  - Number of parameters

Model + Estimated Parameters

- Ex. Gaussian model



$$\widehat{p}_1(\mathbf{x}_p) \geq \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 1$$
$$\widehat{p}_1(\mathbf{x}_p) < \widehat{p}_2(\mathbf{x}_p) \rightarrow y_p = 2$$

Model + Estimated Parameters

- Ex. Gaussian model

$$\widehat{p}(\mathbf{x}) = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma}) \qquad \mathbf{x}, \widehat{\mu} \in \Re^{D}, \; \widehat{\Sigma} \in \Re^{D \times D}$$

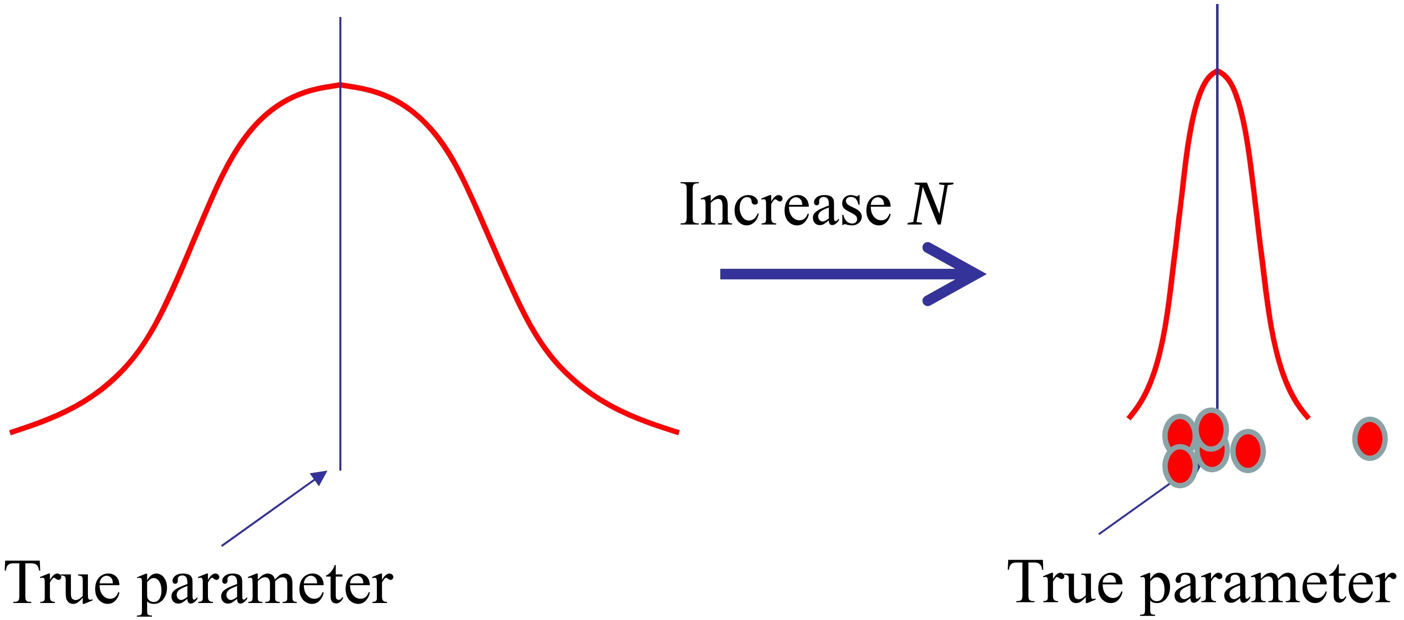
$$= \frac{1}{\sqrt{2\pi}^{D} |\widehat{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x} - \widehat{\mu})^{\top} \widehat{\Sigma}^{-1} (\mathbf{x} - \widehat{\mu})^{\top} \right)$$

- Unbiased estimators

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \qquad \widehat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^{\top}$$

Unbiased Estimation

- Consistency



Theory: Cramer-Rao bound
Minimum variance of covariance estimator: σ^2/n

Covariance Estimation

- In high-dimensional space

$$\Sigma_{D\times D} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1D} \\ \sigma_{21} & \sigma_2^2 & & & \vdots \\ \vdots & & \ddots & \sigma_{ij} & \\ \vdots & & \sigma_{ij} & & \\ \sigma_{D1} & \cdots & & & \sigma_D^2 \end{pmatrix}$$

(D + 1)D/2 number of parameters for covariances

Number of Parameters

- $D = 1000$

  – Number of parameters of a Gaussian:
  1000 + 1001*(1000)/2 = 501,500

Independence

- $$p(\mathbf{x}) = p_1(\mathbf{x}_1)p_2(\mathbf{x}_2)$$

  $$\mathbf{x} \in \mathbb{R}^D, \mathbf{x}_1 \in \mathbb{R}^{D_1}, \mathbf{x}_2 \in \mathbb{R}^{D_2} \qquad D = D_1 + D_2$$

- D_1 = 500, D_2 = 500
  - Number of parameters

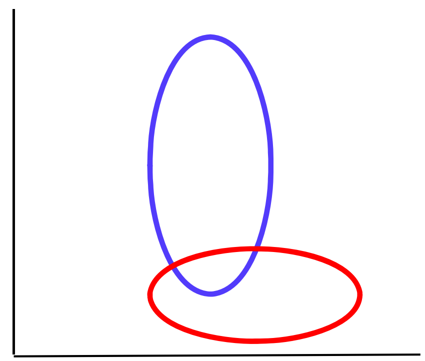
    500 + 501*(500)/2 + 500 + 501*(500)/2

    = 251,500

- Incorporating one independence can reduce the number of parameters into half.

Naïve Bayes As An Extreme Case

- Naïve Bayes



$$p(\mathbf{x}) = \prod_{d=1}^{D} p_d(x_d)$$
$$= p_1(x_1)p_2(x_2)\ldots p_D(x_D)$$

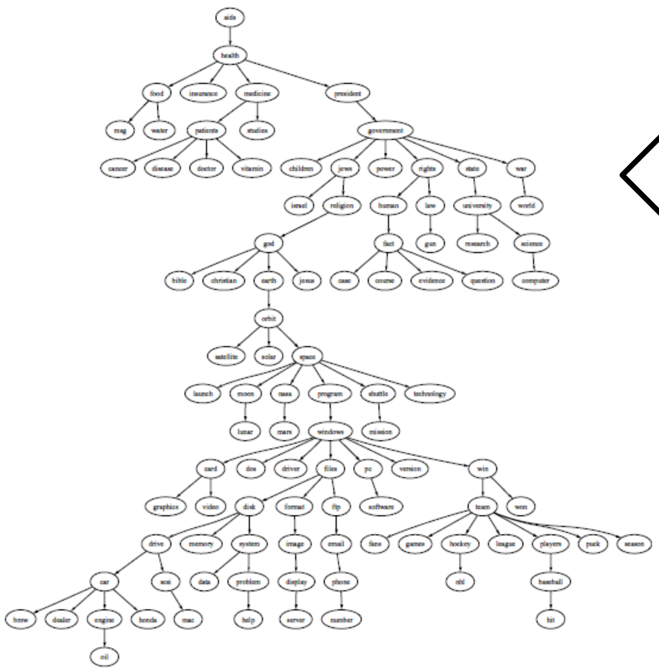
  – Simply ignore every correlation and dependencies between variables

- True decomposition:

$$p(\mathbf{x}) =$$
$$p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_D|x_1, \ldots, x_{D-1})$$

Graphical Models

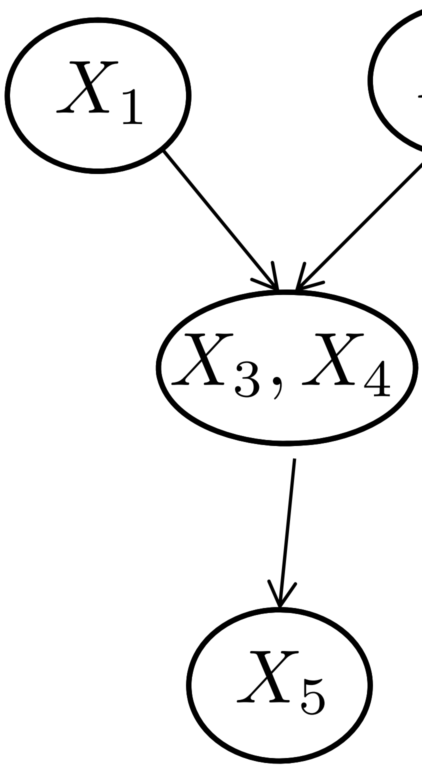
- We utilize probabilities that are represented by the graph structure. (directed & undirected)



Use probabilisic ***independencies*** and ***conditional independencies*** that can be captured by graph structure

Causality Graph

- Directed Acyclic Graph (DAG)

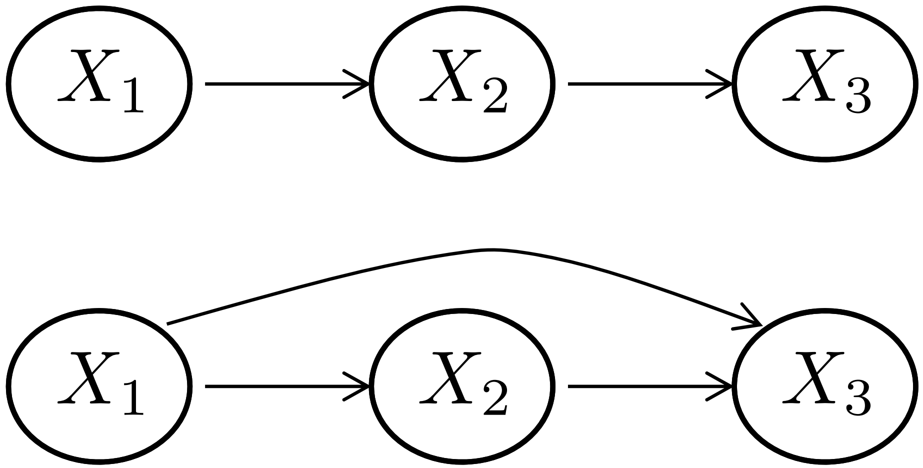


$$P(X_1, X_2, X_3, X_4, X_5) =$$
$$P(X_1)P(X_2)P(X_3, X_4|X_1, X_2)P(X_5|X_3, X_4)$$

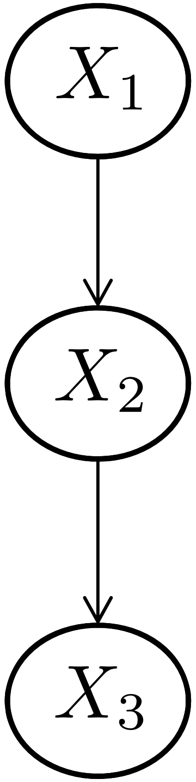
$$P(X_1, \ldots, X_D) = \prod_{i=1}^{D} P(X_i|\mathbf{Pa}_{X_i})$$

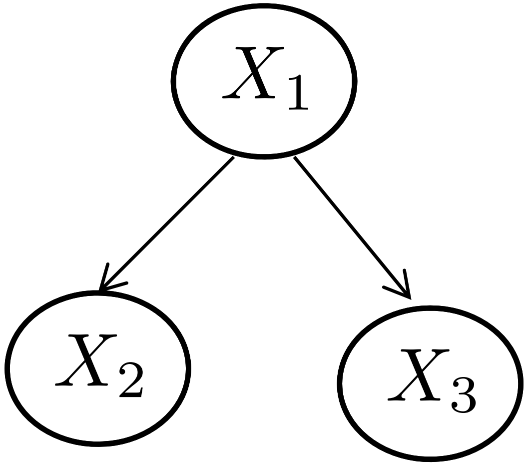
Question?

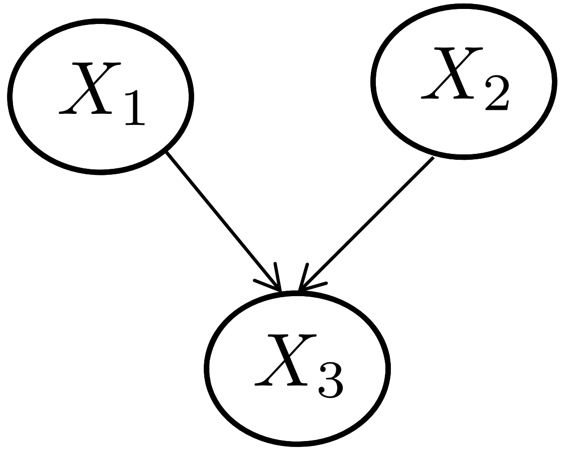
- What is the difference between two graphs?



D-Separations







$X_1 \perp\!\!\!\perp X_3 | X_2$

Causal path

$X_2 \perp\!\!\!\perp X_3 | X_1$

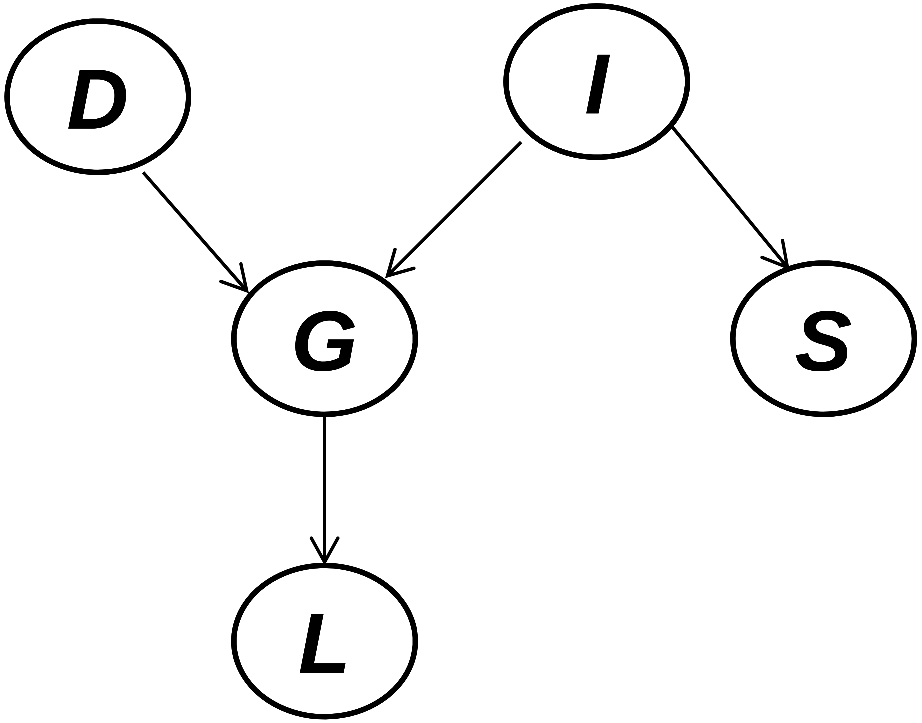
Common cause

$X_1 \perp\!\!\!\perp X_2$

$X_1 \not\perp\!\!\!\perp X_2 | X_3$

Common effect

Want to get a good reference letter?



D: *Difficulty*
I: *Intelligence*
G: *Grade*
S: *SAT*
L: *Reference Letter*

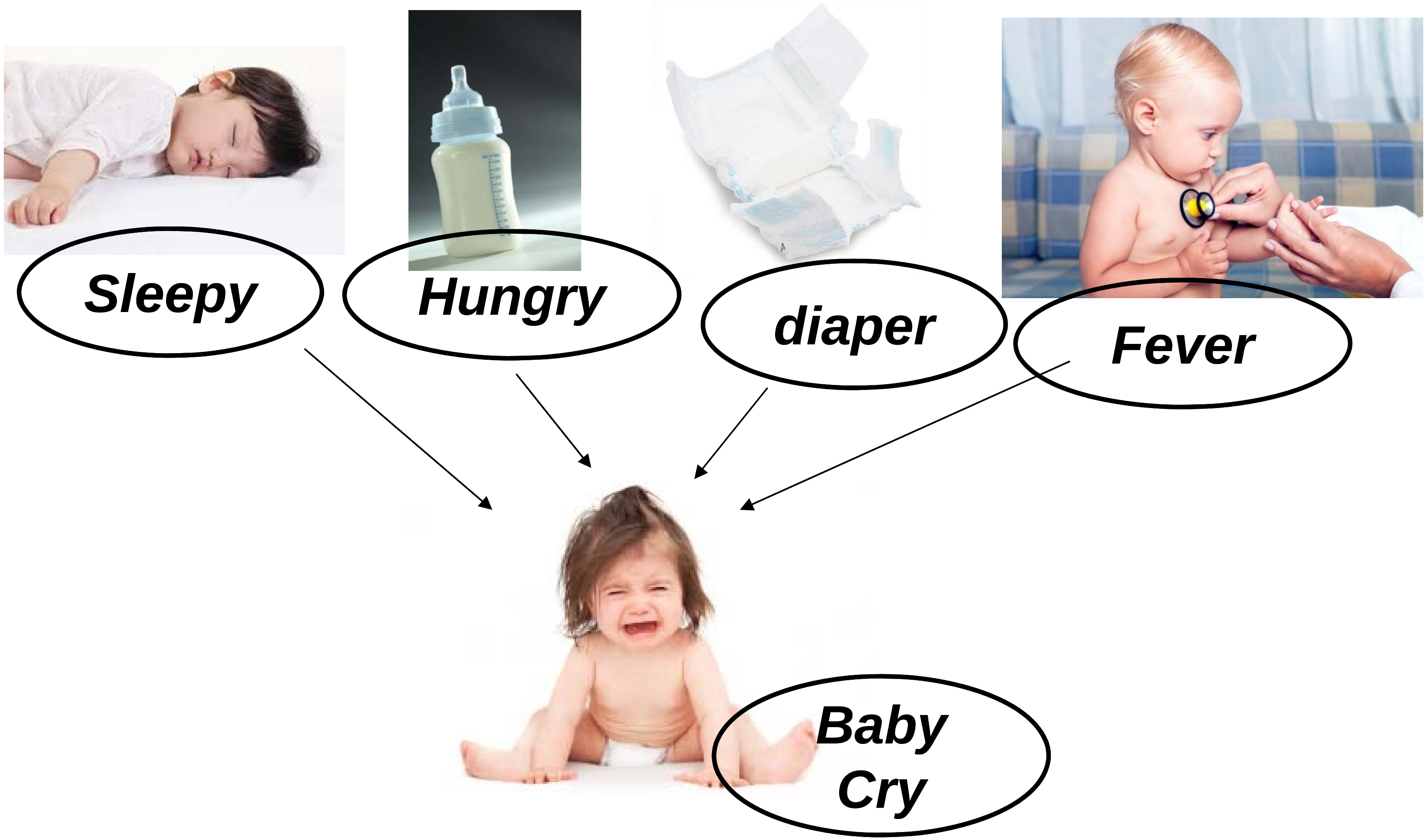
$$I \perp\!\!\!\perp L | G$$

$$G \perp\!\!\!\perp S | I$$

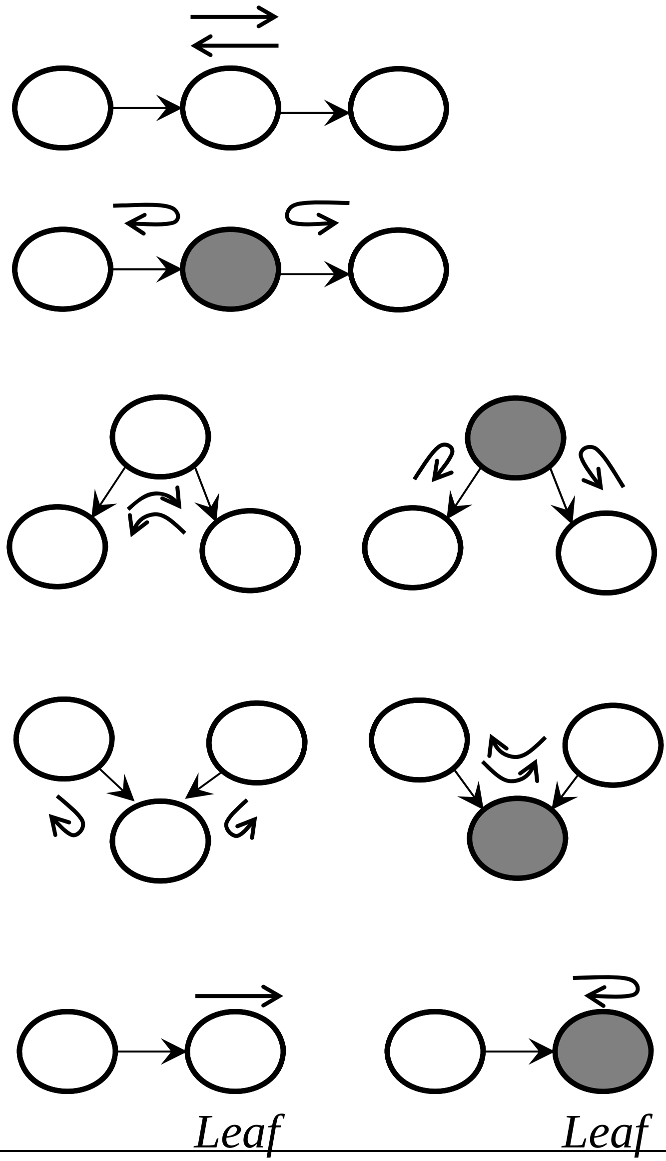
$$D \perp\!\!\!\perp I$$

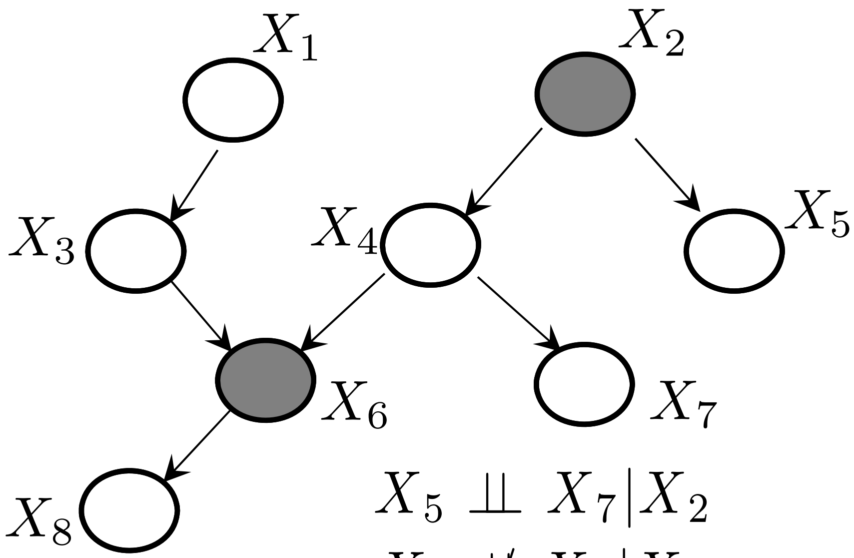
$$D \not\!\perp\!\!\!\perp I | G$$

Infant Rearing

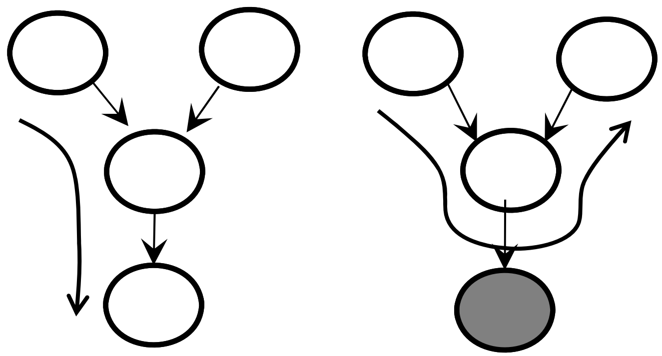


Bayes Ball Theorem



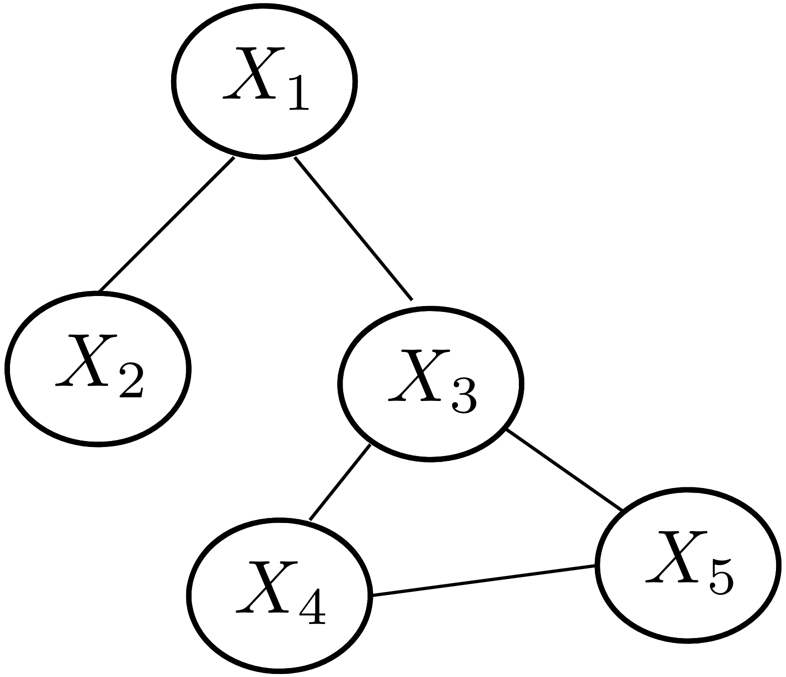


$$X_5 \perp\!\!\!\perp X_7 | X_2$$
$$X_1 \not\perp\!\!\!\perp X_7 | X_6$$
$$X_1 \perp\!\!\!\perp X_8 | X_6$$



Markov Random Field

- Undirected Graph



If there is a direct edge between X_i and X_j:

$$X_i \not\perp\!\!\!\perp X_j | X_{\setminus i,j}$$

If there is no direct edge between X_i and X_j:

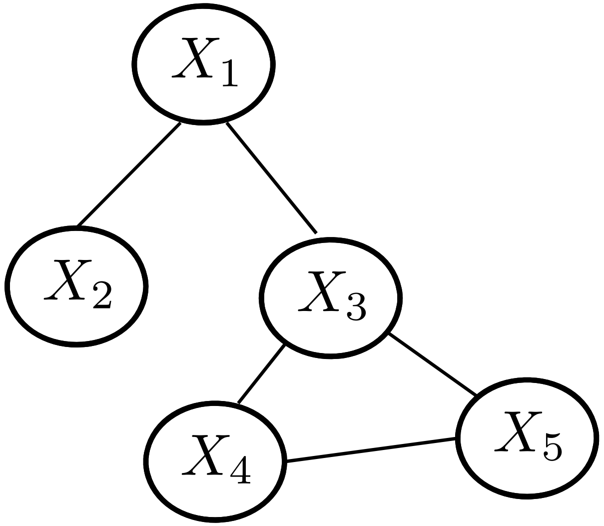
$$X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$$

$$X_1 \not\perp\!\!\!\perp X_3 | X_2, X_4, X_5$$

$$X_1 \perp\!\!\!\perp X_5 | X_3$$

Joint Distribution

- Product of functions on cliques

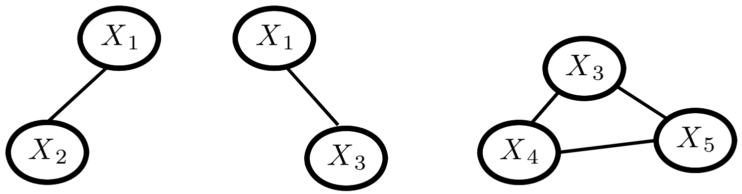


$$P(X_1, X_2, X_3, X_4, X_5) =$$

$$\frac{1}{Z}\ \psi_{1,2}(X_1, X_2)\ \psi_{1,3}(X_1, X_3)\ \psi_{3,4,5}(X_3, X_4, X_5)$$

$$\left(Z = \sum_{X_1, X_2, X_3, X_4, X_5} \psi_{1,2}(X_1, X_2)\psi_{1,3}(X_1, X_3)\psi_{3,4,5}(X_3, X_4, X_5) \right)$$

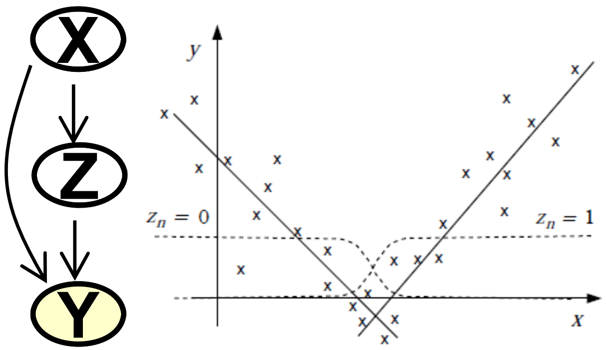
Cliques:

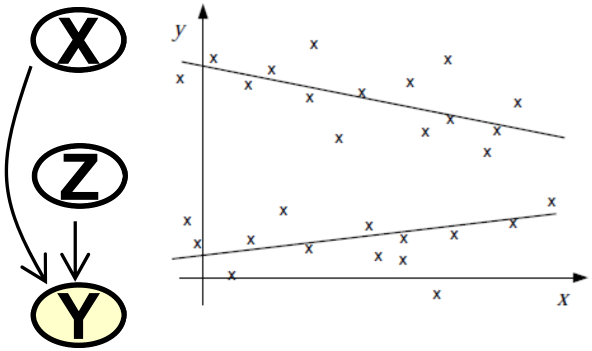


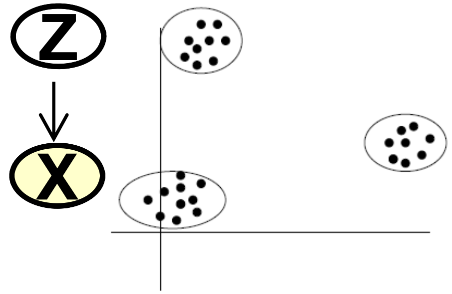
The set of distributions satisfying MRF conditions (Markov random field)
= The set of distributions decomposed by cliques (Gibbs random field)
(Hammersley-Clifford Theorem)

More Fancy Models

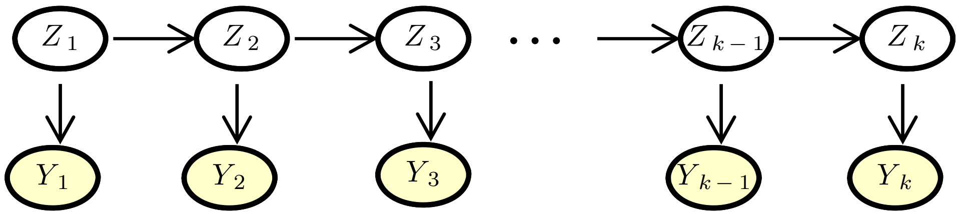
- Latent Variable Model



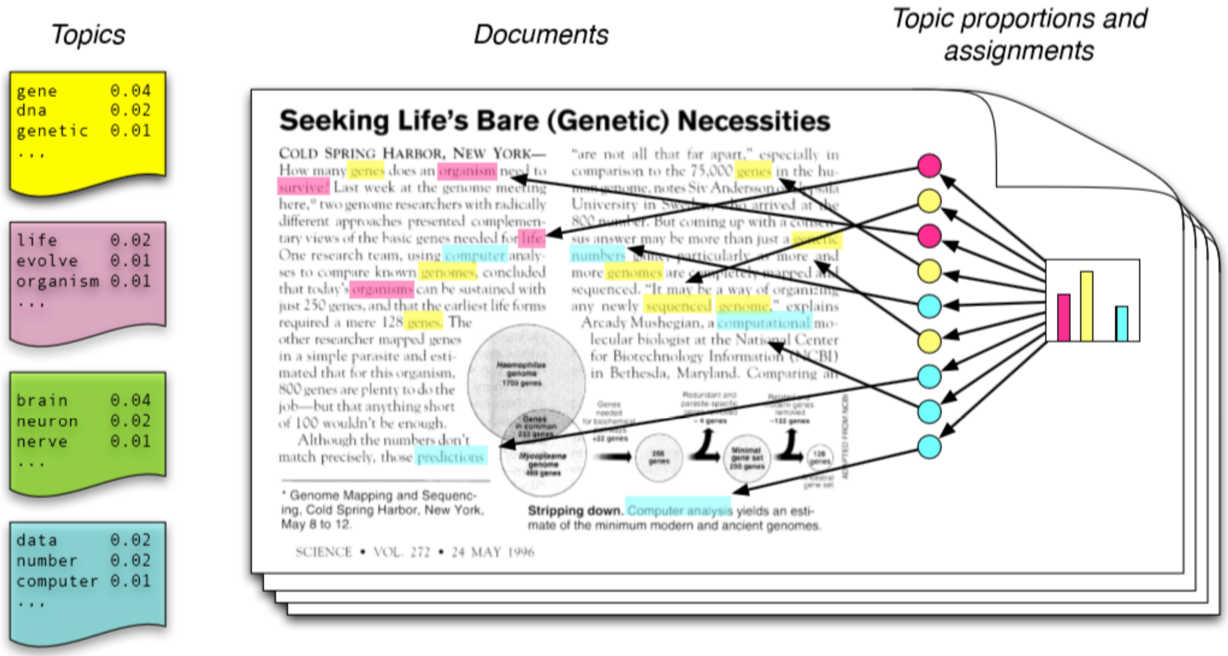


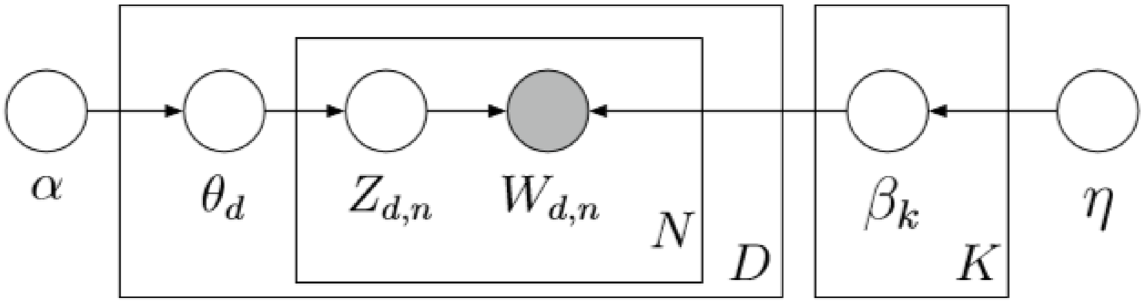


- Filtering



Topic Models





$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Topic Models

NIPS 2012 papers
(in nicer format than this)
maintained by @karpathy
source code on github

Below every paper are TOP 100 most-occuring words in that paper and their color is based on LDA topic model with k = 7.
(It looks like 0 = theory, 1 = reinforcement learning, 2 = graphical models, 3 = deep learning/vision, 4 = optimization, 5 = neuroscience, 6 = embeddings etc.)

Toggle LDA topics to sort by: TOPIC0 TOPIC1 TOPIC2 TOPIC3 TOPIC4 TOPIC5 TOPIC6

Discriminatively Trained Sparse Code Gradients for Contour Detection
Ren Xiaofeng, Liefeng Bo

[pdf] [bibtex] [supplementary]
[rank by tf-idf similarity to this]
[abstract]

[set, algorithm, including] [average, approach, benchmark, evaluation] [comparing, normal, hierarchical] [contour, gpb, local, detection, depth, scg, color, image, oriented, matching, contrast, object, grayscale, precision, recognition, transform, work, learned, pooling, pixel, representation, double, global, learn, accuracy, scale, level, segmentation, figure, feature, nyu, globalization, scene, training, rich, single, automatically, apply, discriminative, codewords, ieee, half, directly, unsupervised, higher, chromaticity] [sparse, dictionary, gradient, pursuit, size, spectral, analysis, edge, step, sparsity] [power, coding, surface, natural] [code, learning, linear, data, orthogonal, dataset, svm, large, better, table, well, datasets]

Deep Learning of Invariant Features via Simulated Fixations in Video
Will Zou, Andrew Ng, Shenghuo Zhu, Kai Yu

[pdf] [bibtex] [supplementary]
[rank by tf-idf similarity to this]
[abstract]

http://cs.stanford.edu/people/karpathy/nipspreview/

GAUSSIAN DENSITY FUNCTION

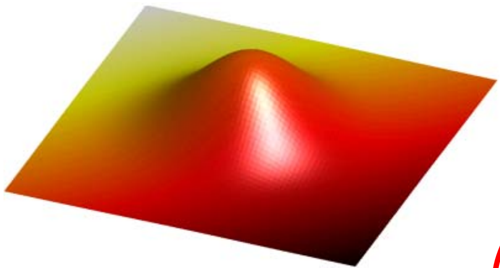
Gaussian Random Variable

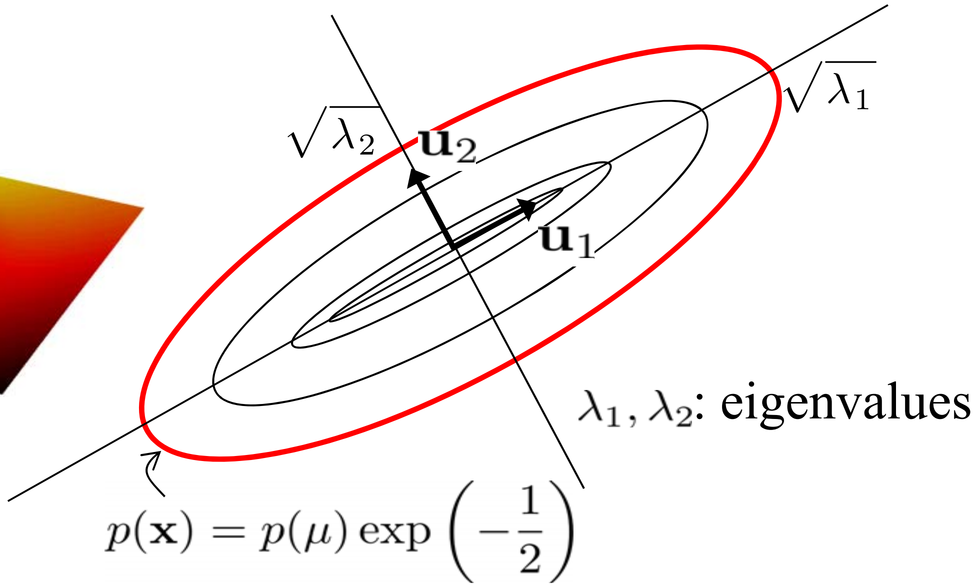
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

Principal axes are the eigenvector directions of Σ

$$\Sigma\,\mathbf{u}_i = \lambda_i \mathbf{u}_i$$





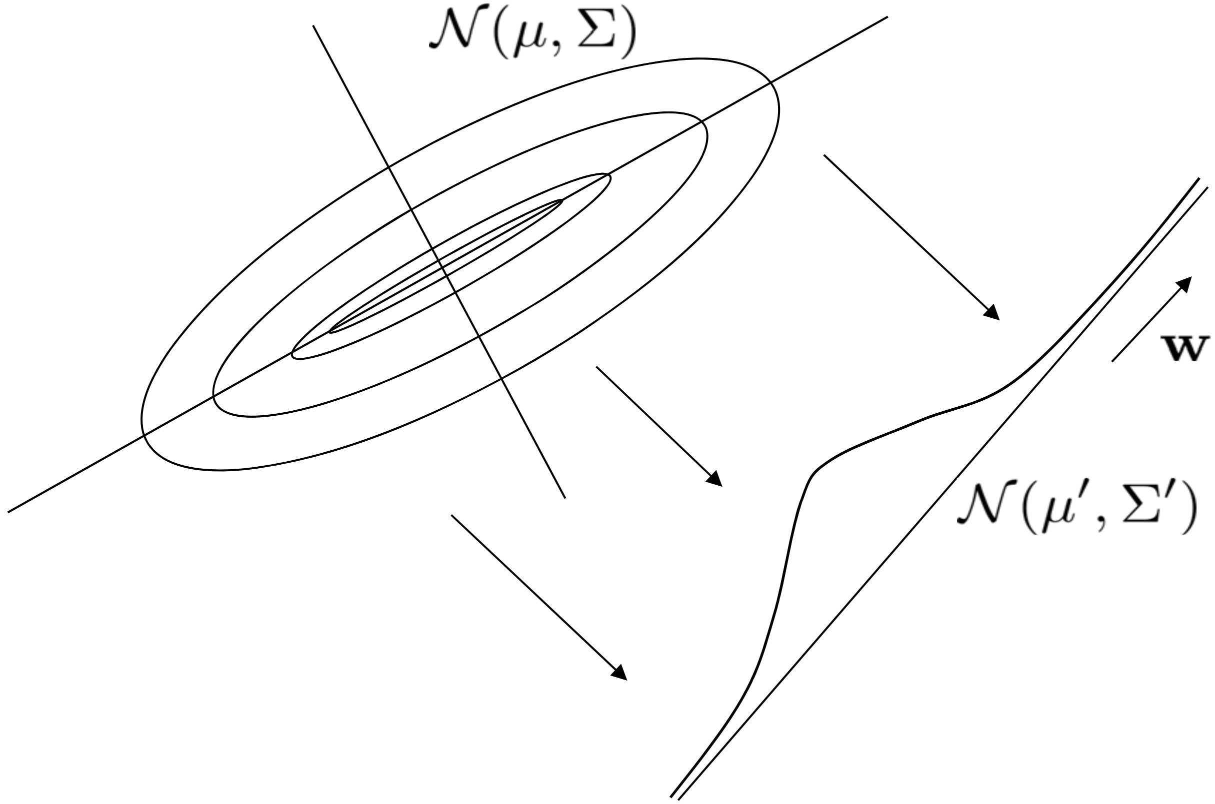
Gaussian Random Variable - Projection

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

Projection to any direction is Gaussian.

$$\mu' = \mathbf{w}^\top \mu$$

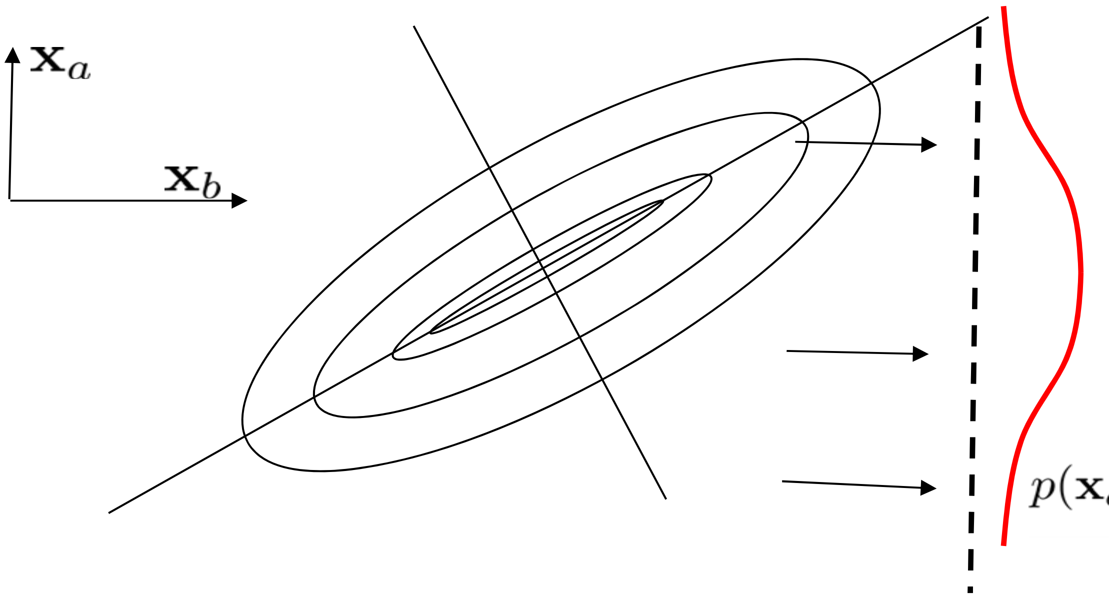
$$\Sigma' = \mathbf{w}^\top \Sigma \mathbf{w}$$



Gaussian Random Variable – Marginal

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

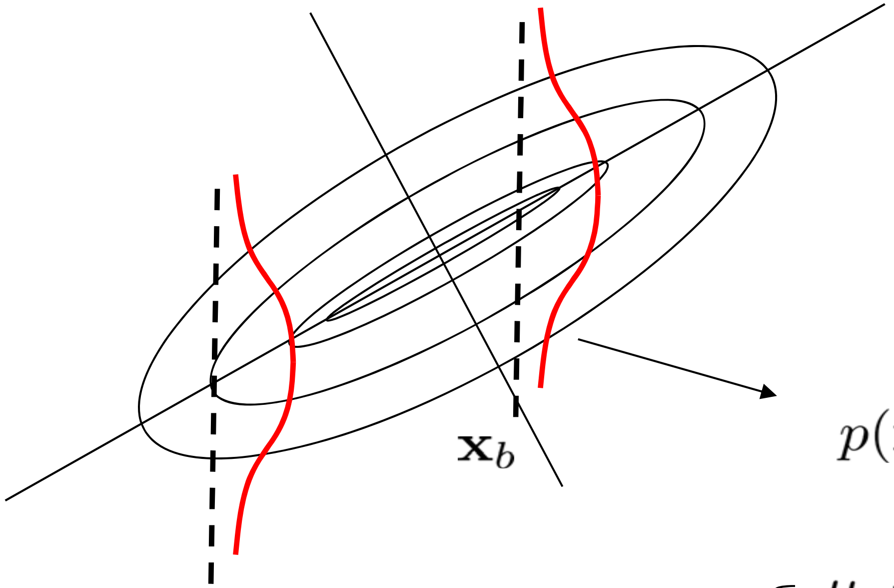
$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \quad \begin{matrix} \mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b} \end{matrix} \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$



$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$
$$= \mathcal{N}(\mu_a, \Sigma_a)$$

Gaussian Random Variable – Conditional

$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right)$$



$$\mathbf{x} = \begin{pmatrix}\mathbf{x}_a \\ \mathbf{x}_b\end{pmatrix} \qquad \begin{matrix}\mathbf{x}_a \in \mathbb{R}^{D_a} \\ \mathbf{x}_b \in \mathbb{R}^{D_b}\end{matrix}$$

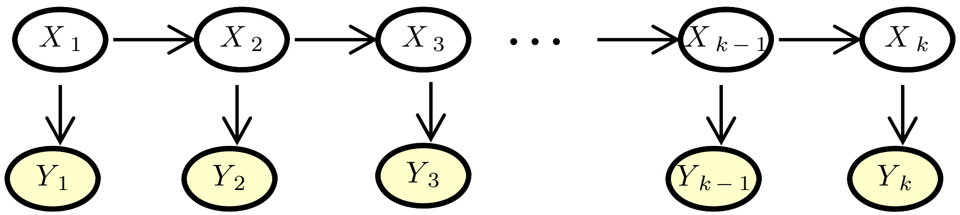
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$$

$$\begin{cases} \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(\mathbf{x}_b - \mu_b) \\ \Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba} \end{cases}$$

KALMAN FILTER

Filtering

- Hidden Markov Models (HMM) /Linear Dynamical Systems (LDS)



$$p(y_1 \ldots, y_K, x_1, \ldots, x_K) = p(x_1)p(y_1|x_1) \prod_{t=1}^{K-1} p(x_{t+1}|x_t)p(y_t|x_t)$$

HMM

$$p(x_1 = j) = \pi_j$$

$$p(x_{t+1}|x_t) = T_{ij}$$

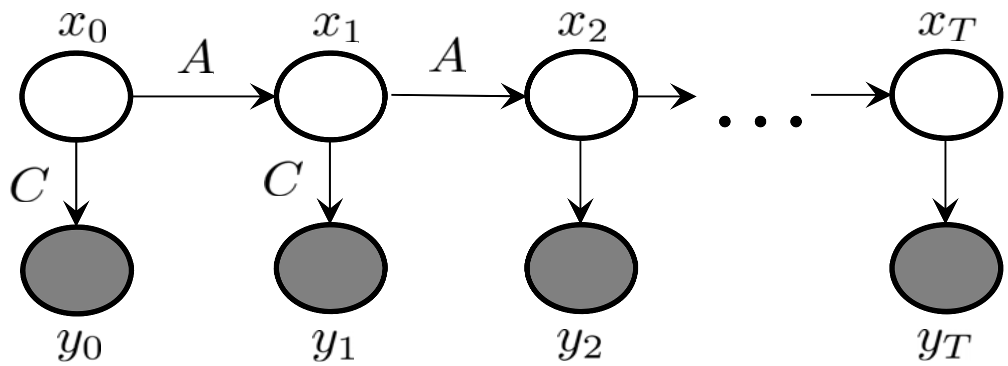
$$p(y_t|x_t) = A_j(y)$$

LDS

$$x_{t+1} = Ax_t + Gw_t \quad w_t \sim \mathcal{N}(0, Q)$$

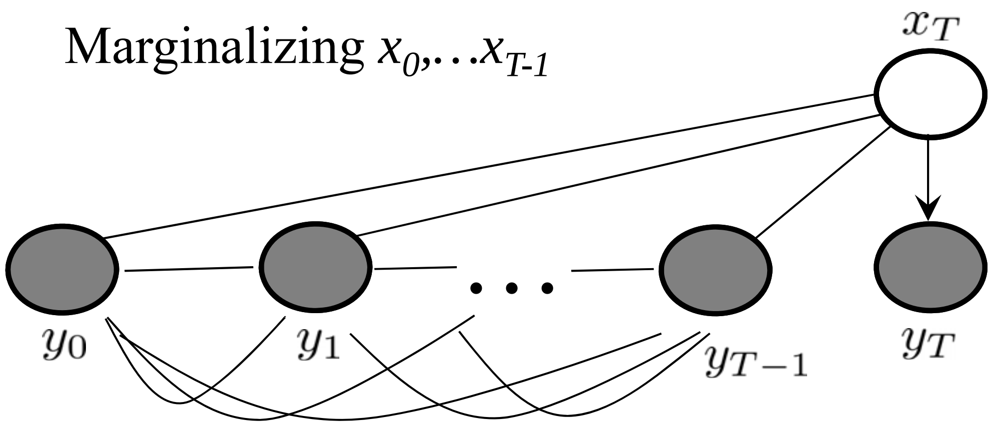
$$y_t = Cx_t + v_t \qquad v_t \sim \mathcal{N}(0, R)$$

Kalman Filter



$$x_{t+1} = Ax_t + Gw_t$$
$$w_t \sim \mathcal{N}(0, Q)$$
$$y_t = Cx_t + v_t$$
$$v_t \sim \mathcal{N}(0, R)$$

Marginalizing $x_0, \ldots x_{T-1}$



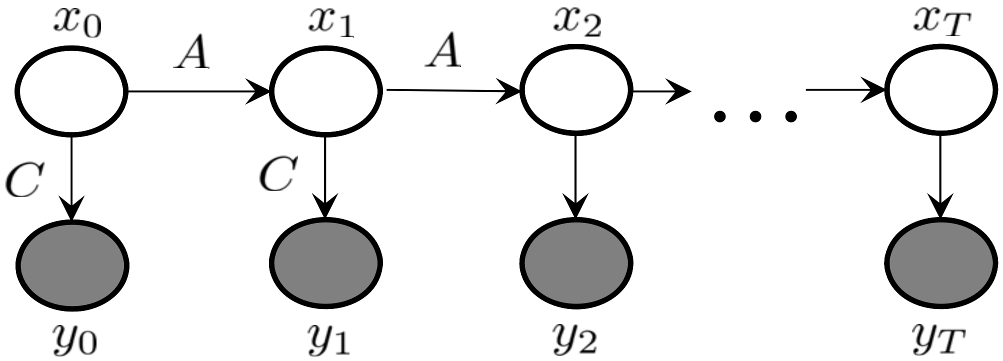
Filtering

$$\hat{x}_{T|T} = \mathbb{E}[x_T | y_0, \ldots, y_T]$$

$$P_{T|T} = \mathbb{E}\Big[(x_T - \hat{x}_{T|T})(x_T - \hat{x}_{T|T})^\top \Big| y_0, \ldots, y_T\Big]$$

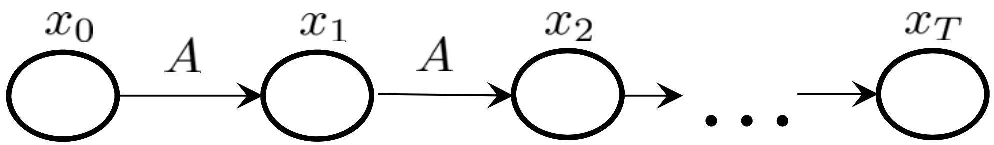
"Conditional marginalization"
Marginalization from the left

Kalman Filter



$$
\begin{aligned}
x_{t+1} &= Ax_t + Gw_t \\
& \quad w_t \sim \mathcal{N}(0, Q) \\
y_t &= Cx_t + v_t \\
& \quad v_t \sim \mathcal{N}(0, R)
\end{aligned}
$$

Unconstrained distribution



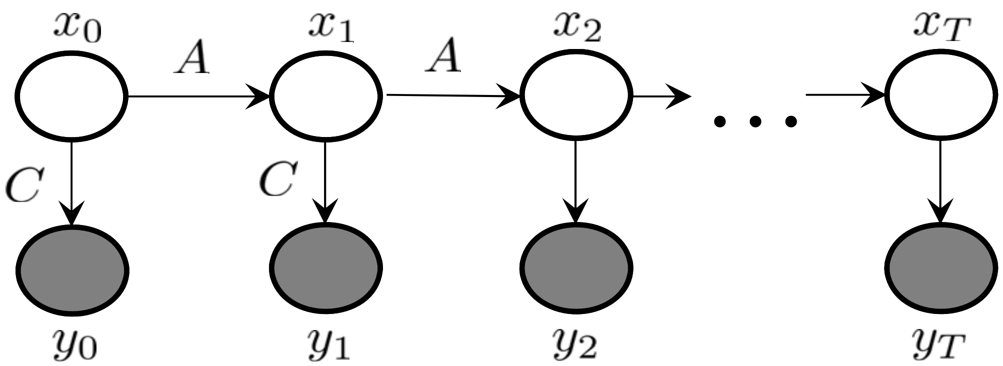
$$\mu_{t+1} = 0$$

$$
\begin{aligned}
\Sigma_{t+1} &= \mathbb{E}[x_{t+1}x_{t+1}^\top] = \mathbb{E}[(Ax_t + Gw_t)(Ax_t + Gw_t)^\top] \\
&= A\mathbb{E}[x_t x_t^\top]A^\top + G\mathbb{E}[w_t w_t^\top]G^\top \\
&= A\Sigma_t A^\top + GQG^\top
\end{aligned}
$$

Also, for joint density if necessary

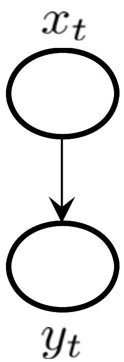
$$\mathbb{E}[x_t x_{t+1}^\top] = \Sigma_t A^\top$$

Kalman Filter



$$x_{t+1} = Ax_t + Gw_t$$
$$w_t \sim \mathcal{N}(0, Q)$$
$$y_t = Cx_t + v_t$$
$$v_t \sim \mathcal{N}(0, R)$$

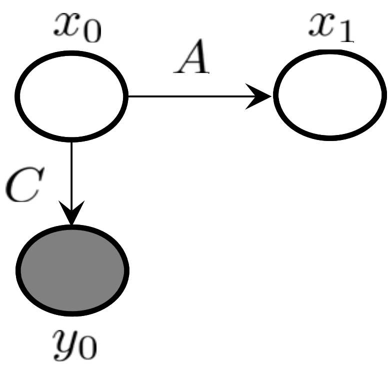
Unconstrained distribution



$$\mu_{y_t} = 0$$

$$\begin{aligned}\Sigma_{y_t} &= \mathbb{E}[y_t y_t^\top] \qquad y_t = Cx_t + v_t \\ &= \mathbb{E}[(Cx_t + v_t)(Cx_t + v_t)^\top] \\ &= C\mathbb{E}[x_t x_t^\top]C^\top + \mathbb{E}[v_t v_t^\top]^\top \\ &= C\Sigma_t C^\top + R\end{aligned}$$

Kalman Filter



$$\begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_{y_0} \end{pmatrix}$$

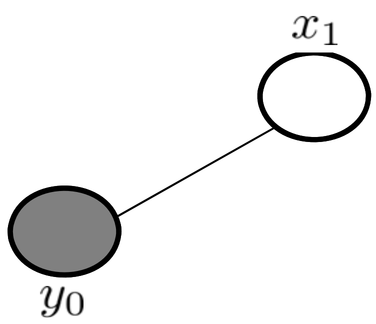
$$\begin{pmatrix} \Sigma_0 & \Sigma_{01} & \Sigma_{0y_0} \\ \Sigma_{10} & \Sigma_1 & \Sigma_{1y_0} \\ \Sigma_{y_0 0} & \Sigma_{y_0 1} & \Sigma_{y_0} \end{pmatrix}$$

Constrained distribution

$$\mu_{1|0} = \mu_{x_1|y_0} = \mu_1 + \Sigma_{1y_0}\Sigma_{y_0}^{-1}(y_0 - \mu_{y_0})$$

$$\Sigma_{1|0} = \Sigma_{x_1|y_0} = \Sigma_1 - \Sigma_{1y_0}\Sigma_{y_0}^{-1}\Sigma_{y_0 1}$$

Marginalizing x_0



$$\begin{pmatrix} \mu_1 \\ \mu_{y_0} \end{pmatrix}$$

$$\begin{pmatrix} \Sigma_1 & \Sigma_{1y_0} \\ \Sigma_{y_0 1} & \Sigma_{y_0} \end{pmatrix}$$

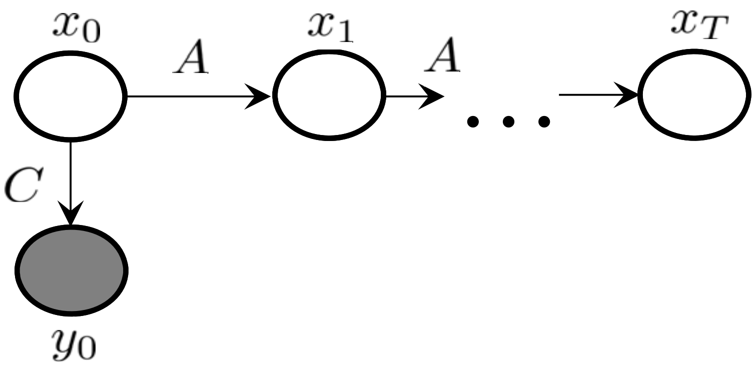
Same

$$\mu_{1|0} = \mu_{x_1|y_0} = \mu_1 + \Sigma_{1y_0}\Sigma_{y_0}^{-1}(y_0 - \mu_{y_0})$$

$$\Sigma_{1|0} = \Sigma_{x_1|y_0} = \Sigma_1 - \Sigma_{1y_0}\Sigma_{y_0}^{-1}\Sigma_{y_0 1}$$

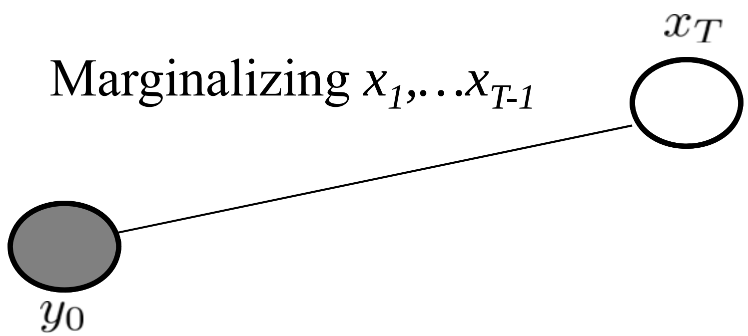
Σ_{y_0} and Σ_1 are from unconstrained distribution. What matters is Σ_{1y_0} .

Kalman Filter



$$
\begin{pmatrix} \mu_0 \\ \mu_1 \\ \vdots \\ \mu_{y_0} \end{pmatrix}
\begin{pmatrix} \Sigma_0 & \Sigma_{01} & \dots & \Sigma_{0T} & \Sigma_{0y_0} \\ \Sigma_{10} & \Sigma_1 & \dots & \Sigma_{1T} & \Sigma_{1y_0} \\ & \dots & & \dots & \\ \Sigma_{y_0 0} & \Sigma_{y_0 1} & \dots & \Sigma_{y_0 T} & \Sigma_{y_0} \end{pmatrix}
$$

Marginalizing $x_1,\ldots x_{T\text{-}1}$



$$
\begin{pmatrix} \mu_T \\ \mu_{y_0} \end{pmatrix}
$$

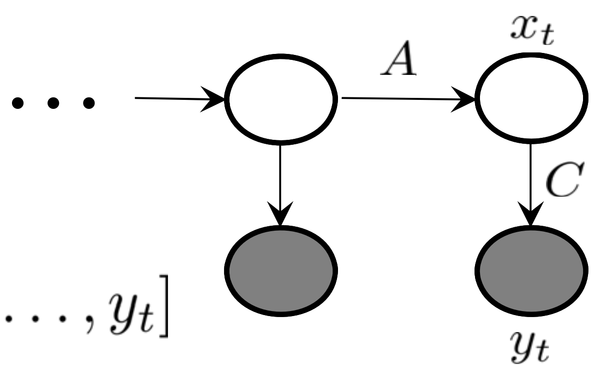
$$
\begin{pmatrix} \Sigma_T & \Sigma_{Ty_0} \\ \Sigma_{y_0 T} & \Sigma_{y_0} \end{pmatrix}
$$

$$
\mu_{T|0} = \mu_T + \Sigma_{Ty_0}\Sigma_{y_0}^{-1}(y_0 - \mu_{y_0})
$$

$$
\Sigma_{T|0} = \Sigma_T - \Sigma_{Ty_0}\Sigma_{y_0}^{-1}\Sigma_{y_0 T}
$$

Kalman Filter

Filtering



$$\widehat{x}_{t|t} = \mathbb{E}[x_t | y_0, \ldots, y_t]$$

$$P_{t|t} = \mathbb{E}[(x_t - \widehat{x}_{t|t})(x_t - \widehat{x}_{t|t})^\top | y_0, \ldots, y_t]$$

"Conditional marginalization"
Marginalization from the left

$$\widehat{x}_{t|t} \ \& \ P_{t|t} \ \longrightarrow \ \widehat{x}_{t+1|t+1} \ \& \ P_{t+1|t+1}$$

Why filtering? Once we know $\hat{x}_{t|t}$ & $P_{t|t}$, we don't have to know (or keep) $y_0, \ldots, y_t$.

Kalman Filter

- Time update

$$p(x_t|y_0,\ldots,y_t) \rightarrow p(x_{t+1}|y_0,\ldots,y_t)$$

- Measurement update

$$p(x_{t+1}|y_0,\ldots,y_t) \rightarrow p(x_{t+1}|y_0,\ldots,y_t,y_{t+1})$$

Time update

$$\begin{aligned}
\widehat{x}_{t+1|t} &= A\widehat{x}_{t|t} \\
P_{t+1|t} &= \mathbb{E}[(x_{t+1}-\widehat{x}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t] \\
&= \mathbb{E}[(Ax_t+Gw_t-A\widehat{x}_{t|t})(Ax_t+Gw_t-A\widehat{x}_{t|t})^\top|y_0,\ldots,y_t] \\
&= AP_{t|t}A^\top + GQG^\top
\end{aligned}$$

Kalman Filter

$$\begin{aligned}\mathbb{E}[y_{t+1}|y_0,\ldots,y_t] &= \mathbb{E}[Cx_{t+1}+v_{t+1}|y_0,\ldots,y_t]\\ &= C\widehat{x}_{t+1|t}\end{aligned}$$

$$\begin{aligned}&\mathbb{E}[(y_{t+1}-\widehat{y}_{t+1|t})(y_{t+1}-\widehat{y}_{t+1|t})^\top|y_0,\ldots,y_t]\\ &\quad=\mathbb{E}[(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]\\ &\quad=CP_{t+1|t}C^T+R\end{aligned}$$

Also,

$$\begin{aligned}&\mathbb{E}[(y_{t+1}-\widehat{y}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]\\ &\quad=\mathbb{E}[(Cx_{t+1}+v_{t+1}-C\widehat{x}_{t+1|t})(x_{t+1}-\widehat{x}_{t+1|t})^\top|y_0,\ldots,y_t]\\ &\quad=CP_{t+1|t}\end{aligned}$$

Joint:

$$\begin{aligned}&p(x_{t+1},y_{t+1}|y_0,\ldots,y_t)\\ &\qquad=\mathcal{N}\left(\begin{pmatrix}\widehat{x}_{t+1|t}\\ C\widehat{x}_{t+1|t}\end{pmatrix},\begin{pmatrix}P_{t+1|t} & P_{t+1|t}C^\top\\ CP_{t+1|t} & CP_{t+1|t}C^\top+R\end{pmatrix}\right)\end{aligned}$$

Kalman Filter

Measurement update (Conditional density)

$$p(x_{t+1}|y_0,\ldots,y_{t+1}) = \mathcal{N}(\widehat{x}_{t+1|t+1}, P_{t+1|t+1})$$

$$\begin{cases} \widehat{x}_{t+1|t+1} = \widehat{x}_{t+1|t} + P_{t+1|t}C^\top (CP_{t+1|t}C^\top + R)^{-1}(y_{t+1} - C\widehat{x}_{t+1|t}) \\ P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^\top (CP_{t+1|t}C^\top + R)^{-1} CP_{t+1|t} \end{cases}$$

- Sum - ups

$$\widehat{x}_{t+1|t} = A\widehat{x}_{t|t}$$

$$P_{t+1|t} = AP_{t|t}A^\top + GQG^\top$$

$$\widehat{x}_{t+1|t+1} = \widehat{x}_{t+1|t} + P_{t+1|t}C^\top (CP_{t+1|t}C^\top + R)^{-1}(y_{t+1} - C\widehat{x}_{t+1|t})$$

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C^\top (CP_{t+1|t}C^\top + R)^{-1} CP_{t+1|t}$$

Kalman Filter

- With different notation,

$$K_{t+1} \equiv P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1}$$

$$\widehat{x}_{t+1|t+1} = \widehat{x}_{t+1|t} + K_{t+1}(y_{t+1} - C\widehat{x}_{t+1|t})$$

- Alternative form of K_{t+1}

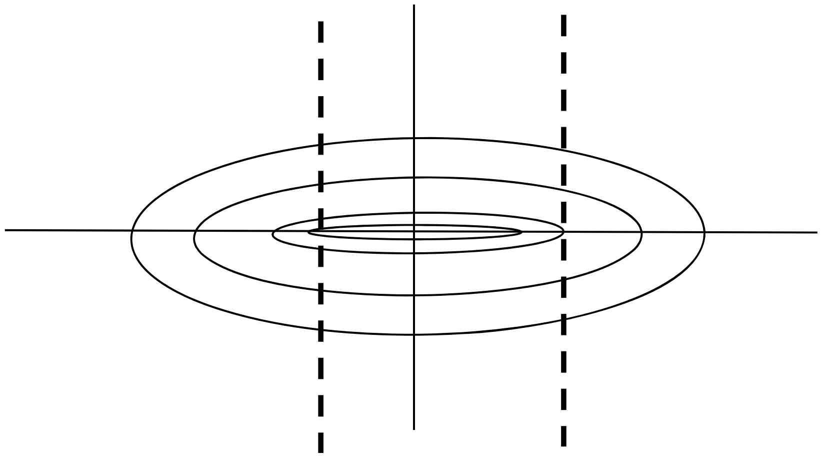
$$\begin{aligned} K_{t+1} &= P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1} \\ &= (P_{t+1|t}^{-1} + C^\top R C)^{-1} C^\top R^{-1} \\ &= (P_{t+1|t} + P_{t+1|t} C^\top (C P_{t+1|t} C^\top + R)^{-1} C P_{t+1|t}) C^\top R^{-1} \\ &= P_{t+1|t+1} C^\top R^{-1} \end{aligned}$$

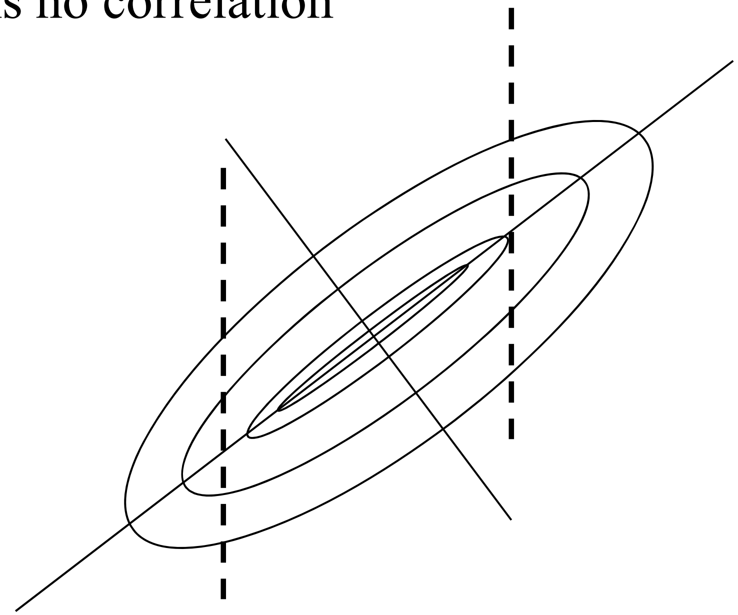
Independency

- Correlation and Independency

$$p(\mathbf{x}_a, \mathbf{x}_b) = p(\mathbf{x}_a)p(\mathbf{x}_b)$$

Independency in Gaussian means no correlation





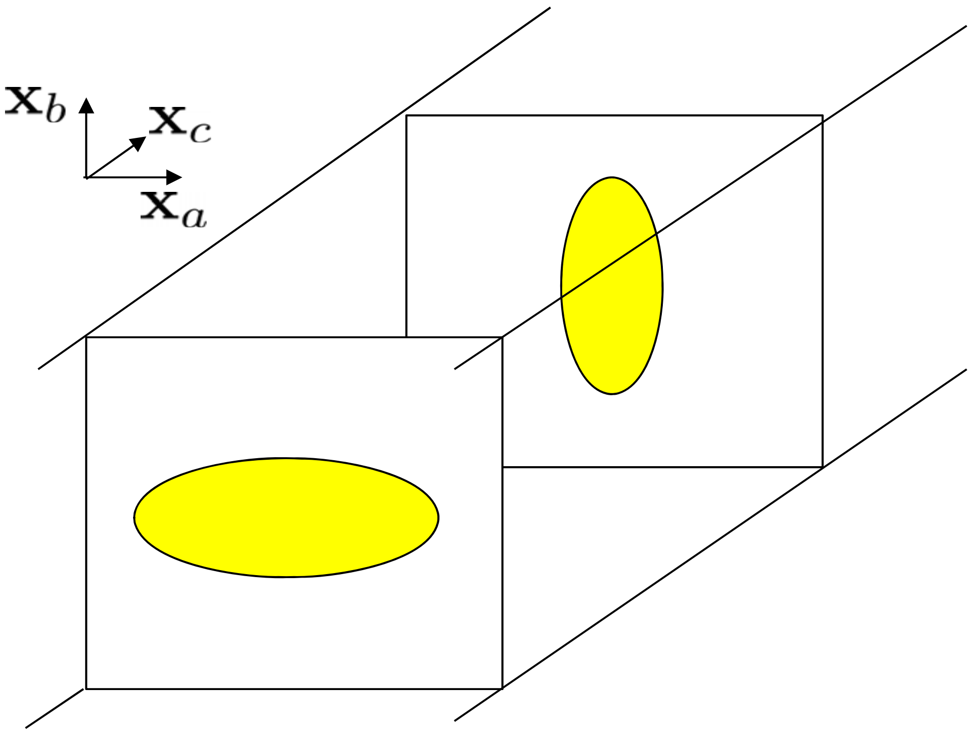
Naïve Bayes?
Mixture of Gaussian?

Conditional Independency

$$p(\mathbf{x}_a, \mathbf{x}_b) = p(\mathbf{x}_a)p(\mathbf{x}_b)$$

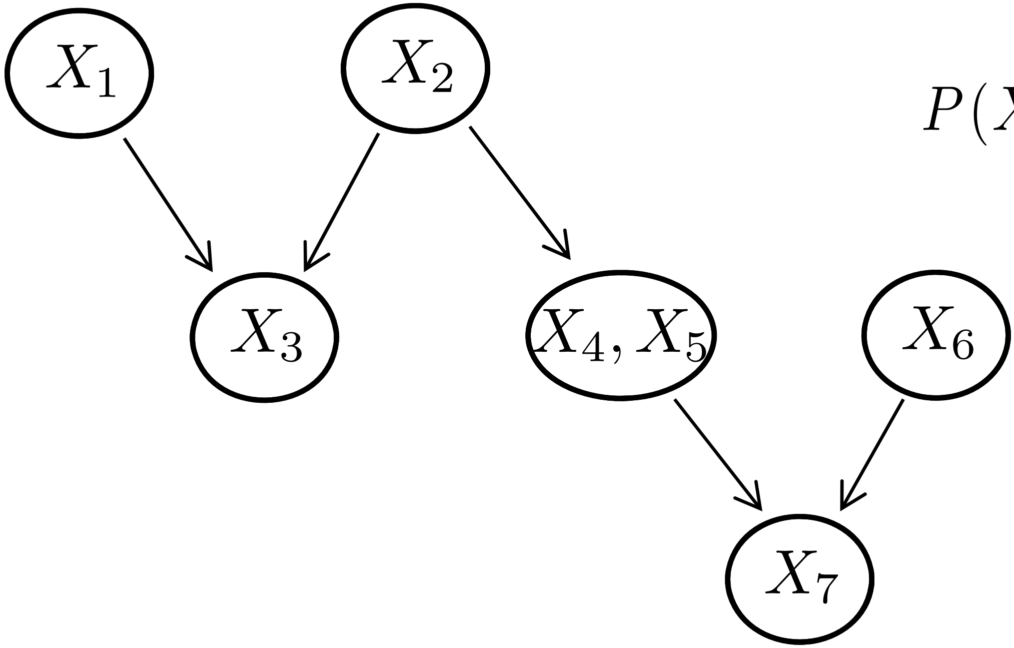
vs.

$$p(\mathbf{x}_a, \mathbf{x}_b|\mathbf{x}_c) = p(\mathbf{x}_a|\mathbf{x}_c)p(\mathbf{x}_b|\mathbf{x}_c)$$



Directed Graphical Models

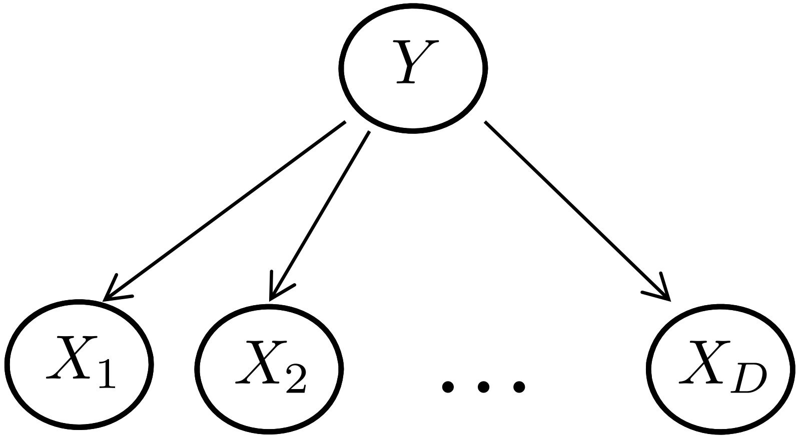
- Factorization of a (large) joint pdf

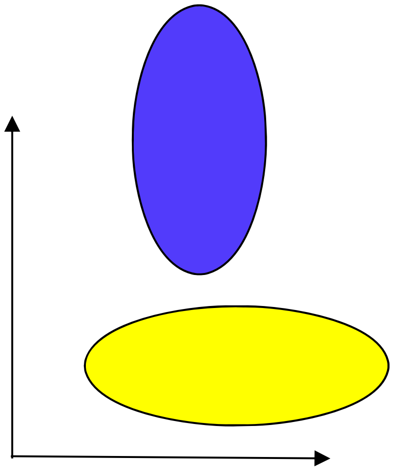


$$
\begin{aligned}
P(X) &= P(X_1, \ldots, X_7) \\
&= P(X_1)P(X_2)P(X_3|X_1, X_2) \\
&\quad P(X_4, X_5|X_2)P(X_6) \\
&\quad P(X_7|X_4, X_5, X_6)
\end{aligned}
$$

- For given data, make a model for each decomposed probability, then estimate parameters separately.

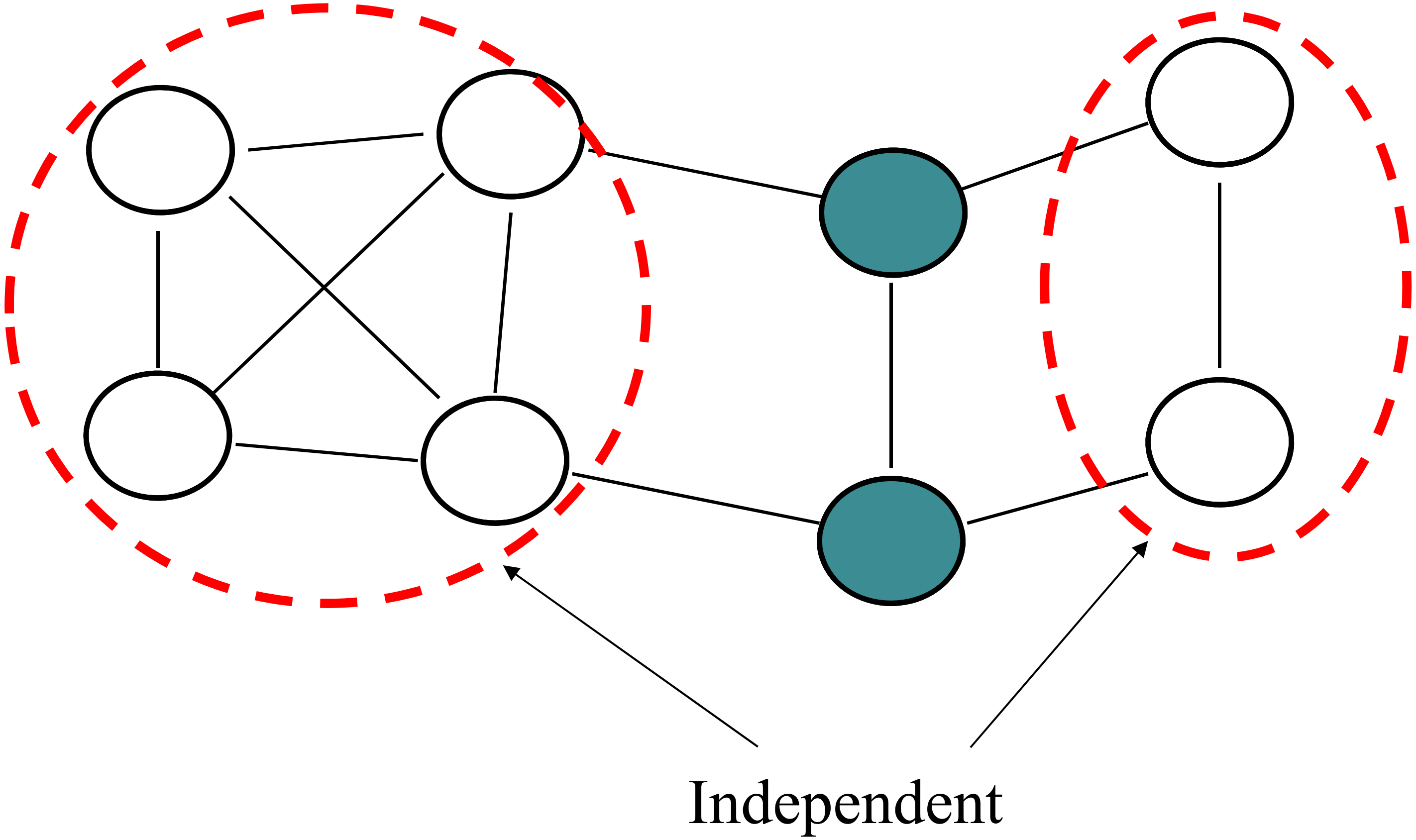
Naïve Bayes for Classification





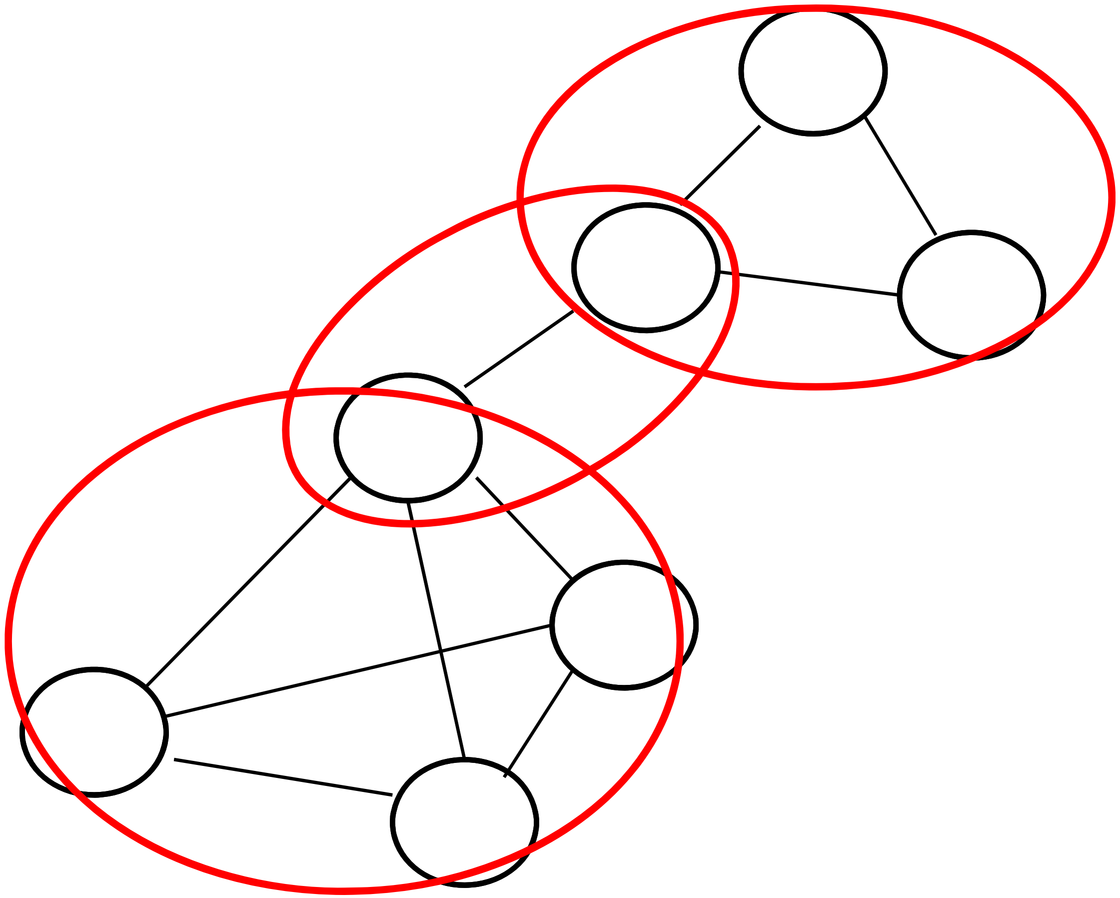
$$P(X, Y) = P(Y)P(X_1|Y)\ldots P(X_D|Y)$$

Undirected Graphical Models



Undirected Graphical Models

- Find all maximal cliques:



Undirected Graphical Models

- Potential functions on cliques

$$\Psi_1(X_1), \Psi_2(X_2), \ldots \qquad (X_1, X_2, \ldots\text{: maximal cliques})$$

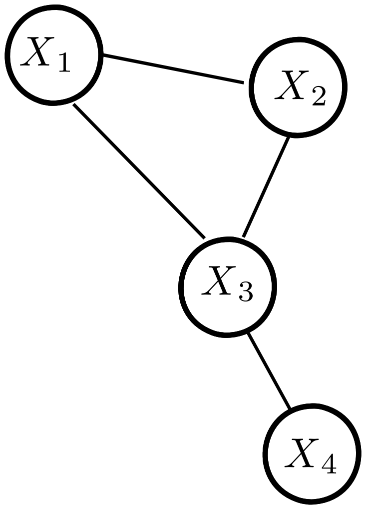
$$P(X) = \frac{1}{Z}\Psi_1(X_1)\Psi_2(X_2)\cdots\Psi_C(X_C)$$

$$\begin{cases} Z = \displaystyle\sum_{X_1, X_2, \cdots, X_D} \Psi_1(X_1)\cdots\Psi_C(X_C) & \textit{Discrete} \\ Z = \displaystyle\int_{X_1, X_2, \cdots, X_D} \Psi_1(X_1)\cdots\Psi_C(X_C) dX_1 \cdots X_C & \textit{Continuous} \end{cases}$$

Partition function

Undirected Graphical Models

2-cliques



X_1	X_2	X_3	Ψ_{X_1,X_2,X_3}
1	1	1	2
1	1	0	0
1	0	1	0
1	0	0	1
0	1	1	2
0	1	0	0
0	0	1	0
0	0	0	1

X_3	X_4	Φ_{X_3,X_4}
1	1	1
1	0	0
0	1	0
0	0	3

$$Z = \sum_{X_1,X_2,X_3,X_4} \Psi_{X_1,X_2,X_3}(X_1,X_2,X_3)\Phi_{X_3,X_4}(X_3,X_4)$$

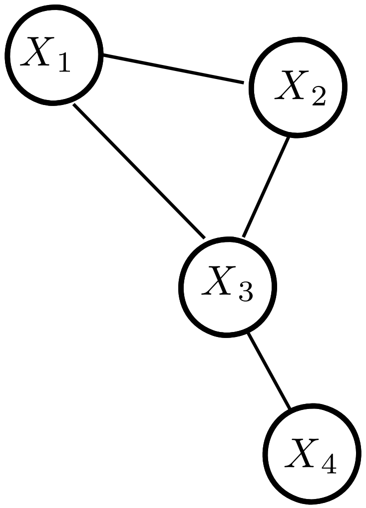
$$= \Psi_{X_1,X_2,X_3}(1,1,1)\Phi_{X_3,X_4}(1,1) + \Psi_{X_1,X_2,X_3}(1,1,1)\Phi_{X_3,X_4}(1,0) + \ldots$$

$$= 2\cdot 1 + 2\cdot 0 + \ldots = 2+3+2+3 = 10$$

Ex. $P(1,0,0,0) = \frac{1}{Z}\Psi_{X_1,X_2,X_3}(1,0,0)\Phi_{X_3,X_4}(0,0) = \frac{1}{10}\cdot 1\cdot 3 = \frac{3}{10}$

Estimating Parameters

2-cliques



X_1	X_2	X_3	
1	1	1	$= \Psi_{1,1,1}$
1	1	0	$= \Psi_{1,1,0}$
1	0	1	$\vdots$
1	0	0	
0	1	1	
0	1	0	
0	0	1	
0	0	0	

X_3	X_4	
1	1	$= \Phi_{1,1}$
1	0	$= \Phi_{1,0}$
0	1	$\vdots$
0	0	

12 parameters

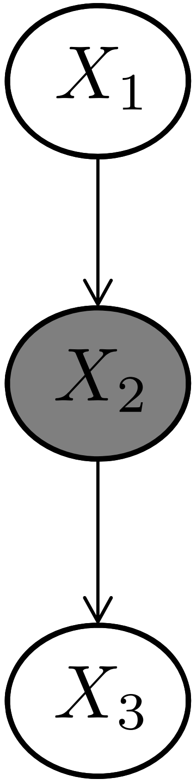
Ψ_{X_1,X_2,X_3} : *8*

Φ_{X_3,X_4} : *4*

Without graphical model: *15 parameters* (*$2^4 - 1$*)

X_1	X_2	X_3	X_4	
1	1	1	1	$= P(1,1,1,1)$
1	1	1	0	$= P(1,1,1,0)$
1	1	0	1	$\vdots$
$\cdots$				

Conditional Independency



$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2)$$

$$P(X_3 = 1|X_1 = 0, X_2 = 1) = ?$$

$$P(X_3 = 1|X_1 = 0, X_2 = 1)$$

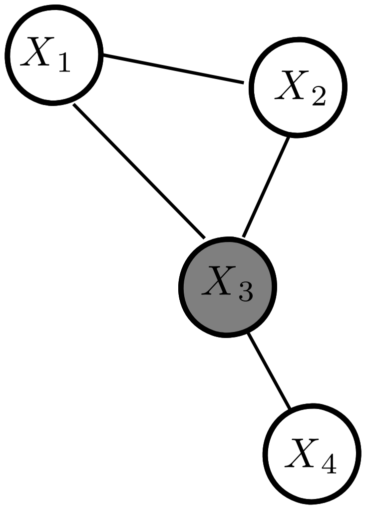
$$= \frac{P(X_1 = 0, X_2 = 1, X_3 = 1)}{P(X_1 = 0, X_2 = 1, X_3 = 1) + P(X_1 = 0, X_2 = 1, X_3 = 0)}$$

$$= \frac{P(X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_3 = 1|X_2 = 1)}{P(X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_3 = 1|X_2 = 1) + P(X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_3 = 0|X_2 = 1)}$$

$$= \frac{P(X_3 = 1|X_2 = 1)}{P(X_3 = 1|X_2 = 1) + P(X_3 = 0|X_2 = 1)} = P(X_3 = 1|X_2 = 1)$$

Conditional Independency

2-cliques



X_1	X_2	X_3	
1	1	1	$=\Psi_{1,1}$
1	0	1	$=\Psi_{1,0}$
0	1	1	$=\Psi_{0,1}$
0	0	1	$=\Psi_{0,0}$

X_3	X_4	
1	1	$=\Phi_1$
1	0	$=\Phi_0$

X_1	X_2	X_3	X_4	
1	1	1	1	$=\Psi_{1,1}\Phi_1$
1	1	1	0	$=\Psi_{1,1}\Phi_0$
1	0	1	1	$=\Psi_{1,0}\Phi_1$
1	0	1	0	$=\Psi_{1,0}\Phi_0$
0	1	1	1	$=\Psi_{0,1}\Phi_1$
0	1	1	0	$=\Psi_{0,1}\Phi_0$
0	0	1	1	$=\Psi_{0,0}\Phi_1$
0	0	1	0	$=\Psi_{0,0}\Phi_0$

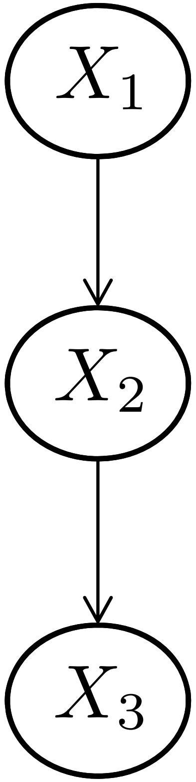
$$P(X_4=1|X_1=0,X_2=1)=?$$

$$P(X_4=1|X_1=0,X_2=0)=?$$

$$P(X_4=1)=?$$

All answers are the same: $\dfrac{\Phi_1}{\Phi_1+\Phi_0}$

Marginalization

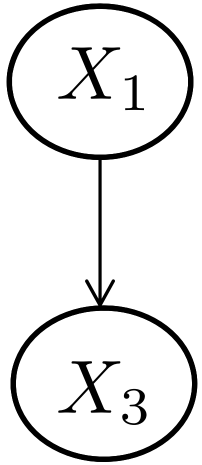


$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2)$$

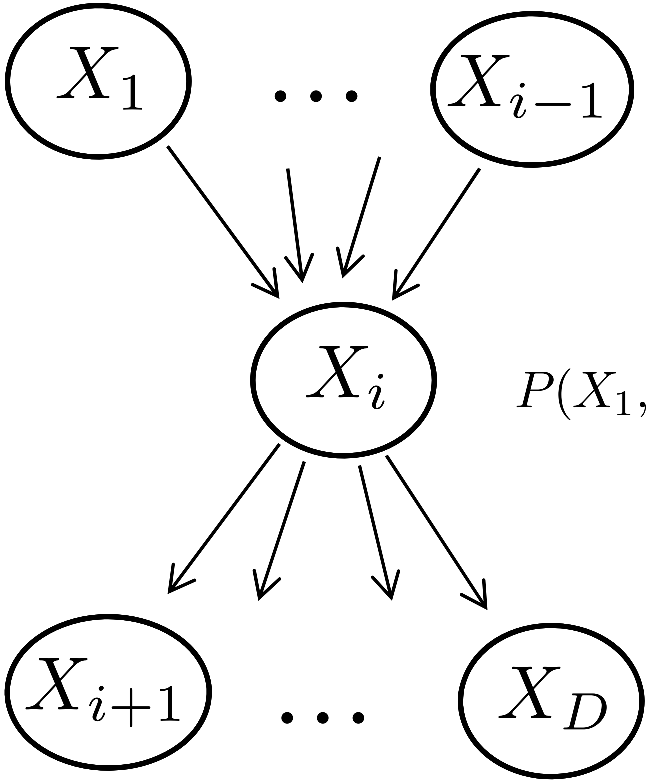
$$P(X_1, X_3) = \int P(X_1, X_2, X_3)dX_2$$

Any good property like

$$P(X_1, X_3) = P(X_1)P(X_3)?$$



Marginalization



$$P(X_1,\ldots,X_D) = P(X_1)\ldots P(X_{i-1})\, P(X_i|X_1,\ldots,X_{i-1})\, P(X_{i+1}|X_i)\ldots P(X_D|X_i)$$

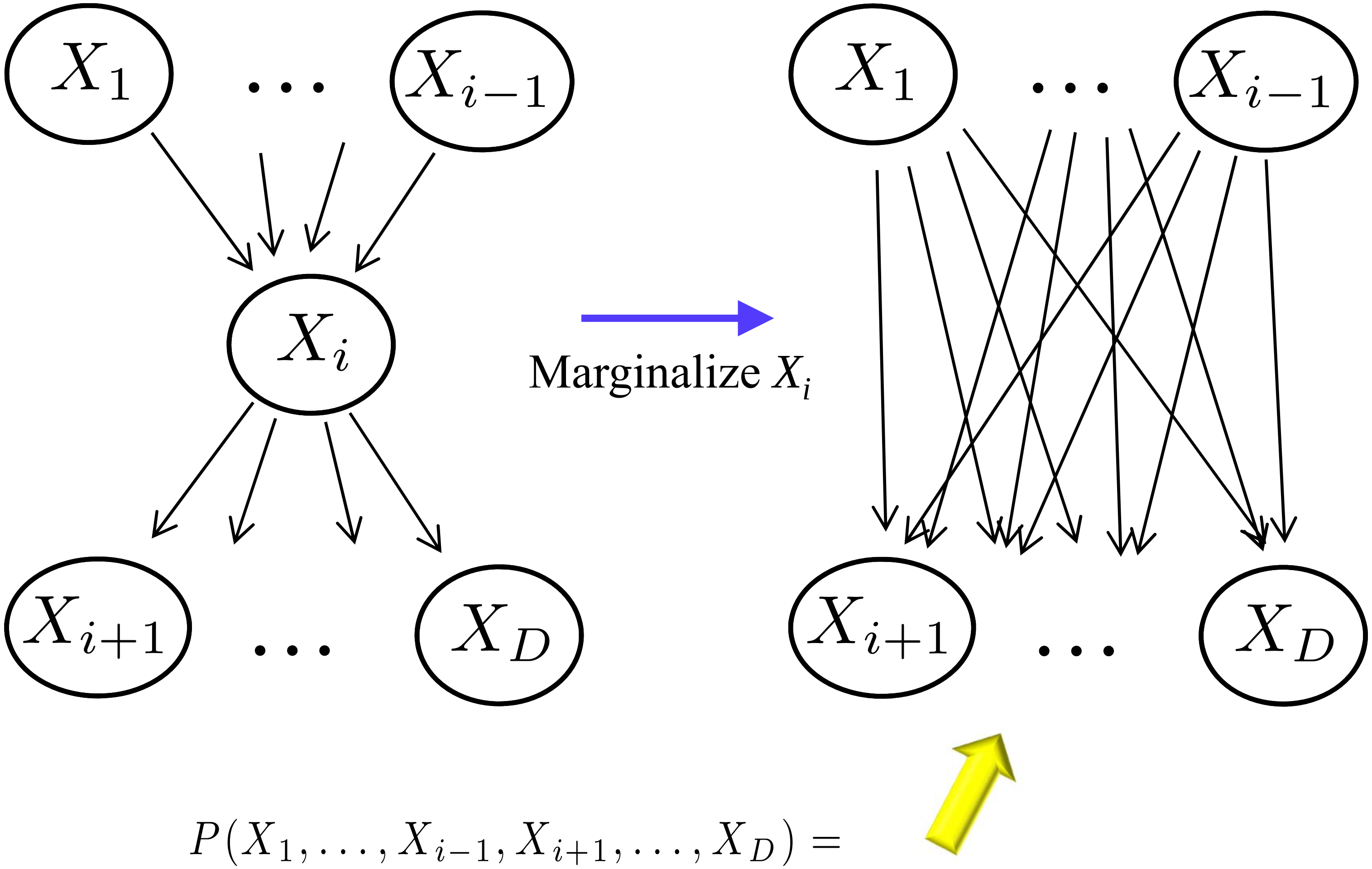
$$P(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_D) = \int P(X_1,\ldots,X_D)dX_i$$

Any decomposition with

$$P(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_D) = ?$$

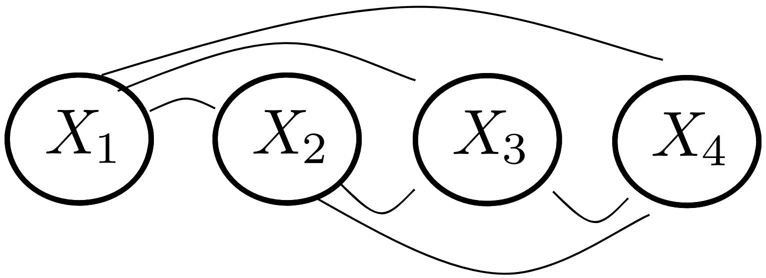
$$P(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_D) = P(X_1)\ldots P(X_{i-1})P(X_{i+1}|X_1,\ldots,X_{i-1})\ldots P(X_D|X_1,\ldots,X_{i-1})$$

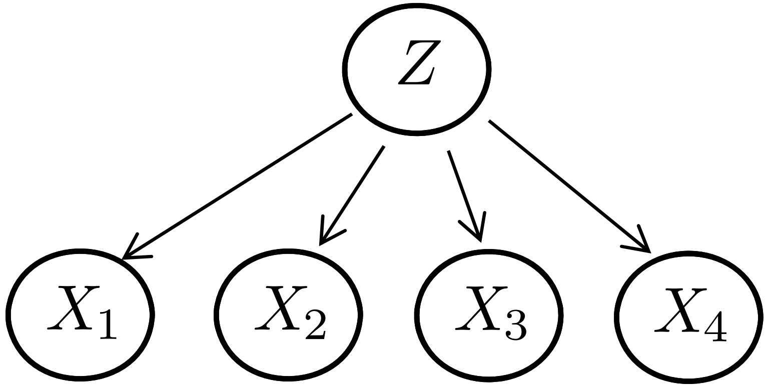
Marginalization



$$P(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_D) =$$
$$P(X_1) \dots P(X_{i-1}) P(X_{i+1} | X_1, \dots, X_{i-1}) \dots P(X_D | X_1, \dots, X_{i-1})$$

Introducing Latent Variables





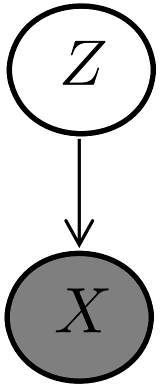
- Issue: how can a model be simplified as much as possible, while the flexibility is kept enough to incorporate the true dependency.

Expectation-Maximization Algorithm

- Parameter estimation with latent variables
  - We don't have data for latent variables
- E-step:
  - Data for latent variables are obtained from expectation with current parameter values.
- M-step:
  - With expected latent variables, parameters are obtained by maximizing the likelihood.
- E-step and M-step are repeated back and forth until the likelihood converges.

Expectation-Maximization Algorithm

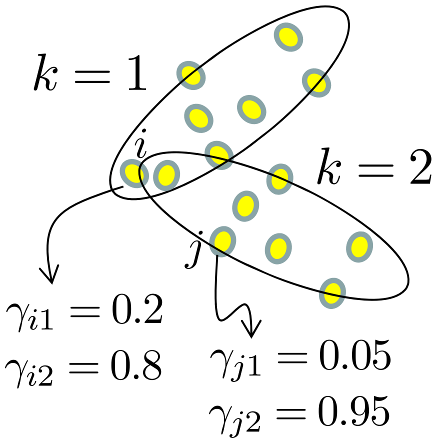
- Gaussian mixture model



Parameters: π_k, μ_k, Σ_k for $k = 1, \ldots, K$

Unknown variables: $z_i = \begin{pmatrix} z_{i1} \\ \vdots \\ z_{iK} \end{pmatrix}$ for $i = 1, \ldots, N$

We are given $\mathbf{x}_i$ for $i = 1, \ldots, N$



E-step: Distribution of $\mathbf{z}_i$ (Responsibilities)

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i | \mu_j, \Sigma_j)}$$

using current parameters π_k, μ_k, Σ_k

Expectation-Maximization Algorithm

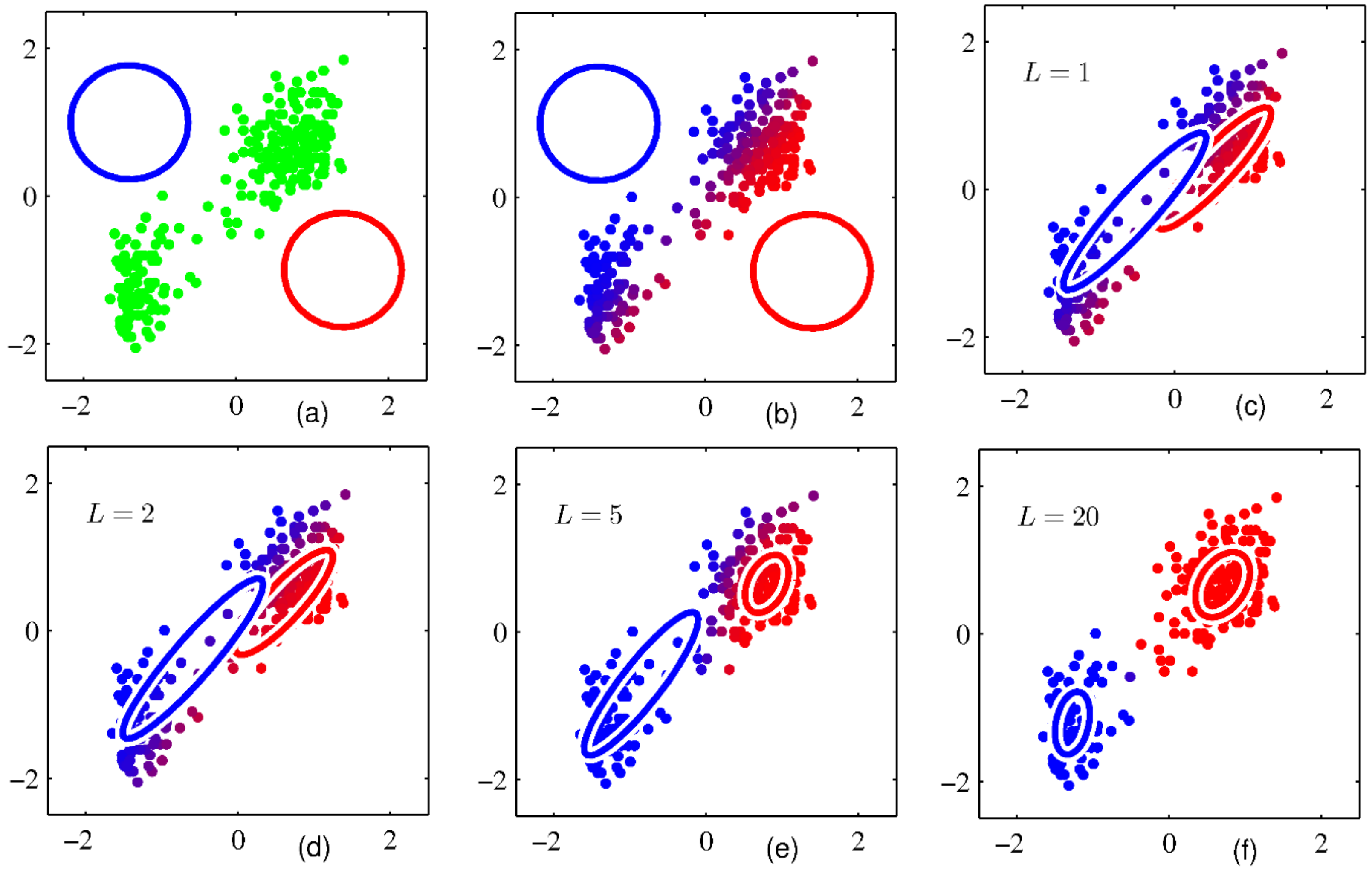
M-step: Estimate parameters.

$$\begin{cases} \mu_k = \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N} \gamma(z_{ik})\ \mathbf{x}_i \\ \Sigma_k = \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N} \gamma(z_{ik})(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top \\ \pi_k = \dfrac{N_k}{N} \end{cases}$$

for $N_k = \sum_{i=1}^{N} \gamma(z_{ik})$

Iterate until $\ln p(X|\pi, \mu, \Sigma) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\mu_k, \Sigma_k) \right)$ converges.

Gaussian Mixture Model With EM



C. Bishop 2007, Figure 9.8(a)-(f)

Sampling and EM Algorithm

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

$$Q(\theta, \theta^{old}) = \int p(Z|X, \theta^{old}) \ln p(Z, X|\theta) dZ$$

cf) $Q(\theta) = \ln p(X|\widehat{Z}, \theta), \quad \widehat{Z} = \int Z \cdot p(Z|X, \theta^{old}) dZ$

- Sampling:

$$Q(\theta, \theta^{old}) \simeq \frac{1}{L} \sum_{l=1}^{L} \ln p(Z^{(l)}, X|\theta) \qquad Z^{(l)} \sim p(Z|X, \theta^{old})$$

Sampling and EM Algorithm (IP Algorithm)

- Imputation-Posterior (IP) algorithm ← more Bayesian
  - I-Step

$$p(Z|X) = \int p(Z|\theta, X) p(\theta|X) d\theta$$

    - $\theta^{(l)}$ are sampled from current estimate for $p(\theta|X)$ then $Z^{(l)}$ are sampled from each $p(Z|\theta^{(l)})$

  - P-step

$$p(\theta|X) = \int p(\theta|Z, X) p(Z|X) dZ$$
$$\simeq \frac{1}{L} \sum_{l=1}^{L} p(\theta|Z^{(l)}, X)$$

ANY QUESTIONS?

THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr