Introduction to Graphical Models

The 15th Winter School of Statistical Physics
POSCO International Center & POSTECH, Pohang
2018. 1. 8-12 (Mon.-Fri.)

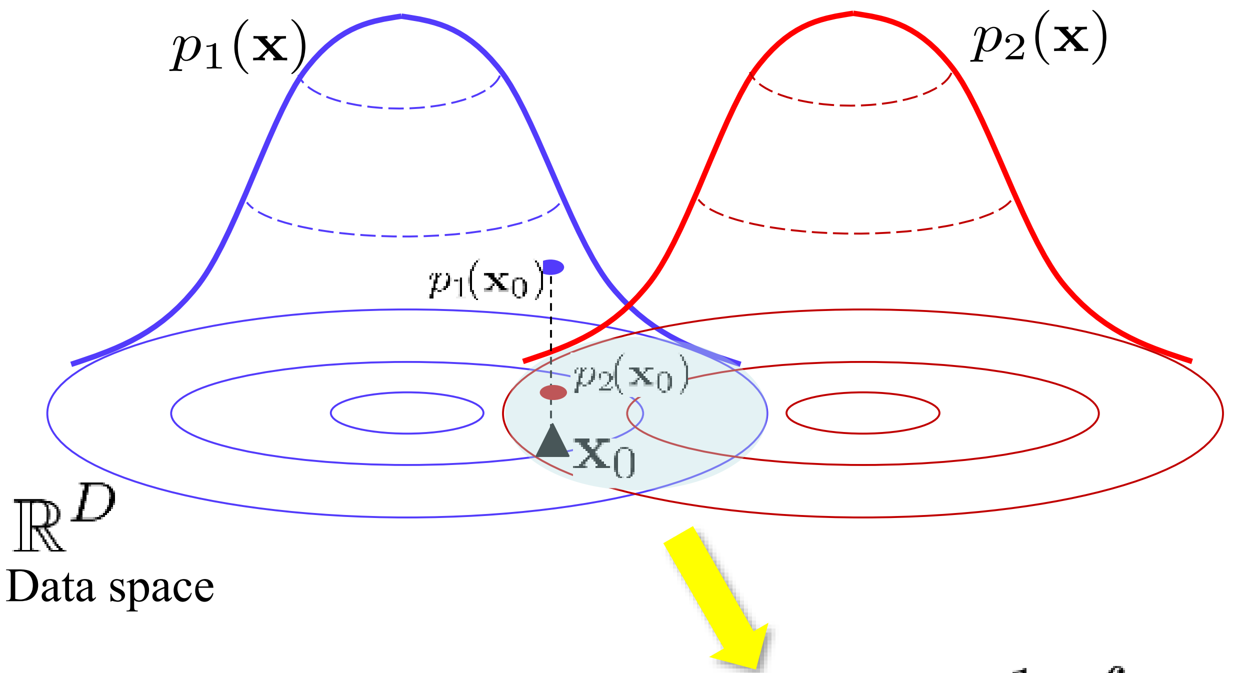
Yung-Kyun Noh
Seoul National University

https://github.com/nohyung/SPWS2018

GENERALIZATION FOR PREDICTION

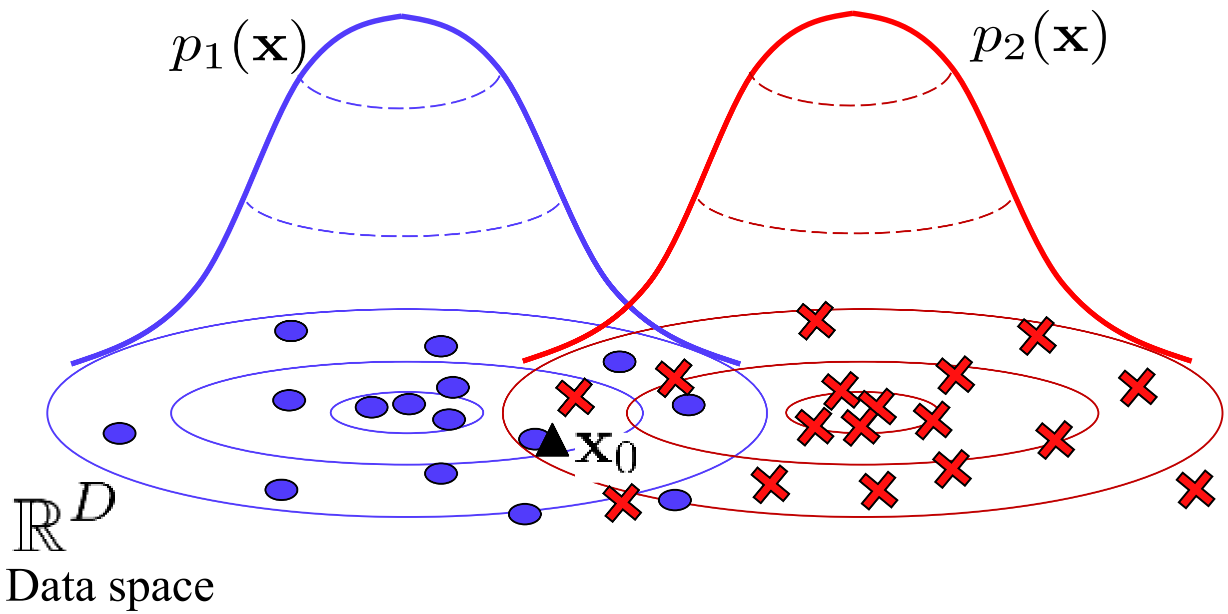
Probabilistic Assumption and Bayes Classification

- Bayes Error



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

Probabilistic Assumption and Bayes Classification



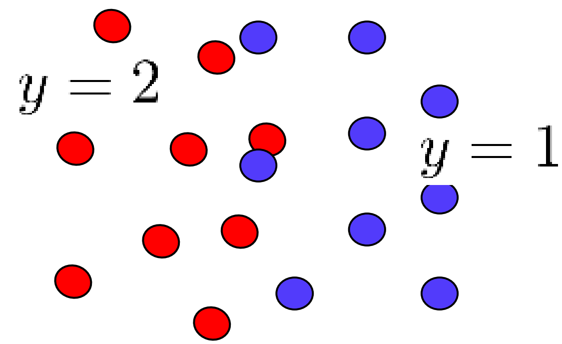
$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim p_1(\mathbf{x}), p_2(\mathbf{x})$$

Learning

- Data

(Regularity)

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim P$$



- Prediction

$$\mathbf{x} \in \mathbb{R}^D \xrightarrow{y = f(\mathbf{x})} \begin{array}{l} y \in \{1, 2, \ldots, C\} \\ y \in \mathbb{R} \end{array}$$

- Learning
  - Learn prediction function $f(\mathbf{x}) \in \mathcal{H}$ from data $\mathcal{D}$

    ($\mathcal{H}$: Hypothesis set/Candidate set)

Quantify the Evaluation

- Measure of quality: expected loss

$$L = \mathbb{E}_P[l(y, f(\mathbf{x}))] \qquad l(y, y')\text{: loss function}$$

- Estimated error

$$\hat{L} = \sum_n l(y_n, f(\mathbf{x}_n)), \qquad f(\mathbf{x}) \in \mathcal{H}$$

- Examples
  - Classification $\quad l(y, f(\mathbf{x})) = \mathbb{I}(y \neq f(\mathbf{x}))$
  - Regression $\quad l(y, f(\mathbf{x})) = ||y_n - f(\mathbf{x}_n)||^2$

Consistent Learner

- $\mathcal{H}$ satisfies $$\boxed{\widehat{L} \underset{N \to \infty}{\longrightarrow} L}$$

$$P\{\sup_{f \in \mathcal{H}} (L(f) - \widehat{L}(f)) > \epsilon\} \to 0 \quad \text{for} \quad \epsilon > 0$$

<Uniform convergence>

- Caution:
  - The definition of consistency is *not*

$$\widehat{L}(f) \to L(f) \quad \text{for} \quad f \in \mathcal{H}$$

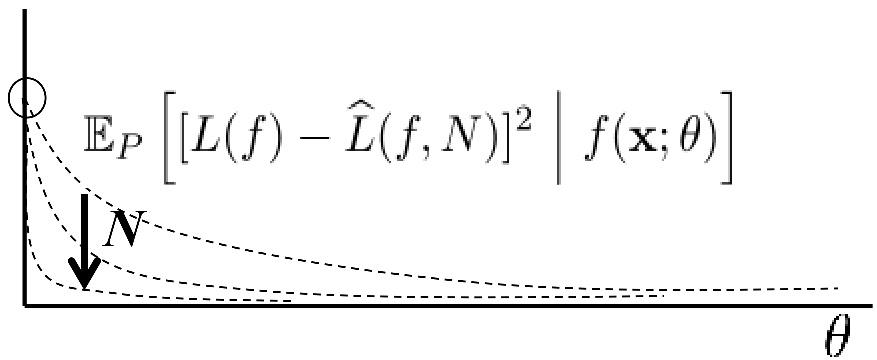
Quiz 1

- Consider a hypothesis set $\mathcal{H}$ which satisfies

$$\mathbb{E}_P\left[[L(f)-\widehat{L}(f,N)]^2 \;\middle|\; f(\mathbf{x};\theta)\right] = \left(\frac{1}{N}\right)^{\theta}$$

$$\mathcal{H} = \{f(\mathbf{x};\theta)|\theta > 0\}$$

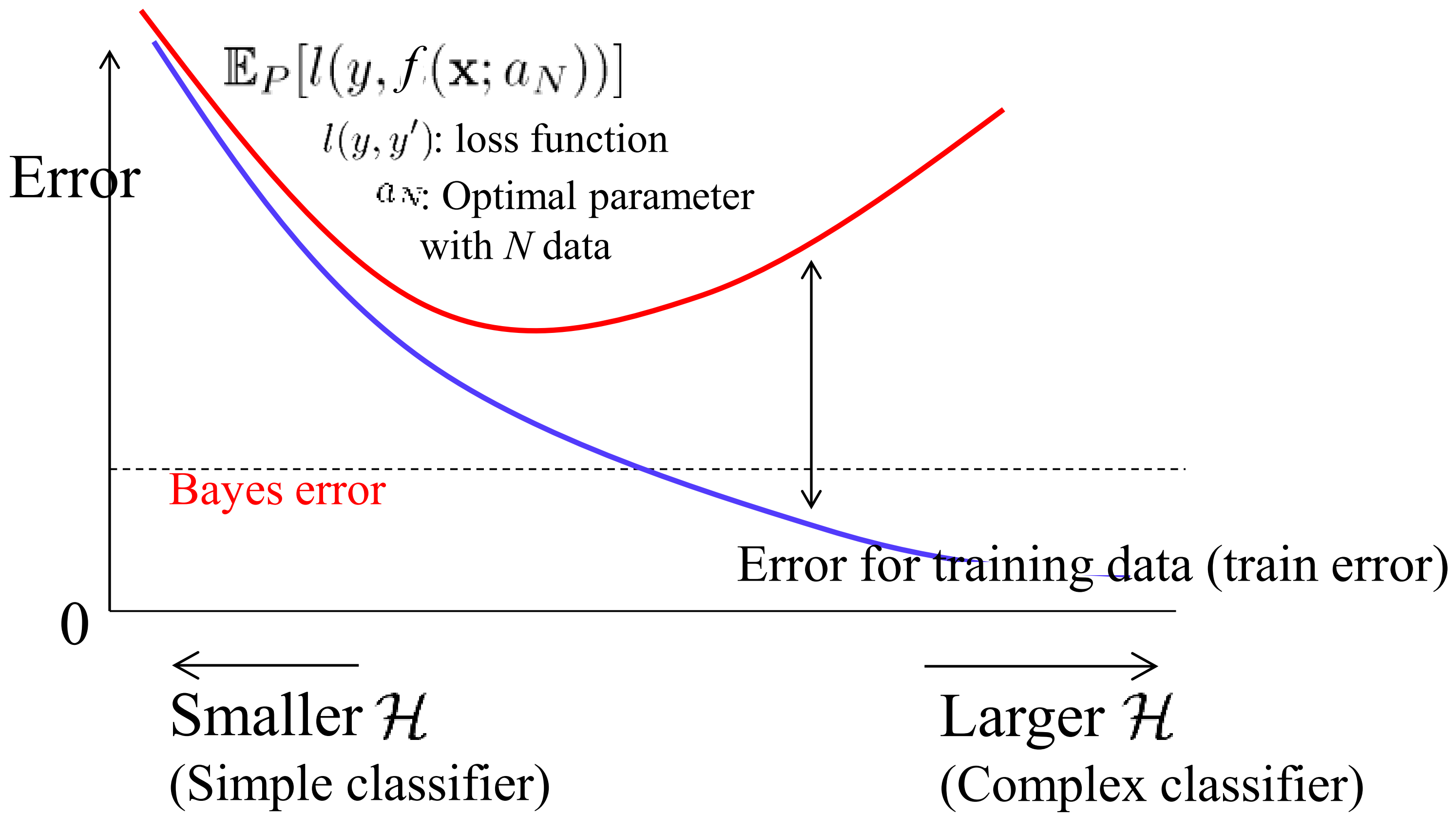
Explain that learning with $\mathcal{H}$ is not consistent though it satisfies $\widehat{L}(f) \to L(f)$.



What is the possible problem in this case?

Consistency and Bayes Error

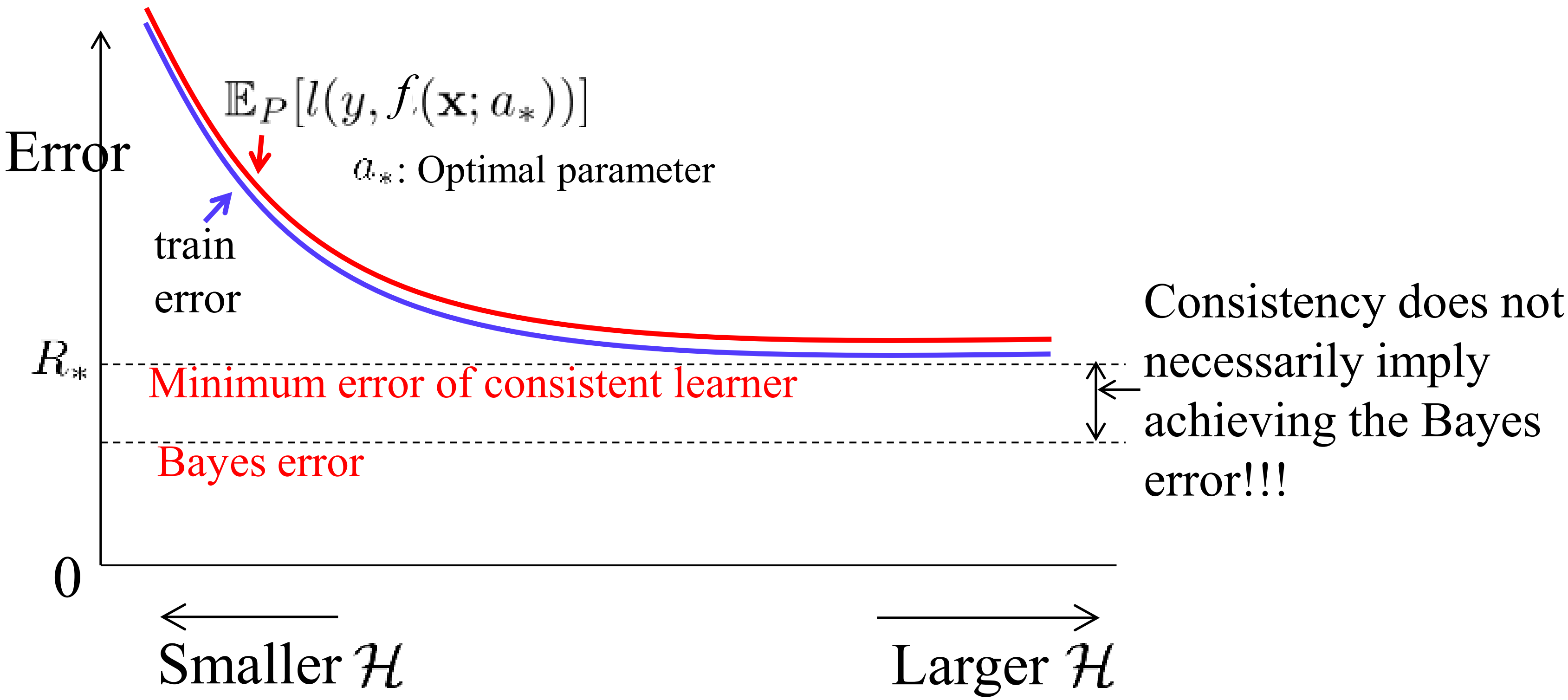
- Minimizing expected error (objective) vs. minimizing estimated error



(For example, a linear classifier with regularization)

Consistency and Bayes Error

- Consistent learner with many data



(For example, a linear classifier with regularization)

Related Terms with Confining $\mathcal{H}$

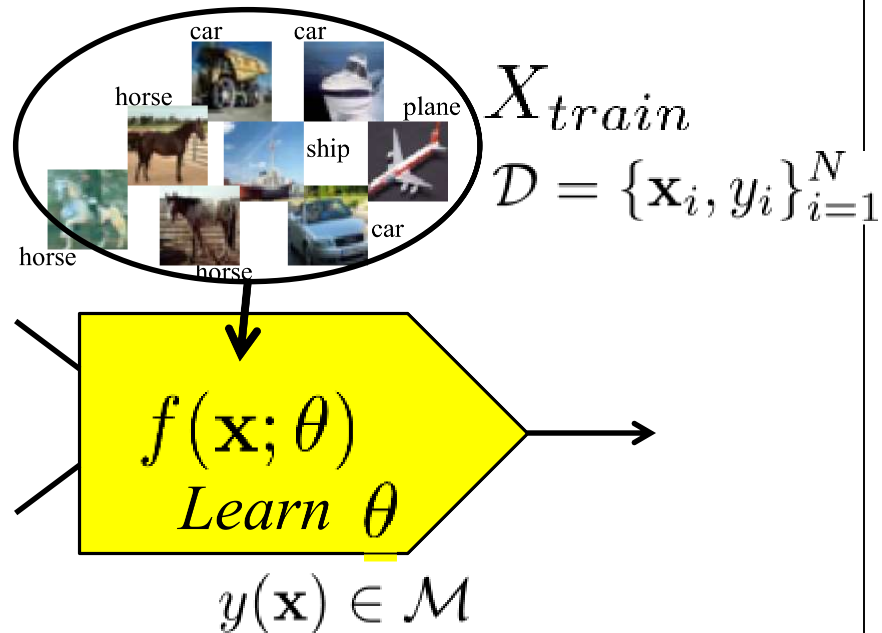
- Linear model →

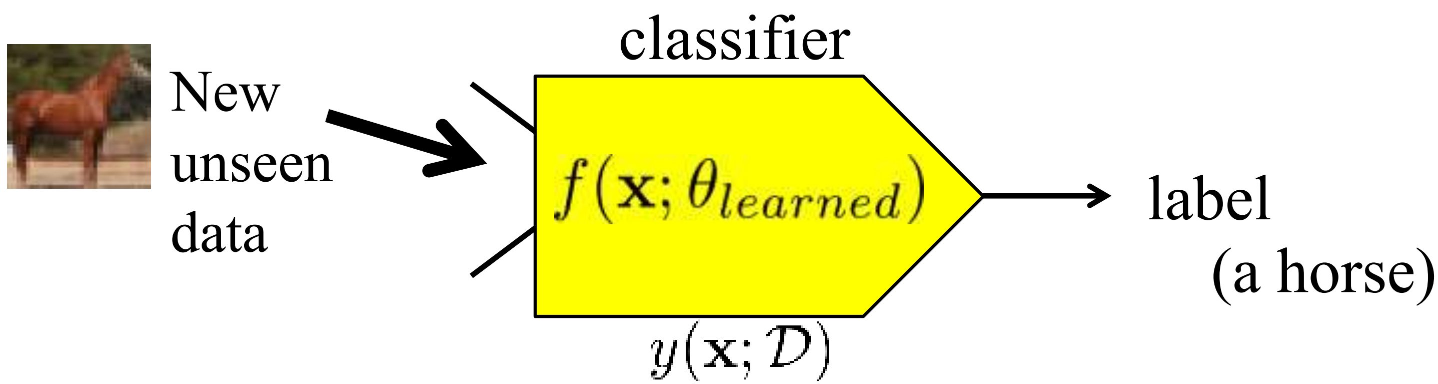
  VC-dim for classification = Dimensionality + 1
- Small number of parameters
- Large margin
- Regularization
- Bias-Variance trade-off
- Generalization ability, overfitting

➔ Many terms are theoretically connected

Supervised Learning (Prediction)

- Method:
  - Learning from *examples* and can classify an *unseen data*

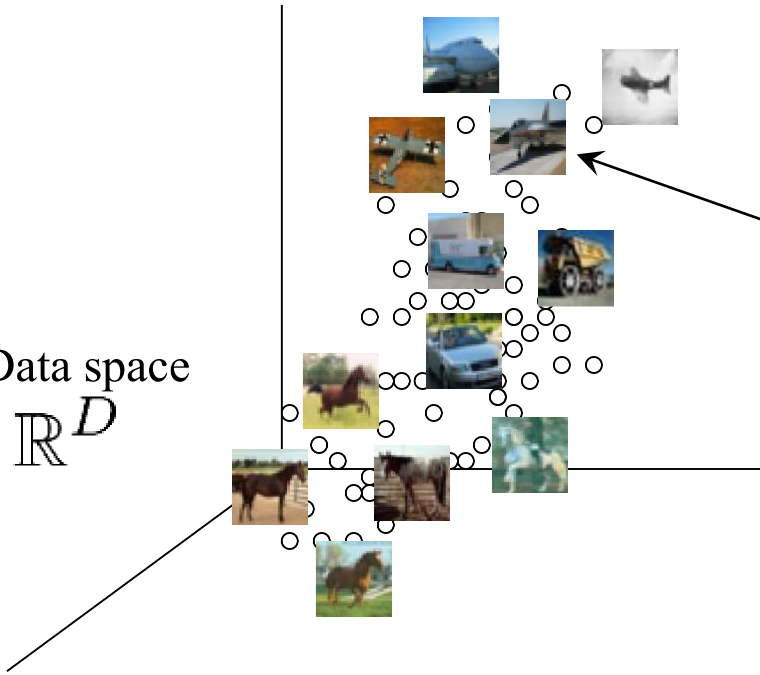


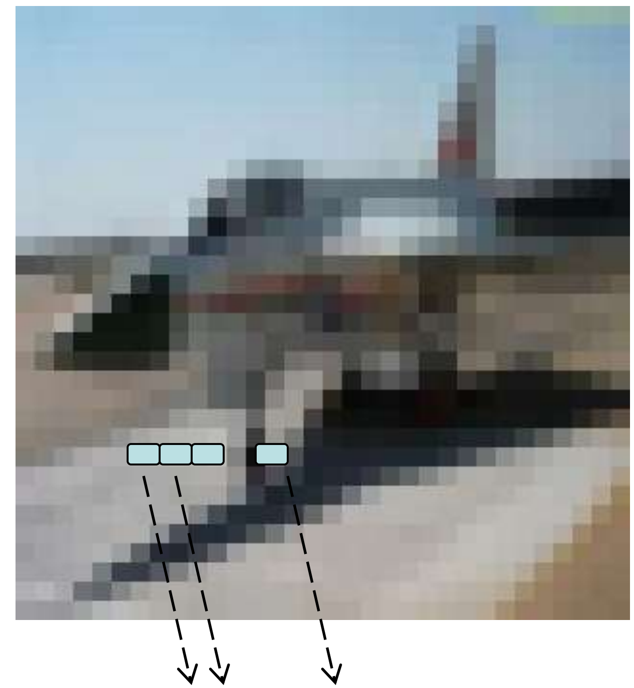


[Based on the assumption of regularity]

Representation of Data

Data space
$\mathbb{R}^D$





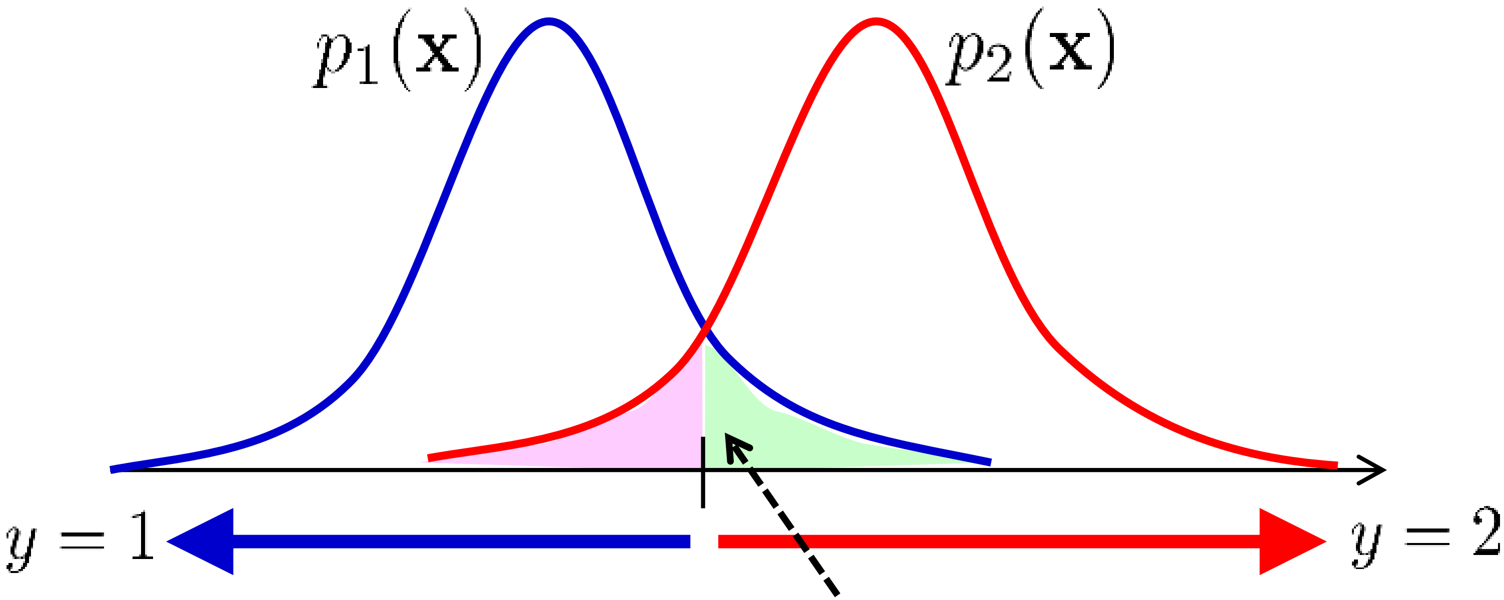
- Each datum is one point in a data space

$$=\underbrace{[1, 2, 5, 10, \ldots]^T}_{D \text{ elements}}$$

Classification

Bayes Optimal Classifier

- Our ultimate goal is *not a zero error.*



$$E_{Bayes} = \frac{1}{2} \int \min[p_1(\mathbf{x}), p_2(\mathbf{x})] d\mathbf{x}$$

Figure credit: Masashi Sugiyama

FISHER DISCRIMINANT ANALYSIS

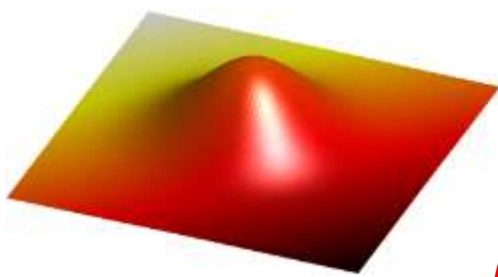
Gaussian Random Variable

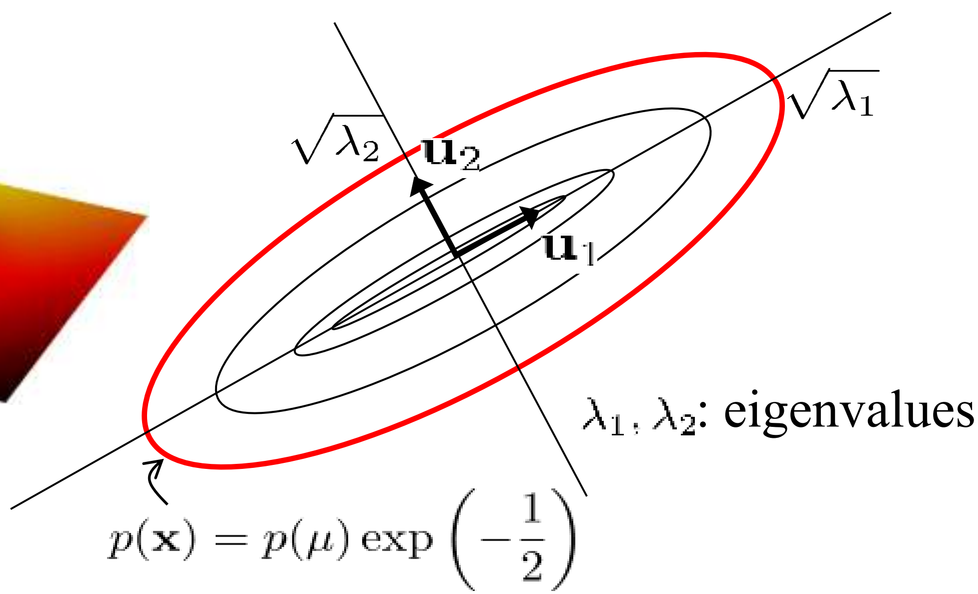
$$p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

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

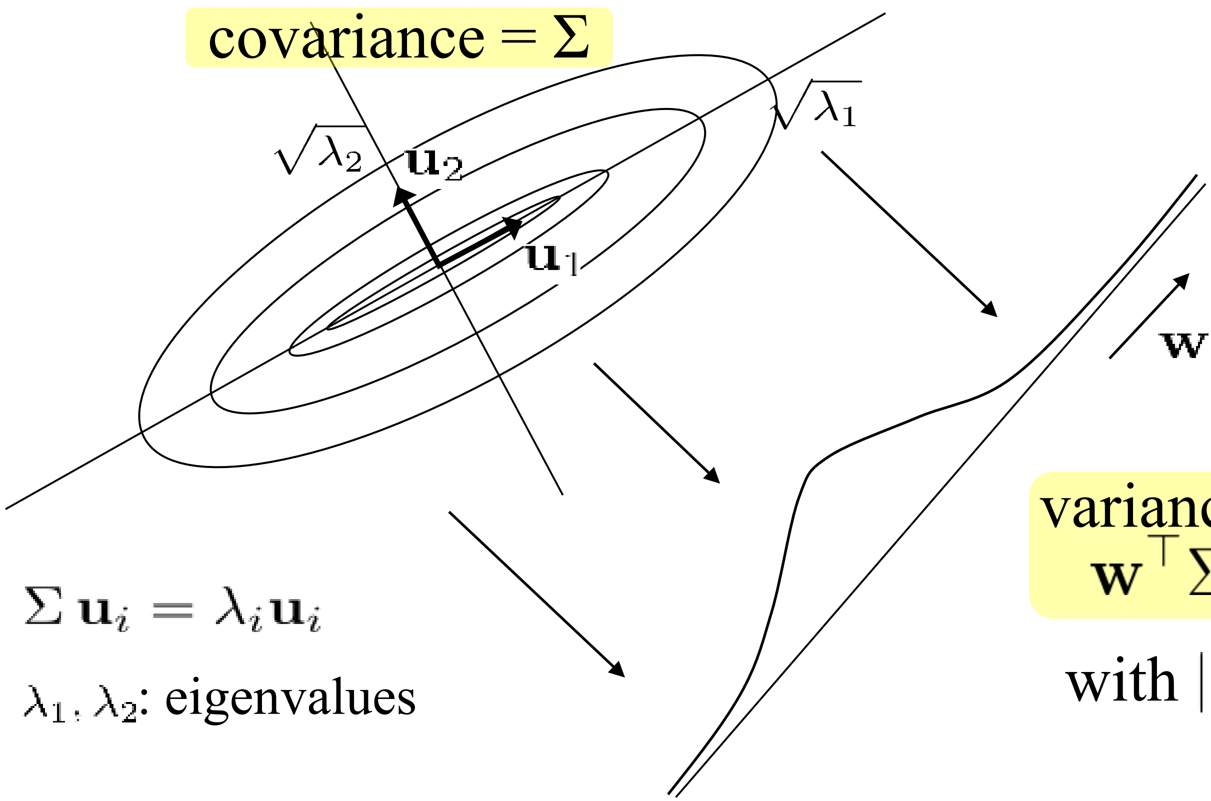
Principal axes are the eigenvector directions of Σ

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





Covariance Matrix and Projection



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

λ_1, λ_2: eigenvalues

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

with $||\mathbf{w}||^2 = 1$

Parameter Estimation

- Maximum Likelihood Estimation

Data: $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$

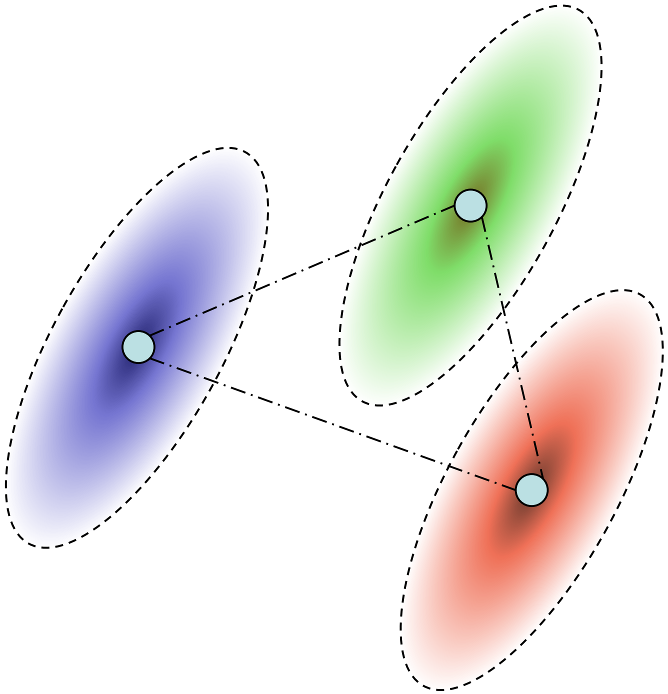
Mean vector: $\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \in \mathbb{R}^D$

Covariance matrix: $\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^\top \in \mathbb{R}^{D \times D}$

Fisher Discriminant Analysis - Consider Two Statistics

- Between-class variance

$$\boxed{var_B(\mathbf{w}) = \mathbf{w}^\top S_B \mathbf{w}}$$

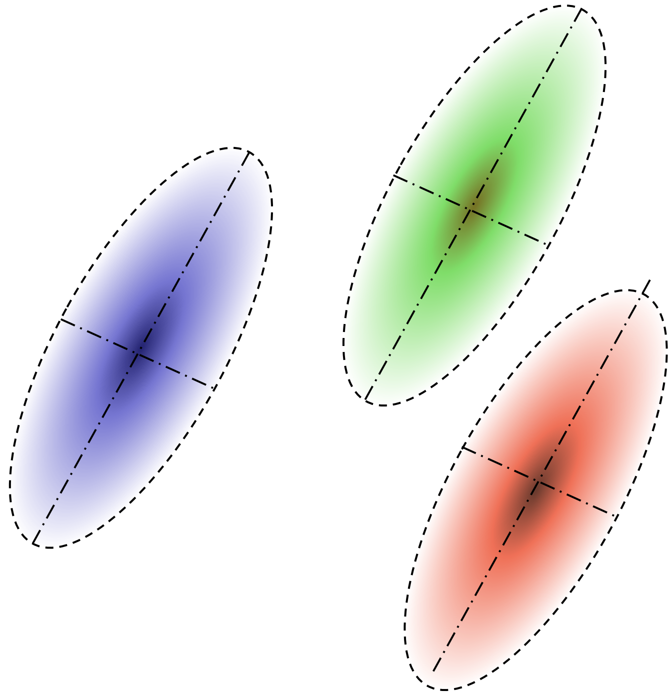


$$S_B = \sum_{c=1}^{C} \frac{N_c}{N} (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^\top$$

$$\mathbf{m} = \sum_{i=1}^{N} \frac{1}{N} \mathbf{x}_i \qquad \mathbf{m}_c = \frac{1}{N_c} \sum_{i \in C_c} \mathbf{x}_i$$

Fisher Discriminant Analysis - Consider Two Statistics

- Within-class variance



$$\boxed{var_W(\mathbf{w}) = \mathbf{w}^\top S_W \mathbf{w}}$$

$$S_W = \sum_{c=1}^{C} \frac{N_c}{N} S_c$$

S_c: covariance matrix of each class

$$S_c = \frac{1}{N_c} \sum_{i \in C_c} (\mathbf{x}_i - \mathbf{m}_c)(\mathbf{x}_i - \mathbf{m}_c)^\top \qquad \mathbf{m}_c = \frac{1}{N_c} \sum_{i \in C_c} \mathbf{x}_i$$

Total Variance

- Total covariance matrix:

$$S_T = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^\top$$

$$S_T = S_W + S_B$$

- Total variance along $\mathbf{w}$:

$$var(\mathbf{w}) = var_W(\mathbf{w}) + var_B(\mathbf{w})$$

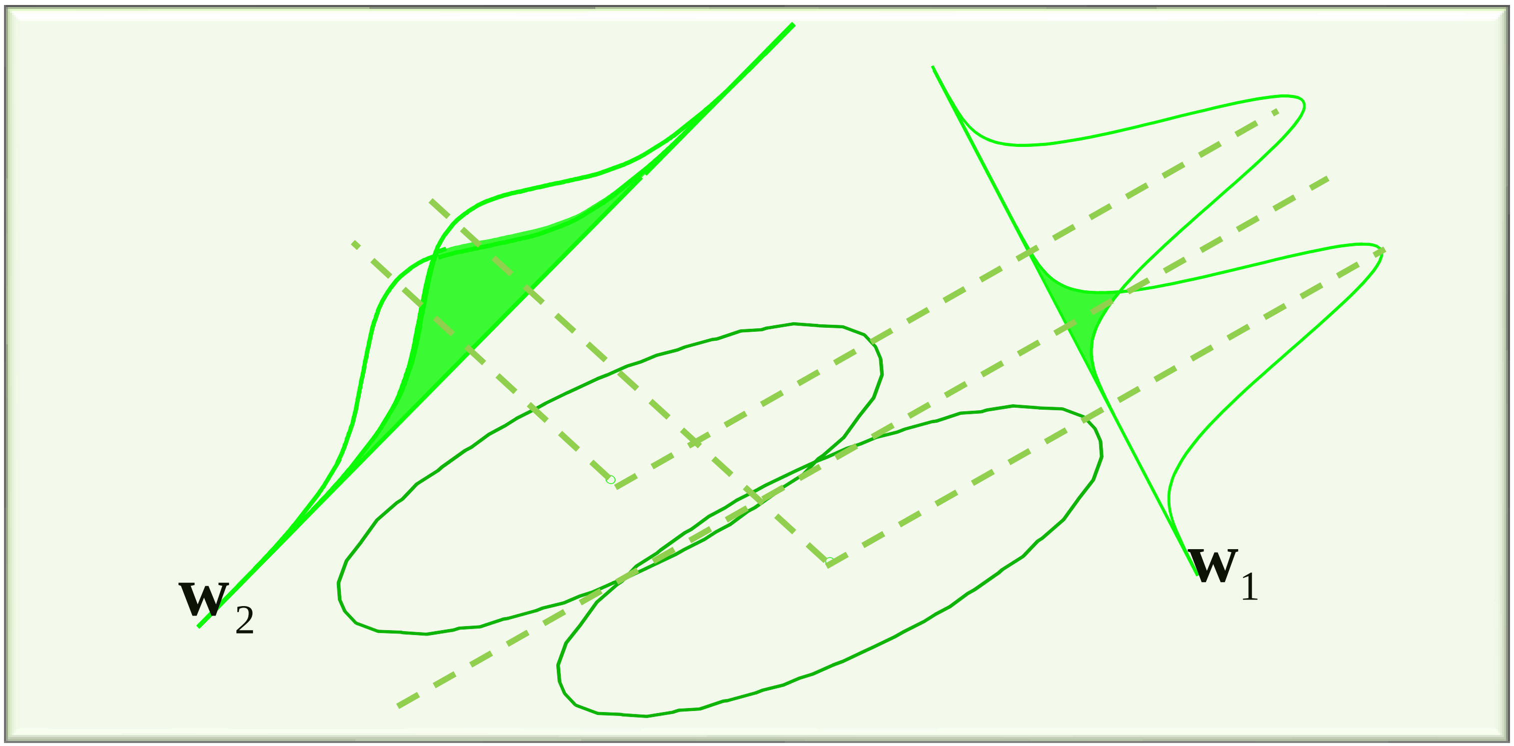
Fisher Discriminant Analysis (FDA)

- Find $\mathbf{w}$ having maximal $var_B(\mathbf{w})$ for give $var_W(\mathbf{w})$.

Between class variance

$$\arg\max_{\mathbf{w}} J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$$

Within class variance



Take the Derivative

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{var_B(\mathbf{w})}{var_W(\mathbf{w})} = \arg\max \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$$

Take the derivative

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$$

$$J'(\mathbf{w}) = \frac{1}{(\mathbf{w}^\top S_W \mathbf{w})^2}\left[2S_B\mathbf{w}(\mathbf{w}^\top S_W \mathbf{w}) - 2\mathbf{w}^\top S_W(\mathbf{w} S_B \mathbf{w})\right] = 0$$

$$S_B\mathbf{w} = \left(\frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}\right) S_W\mathbf{w} = \lambda S_W \mathbf{w}$$

➔ Generalized eigenvector problem

Quiz 1: Closed Form Solution

- Problem: $S_B \mathbf{w} = \lambda S_W \mathbf{w}$
- For two class problem with $N_1 = N_2$:

$$\begin{aligned} S_B &= \sum_{c=1}^{2} \frac{N_c}{N} (\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^\top \\ &= \frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top \end{aligned}$$

Assume S_W is a full-rank matrix.

- Find the closed form solution:

$$\mathbf{w} = ?$$

Quiz 2: How to solve the generalized eigenvector problem?

- Problem: $S_B \mathbf{w} = \lambda S_W \mathbf{w}$

Are the solution eigenvalues real? (non-complex)?

If not, how are these eigenvalues are treated?

Quiz 3: Two different FDAs

- Some papers consider the criterion

$$J_T(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_T \mathbf{w}}$$

instead of $J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}}$

What is the difference?

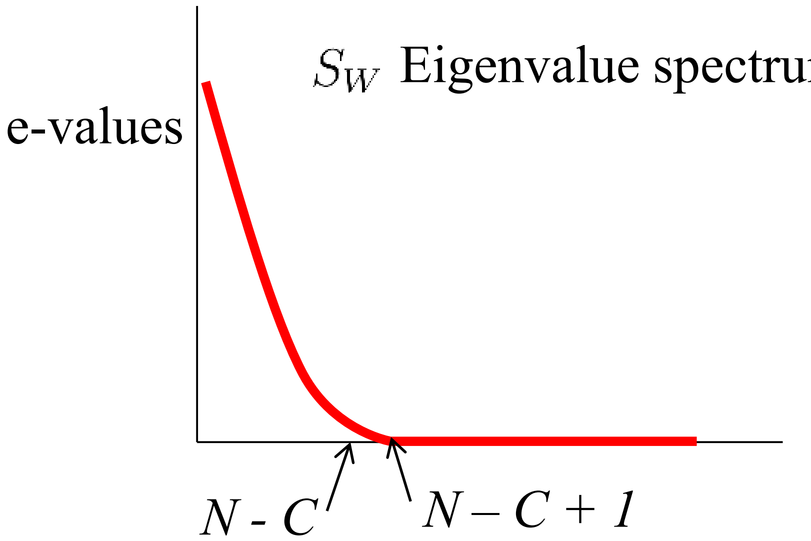
In High Dimensional Space (1/2)

$$D > N$$

$$rank(S_T = S_W + S_B) = N - 1$$

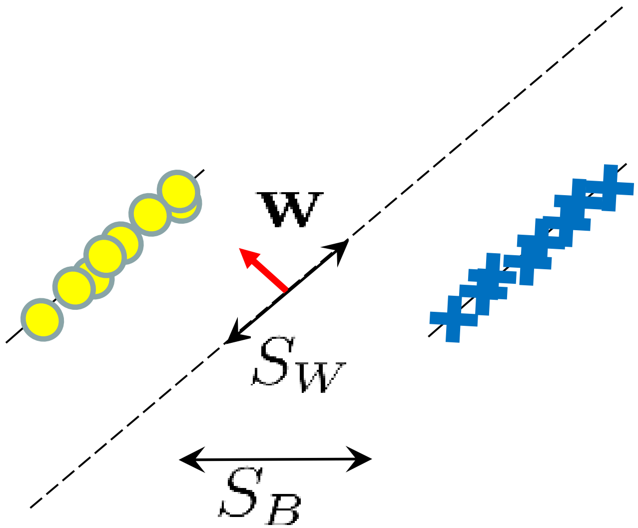
$$rank(S_W) = N - C \longrightarrow$$ S_W: not a full rank matrix

$$rank(S_B) = C - 1$$



In High Dimensional Space (2/2)

- FDA will trivially pick up the null space of S_W as solution.



$$J(\mathbf{w}) = \infty$$

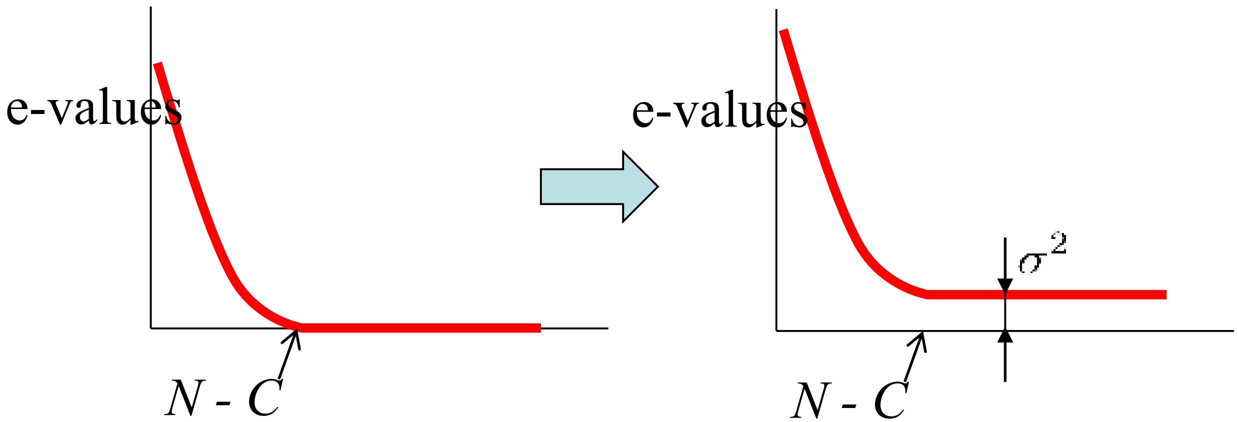
More seriously, the nullspace varies by sampling

➔ Ill-posed problem

Regularization in FDA

- Regularize

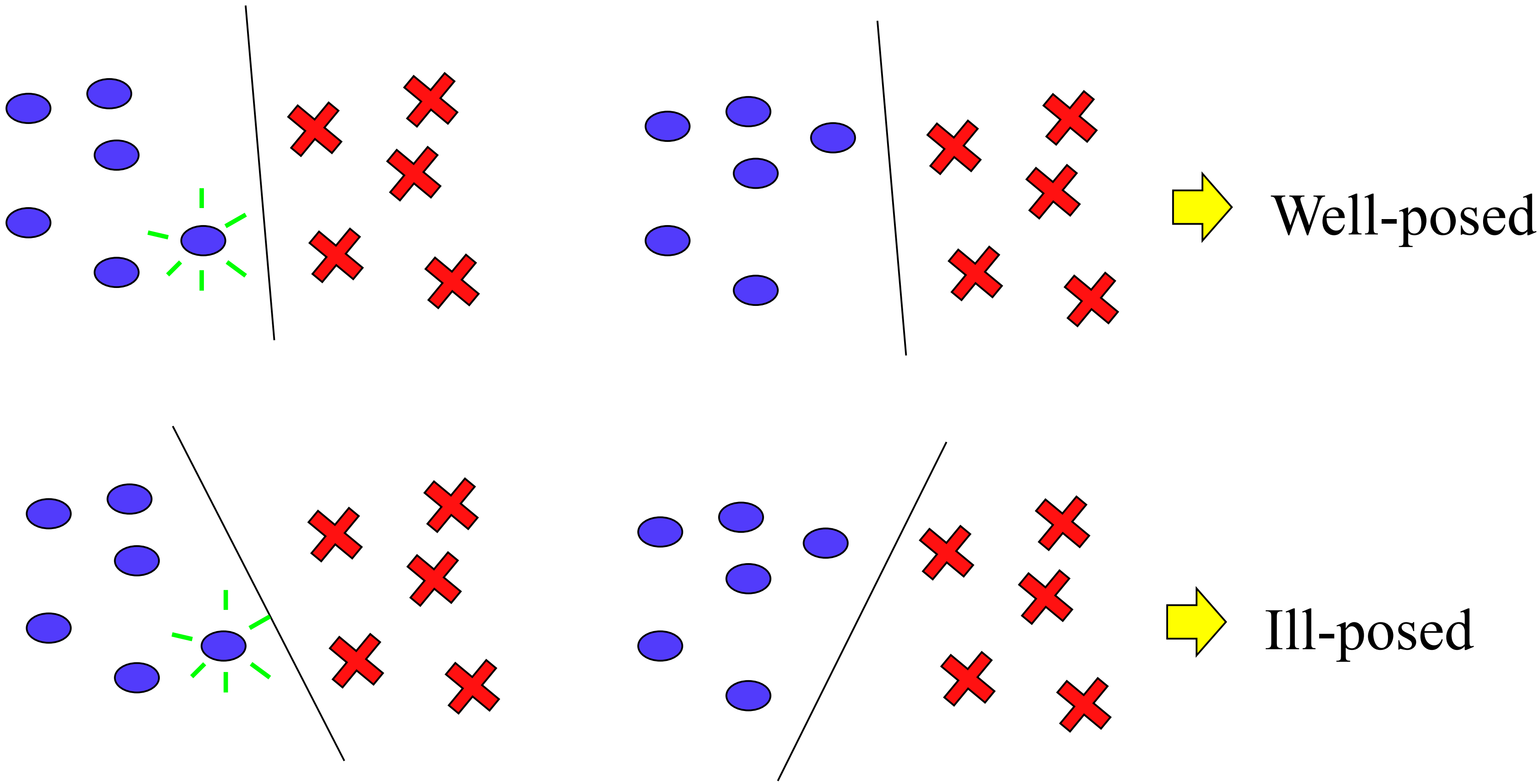
$$S_W \to S_W + \sigma^2 I$$



The problem becomes well-posed.

New solution: $\mathbf{w} = (S_W + \sigma^2 I)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$
for two classes

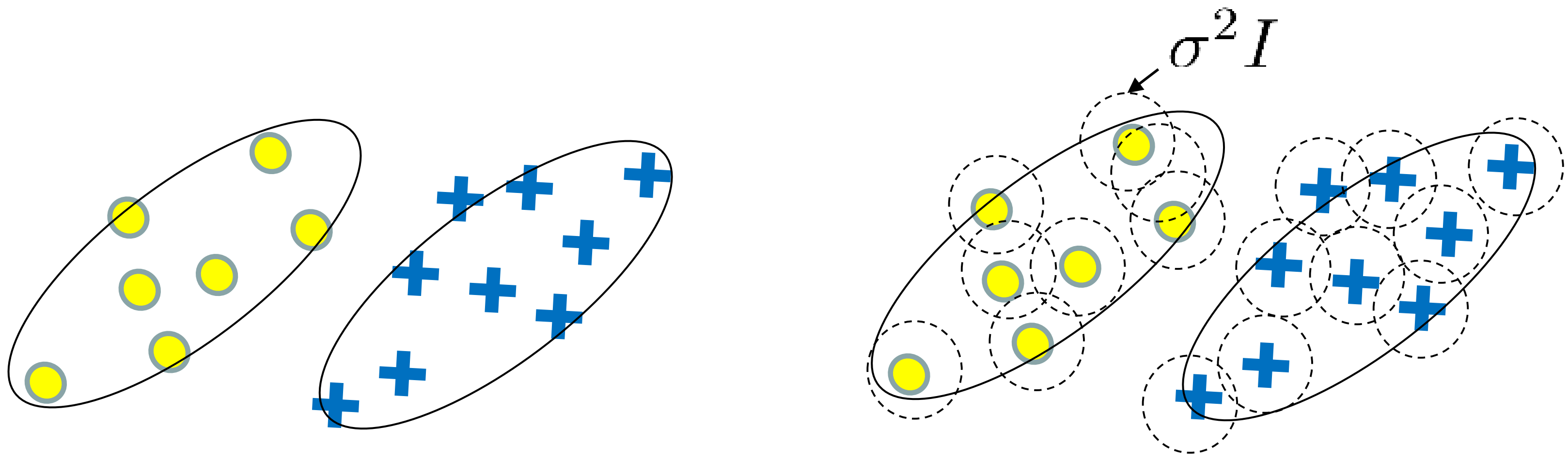
Well-posed Problem vs. Ill-Posed Problem



The variation of configuration may come from the sampling variation.

Quiz 4: New Sw with Infinite Data

- If infinite number of data are generated around each datum with isotropic Gaussian



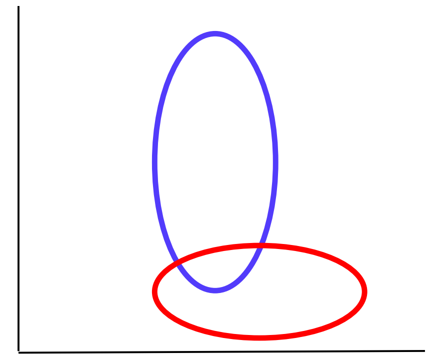
$$S_W = \frac{1}{N}\sum_{c=1}^{C}\sum_{i\in C_c}(\mathbf{x}_i - \mathbf{m}_c)(\mathbf{x}_i - \mathbf{m}_c)^\top \qquad S_W \to \; S_W + \sigma^2 I$$

GRAPHICAL MODELS

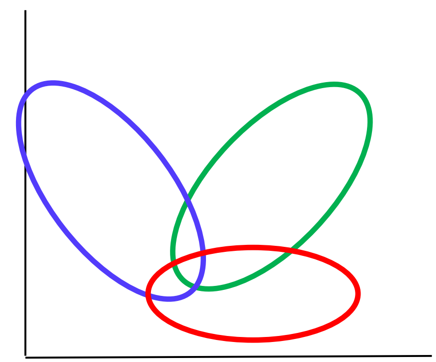
Naïve Bayes As An Extreme Case

- Naïve Bayes

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



  – Simply ignore every correlation and dependencies between variables



- True decomposition:

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

Number of Parameters

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

$$\mu = \quad \in \mathbb{R}^D \qquad \Sigma = \quad \in \mathbb{R}^{D \times D}$$

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

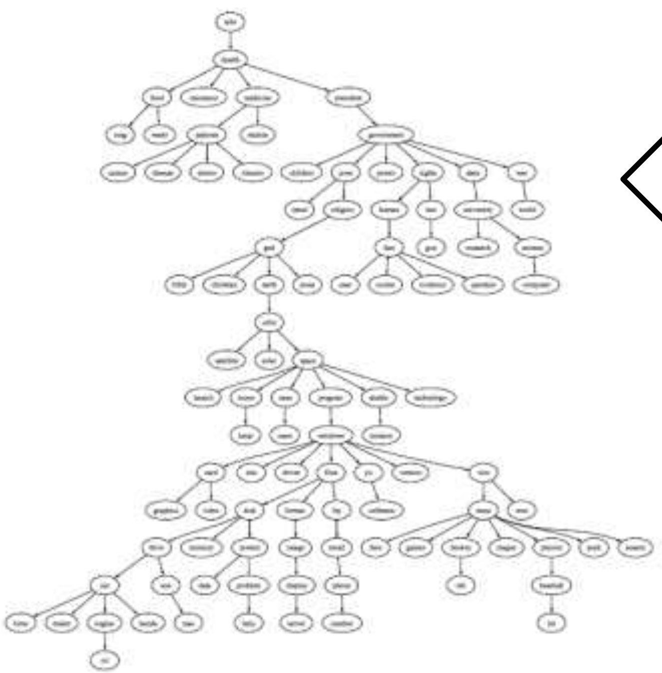
    500 + 501*(500)/2 + 500 + 501*(500)/2

    = 251,500

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

Graphical Models

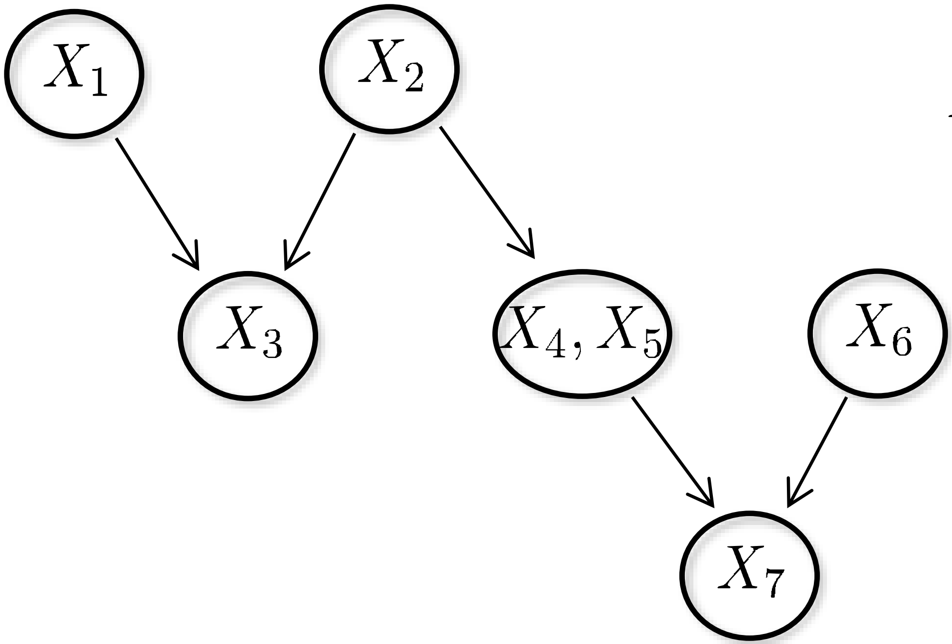
- We utilize probabilities that are represented by the graph structure. (directed & undirected)



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

Directed Graphical Models

- Factorization of a (large) joint pdf

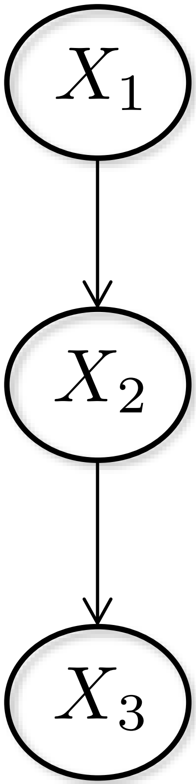


$$
\begin{aligned}
P(X) &= P(X_1, \ldots, X_7) \\
&= P(X_1)P(X_2)P(X_3|X_1, X_2) \\
&\quad P(X_4, X_5|X_2)P(X_6) \\
&\quad P(X_7|X_4, X_5, X_6)
\end{aligned}
$$

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

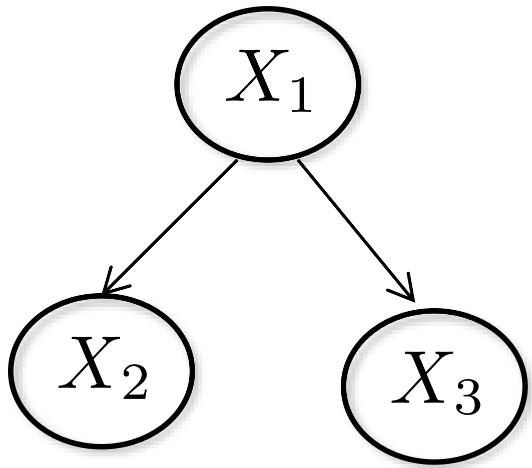
- For given data, make a model for each decomposed probability, then estimate parameters separately.

D-Separations



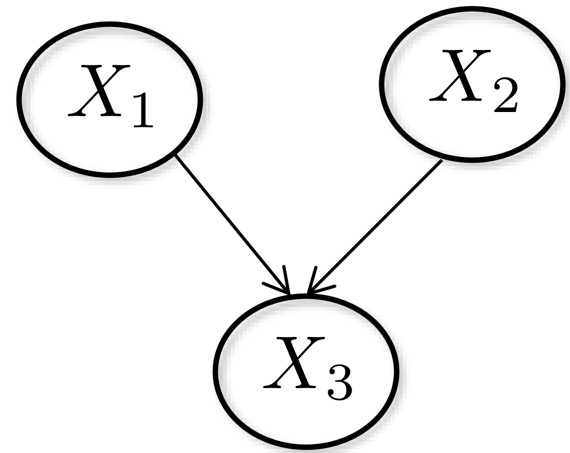
$$X_1 \perp\!\!\!\perp X_3 | X_2$$

Causal path



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

Common cause

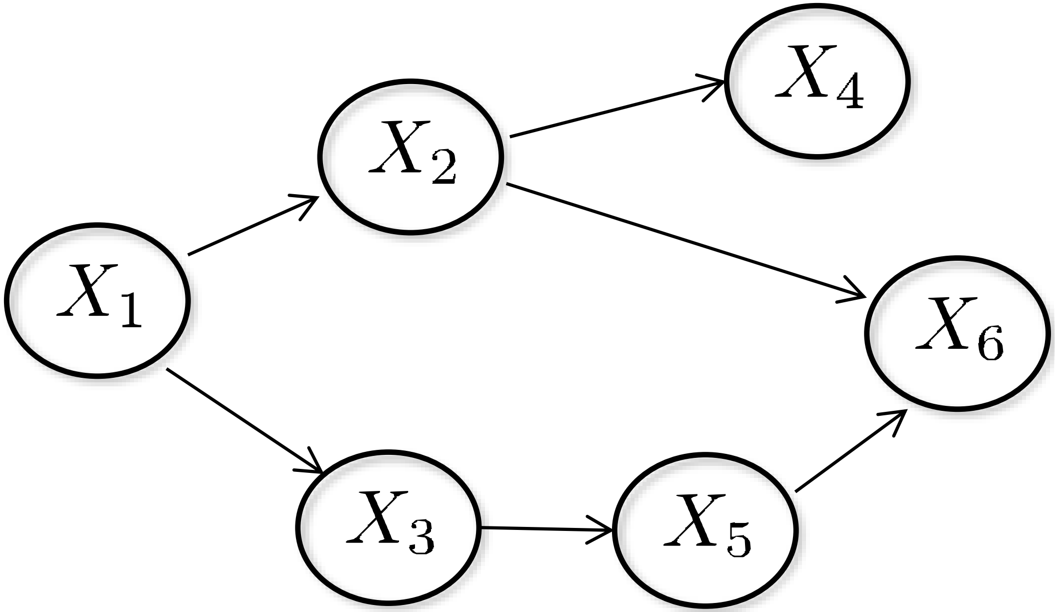


$$X_1 \perp\!\!\!\perp X_2$$

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

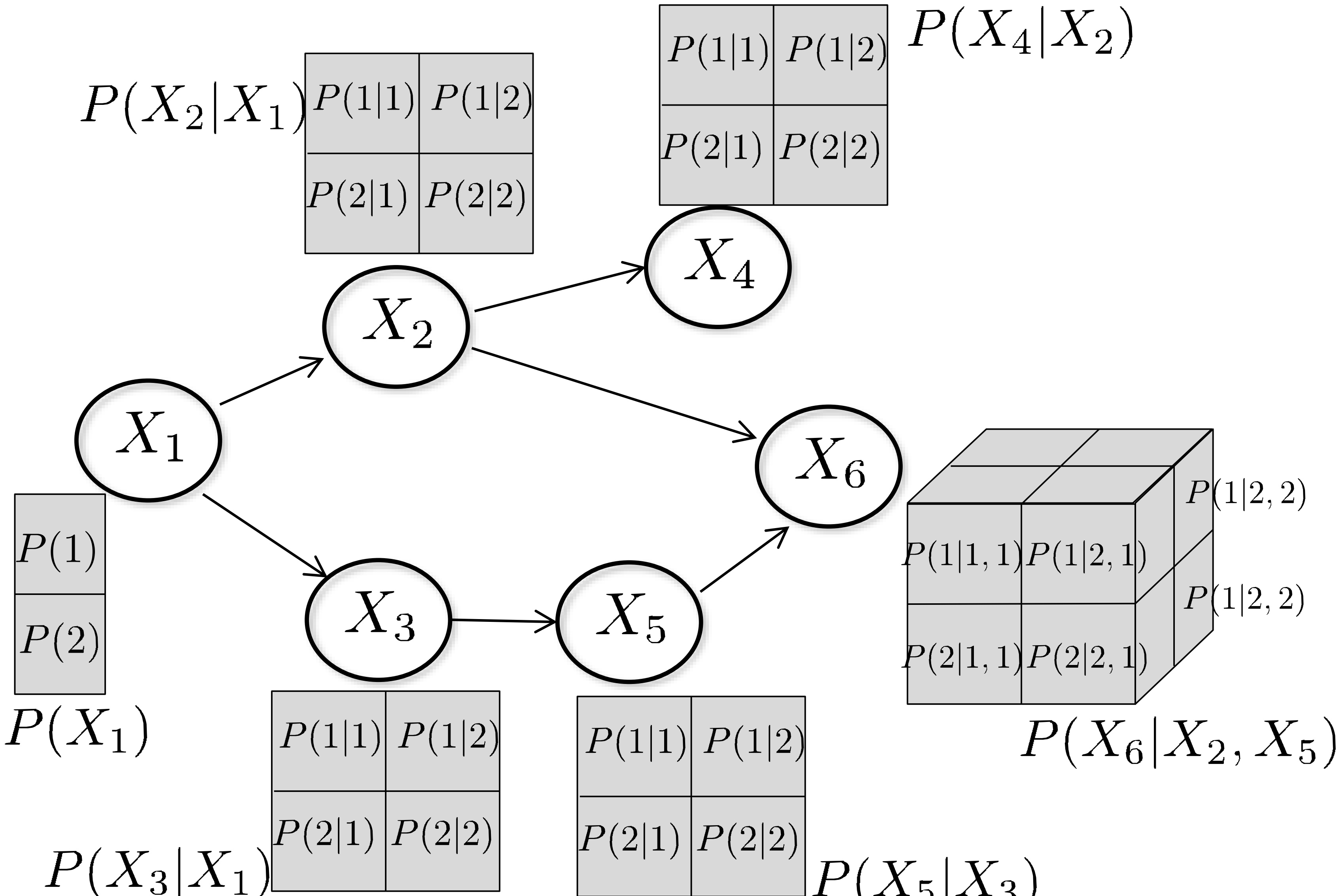
Common effect

Discrete Random Variables

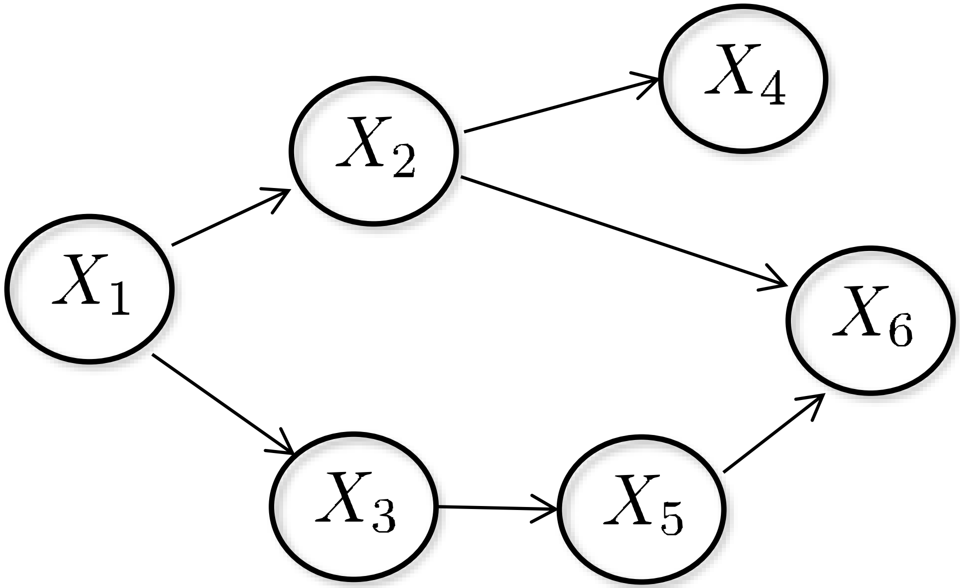


$$P(X_1, \ldots, X_6) = P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)$$
$$P(X_5|X_3)P(X_6|X_2, X_5)$$

Discrete Random Variables



Inference with Discrete Variables



$P(X_1|X_6)?$

$$P(X_1, X_6) = \sum_{X_2, X_3, X_4, X_5} P(X_1, \ldots, X_6)$$

$$= \sum_{X_2, \ldots, X_5} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2) \\ P(X_5|X_3)P(X_6|X_2, X_5)$$

Inference with Discrete Variables

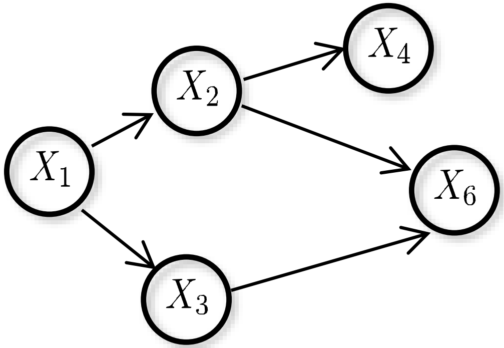
$$\sum_{X_2,\ldots,X_5} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)P(X_5|X_3)P(X_6|X_2,X_5)$$

$$= \sum_{X_2,X_3,X_4} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)\sum_{X_5} P(X_5|X_3)P(X_6|X_2,X_5)$$

- Marginalize X_5

$$\sum_{X_5} P(X_5|X_3)P(X_6|X_2,X_5) \rightarrow \sum_{X_5} P(X_5,X_6|X_2,X_3)$$

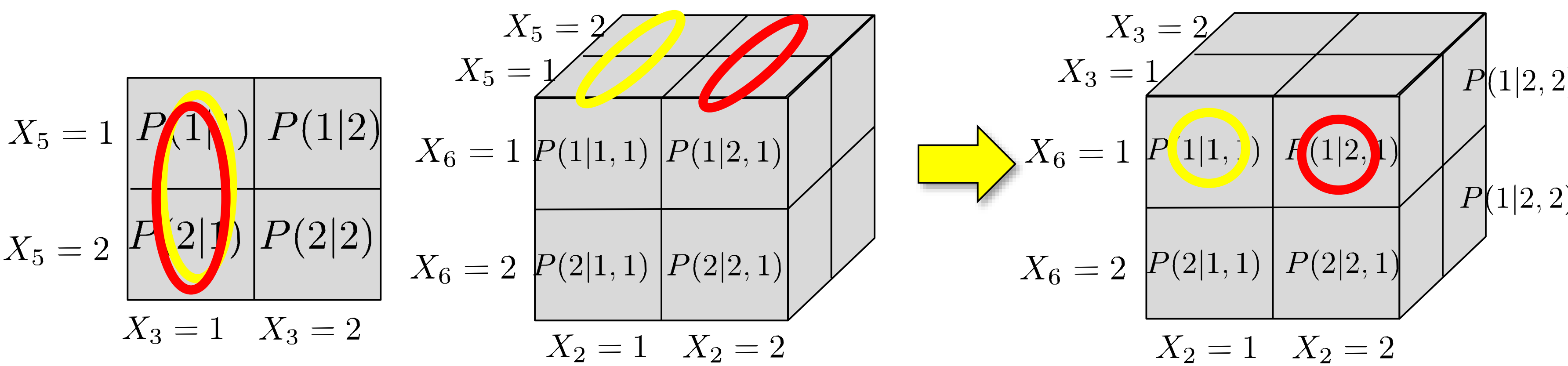
$$= P(X_6|X_2,X_3)$$



Inference with Discrete Variables

- Marginalize X_5

$$P(X_6|X_2,X_3) = \sum_{X_5} P(X_5|X_3)P(X_6|X_2,X_5)$$



$$\sum_{X_5} P(X_5|X_3) \qquad P(X_6|X_2,X_5) \qquad P(X_6|X_2,X_3)$$

Inference with Discrete Variables

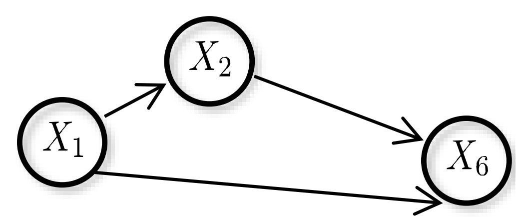
$$\sum_{X_2,X_3,X_4} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)\underline{\sum_{X_5} P(X_5|X_3)P(X_6|X_2,X_5)}$$

$$= \sum_{X_2,X_3,X_4} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_4|X_2)\underline{P(X_6|X_2,X_3)}$$

$$= \sum_{X_2,X_3} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_6|X_2,X_3)\sum_{X_4} P(X_4|X_2)$$

$$= \sum_{X_2,X_3} P(X_1)P(X_2|X_1)P(X_3|X_1)P(X_6|X_2,X_3)$$

$$= \sum_{X_2} P(X_1)P(X_2|X_1)\sum_{X_3} P(X_3|X_1)P(X_6|X_2,X_3)$$



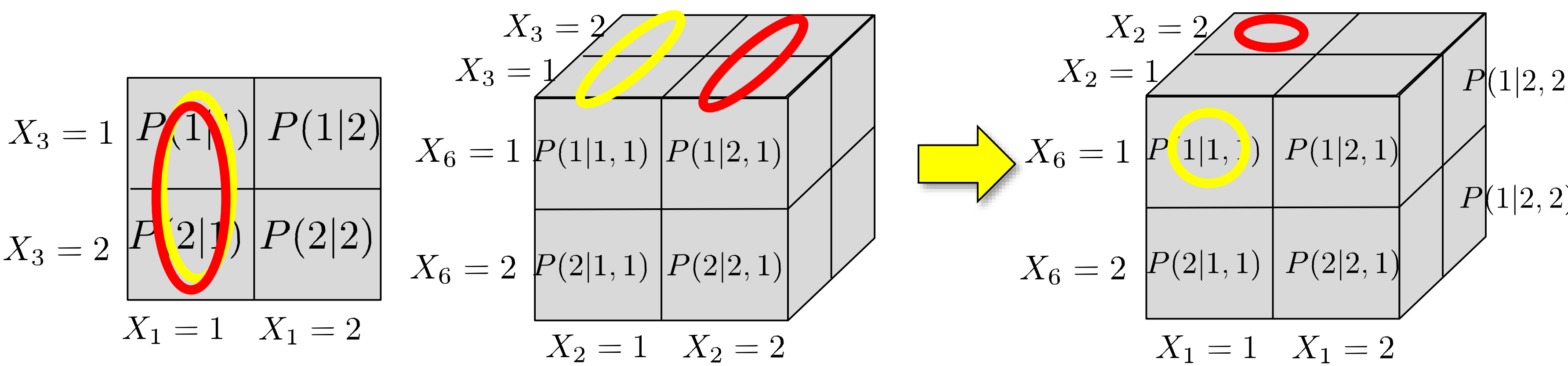
- Marginalize X_3

$$\sum_{X_3} P(X_3|X_1)P(X_6|X_2,X_3) \rightarrow \sum_{X_3} P(X_3,X_6|X_1,X_2) = P(X_6|X_1,X_2)$$

Inference with Discrete Variables

- Marginalize X_3

$$P(X_6|X_1, X_2) = \sum_{X_3} P(X_3|X_1)P(X_6|X_2, X_3)$$



$$\sum_{X_3} P(X_3|X_1) \qquad P(X_6|X_2, X_3) \quad \rightarrow \quad P(X_6|X_1, X_2)$$

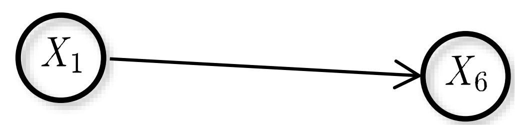
Inference with Discrete Variables

$$\sum_{X_2} P(X_1)P(X_2|X_1)\sum_{X_3} P(X_3|X_1)P(X_6|X_2,X_3)$$

$$= \sum_{X_2} P(X_1)P(X_2|X_1)P(X_6|X_1,X_2)$$

$$= P(X_1)\sum_{X_2} P(X_2|X_1)P(X_6|X_1,X_2)$$

- Marginalize X_2



$$P(X_6|X_1) = \sum_{X_2} P(X_2|X_1)P(X_6|X_1,X_2)$$

Inference with Discrete Variables

$$P(X_1)\sum_{X_2} P(X_2|X_1)P(X_6|X_1,X_2)$$

$$\rightarrow P(X_1)\sum_{X_2} P(X_2,X_6|X_1) = P(X_1)P(X_6|X_1)$$

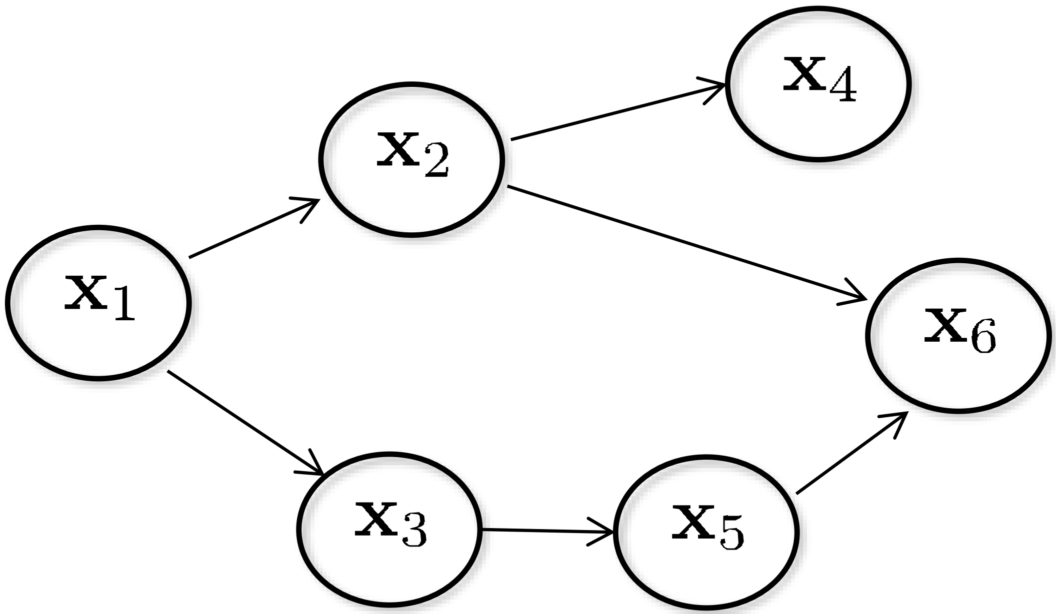
$$= P(X_1,X_6)$$

From the joint distribution,

$$P(X_6) = \sum_{X_1} P(X_1,X_6)$$

$$P(X_1|X_6) = \frac{P(X_1,X_6)}{P(X_6)}$$

Continuous Random Variables



$$p(\mathbf{x}_1, \ldots, \mathbf{x}_6) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_3|\mathbf{x}_1)p(\mathbf{x}_4|\mathbf{x}_2) \\ p(\mathbf{x}_5|\mathbf{x}_3)p(\mathbf{x}_6|\mathbf{x}_2, \mathbf{x}_5)$$

- Each probability density is a Gaussian

Continuous Random Variables

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_6) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_3|\mathbf{x}_1)p(\mathbf{x}_4|\mathbf{x}_2)$$
$$p(\mathbf{x}_5|\mathbf{x}_3)p(\mathbf{x}_6|\mathbf{x}_2, \mathbf{x}_5)$$

- Jointly Gaussian with 6 variables
  - Need 6 + 6*7/2 = 27 parameters
- Using graphical model:
  - Need to obtain the parameters of

$$p(\mathbf{x}_1, \mathbf{x}_2), p(\mathbf{x}_1, \mathbf{x}_3), p(\mathbf{x}_2, \mathbf{x}_4), p(\mathbf{x}_3, \mathbf{x}_5), p(\mathbf{x}_2, \mathbf{x}_5, \mathbf{x}_6)$$

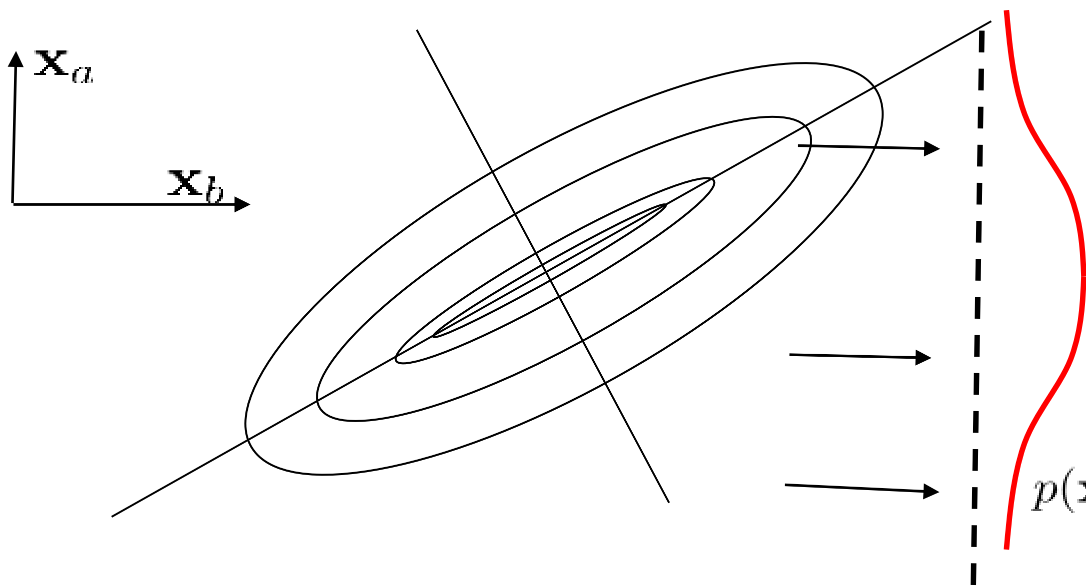
  - We need to estimate only the following 19 parameters

$$\mu_1, \ldots, \mu_6, \sigma_1^2, \ldots, \sigma_6^2, \sigma_{12}, \sigma_{13}, \sigma_{24}, \sigma_{35}, \sigma_{25}, \sigma_{26}, \sigma_{56}$$

Gaussian Random Variable – Marginal

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

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



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

Gaussian Random Variable – Marginal

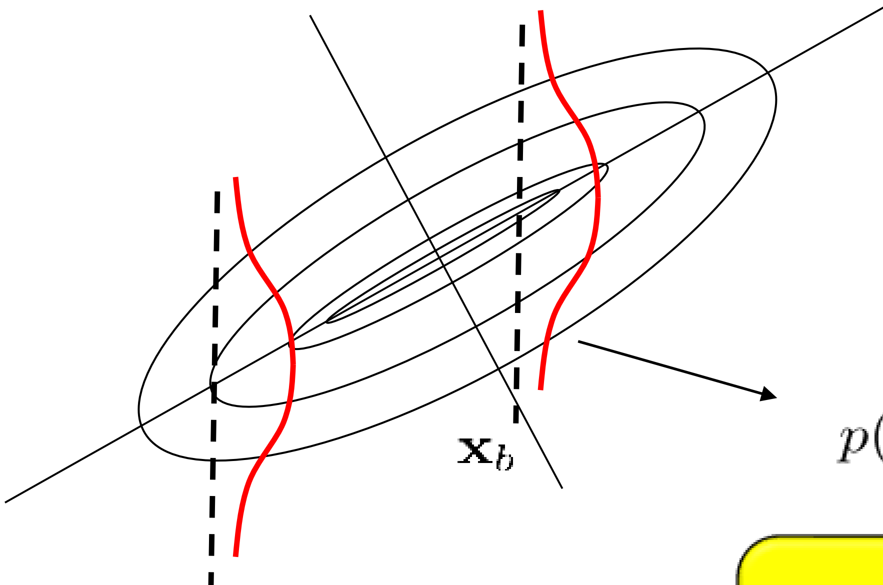
$$p(\mathbf{x}_a, \mathbf{x}_b) = \frac{1}{\sqrt{2\pi}^D \left| \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix} \right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \begin{pmatrix} \mathbf{x}_a - \mu_a \\ \mathbf{x}_b - \mu_b \end{pmatrix}^\top \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_a - \mu_a \\ \mathbf{x}_b - \mu_b \end{pmatrix} \right)$$

$$\int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b = \frac{1}{\sqrt{2\pi}^D \left|\Sigma_a\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\mathbf{x}_a - \mu_a)^\top \Sigma_a^{-1} (\mathbf{x}_a - \mu_a) \right)$$

$$= \mathcal{N}(\mu_a, \Sigma_a)$$

Gaussian Random Variable – Conditional

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



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

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

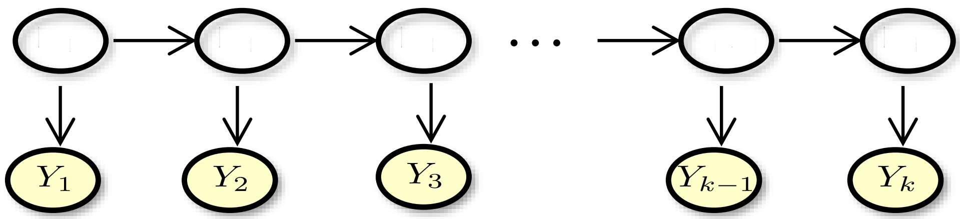
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mu_{a|b}, \Sigma_{a|b})$$

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

KALMAN FILTER

Filtering

- Linear Dynamical Systems (LDS) with Gaussian noise

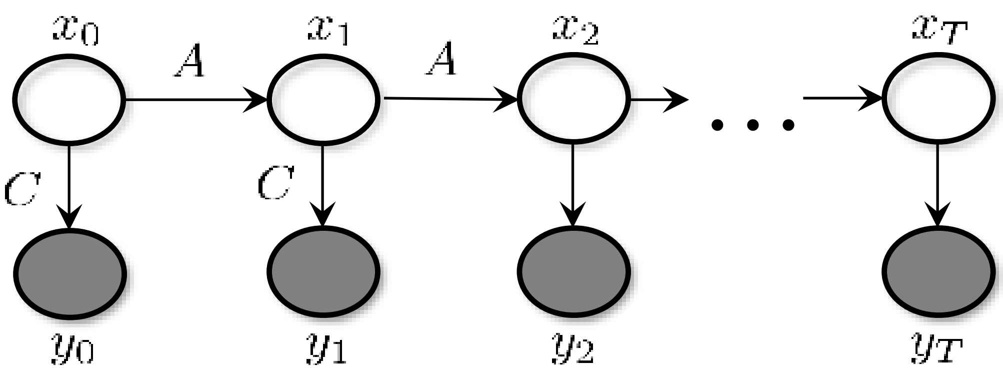


$$p(y_1 \ldots, y_K, x_1, \ldots, x_K) = p(x_1)p(y_1|x_1) \prod_{t=1}^{K-1} p(x_{t+1}|x_t)p(y_t|x_t)$$

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

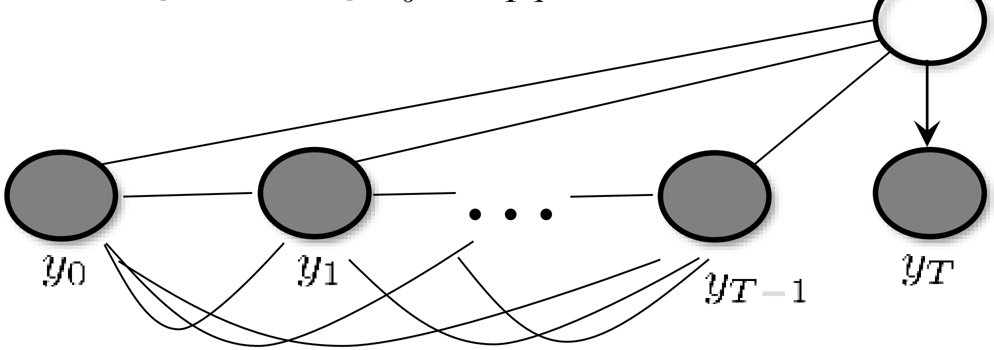
$$y_t = Cx_t + v_t \qquad v_t \sim \mathcal{N}(0, R)$$

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

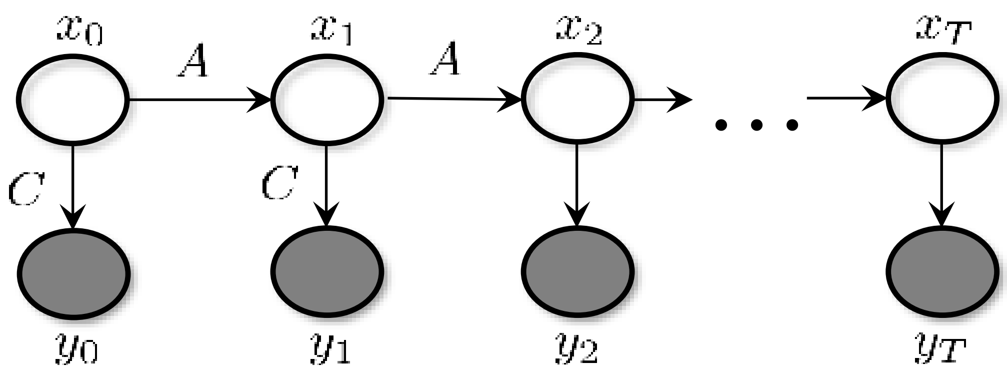


Filtering

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

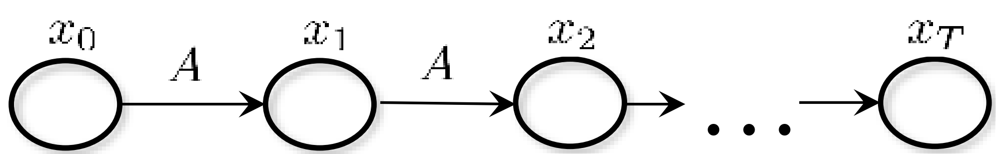
"Conditional marginalization"
Marginalization from the left

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

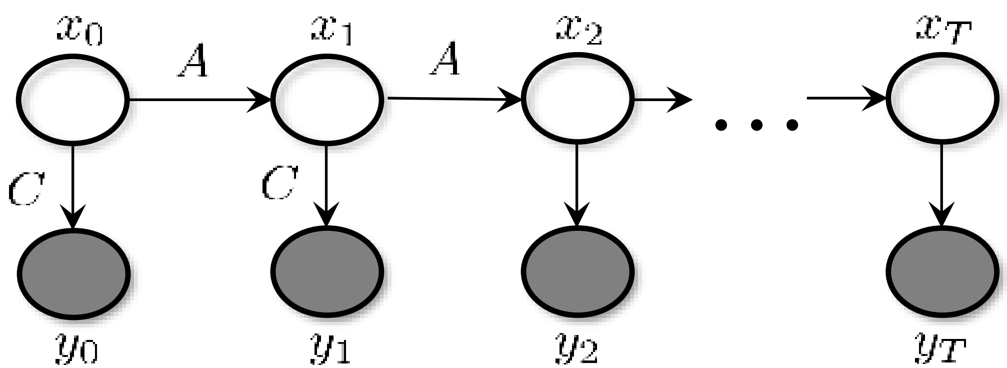


$$\mu_{t+1} = 0$$
$$\begin{aligned}\Sigma_{t+1} &= \mathbb{E}[x_{t+1}x_{t+1}^\top] = \mathbb{E}[(Ax_t + Gw_t)(Ax_t + Gw_t)^\top] \\ &= A\mathbb{E}[x_t x_t^\top]A^\top + G\mathbb{E}[w_t w_t^\top]G^\top \\ &= A\Sigma_t A^\top + GQG^\top\end{aligned}$$

Also, for joint density if necessary

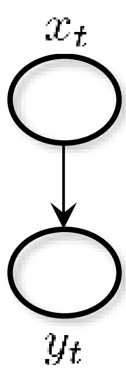
$$\mathbb{E}[x_t x_{t+1}^\top] = \Sigma_t A^\top$$

Kalman Filter



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

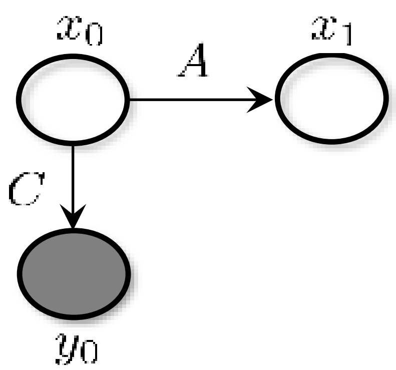
Unconstrained distribution



$$\mu_{y_t} = 0$$

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

Kalman Filter



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

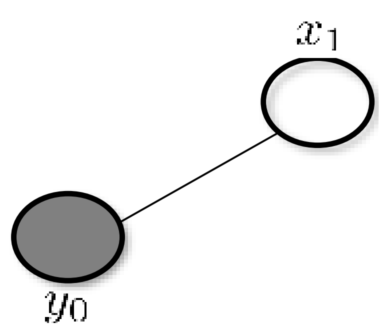
$$\begin{pmatrix} \Sigma_0 & \Sigma_{01} & \Sigma_{0y_0} \\ \Sigma_{10} & \Sigma_1 & \Sigma_{1y_0} \\ \Sigma_{y_0 0} & \Sigma_{y_0 1} & \Sigma_{y_0} \end{pmatrix}$$

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



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

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

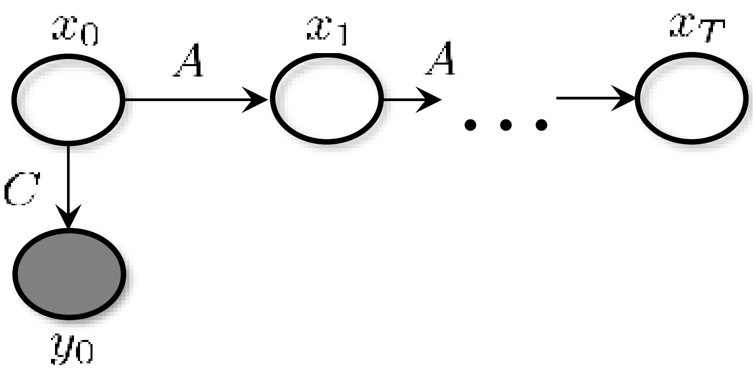
Same

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

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

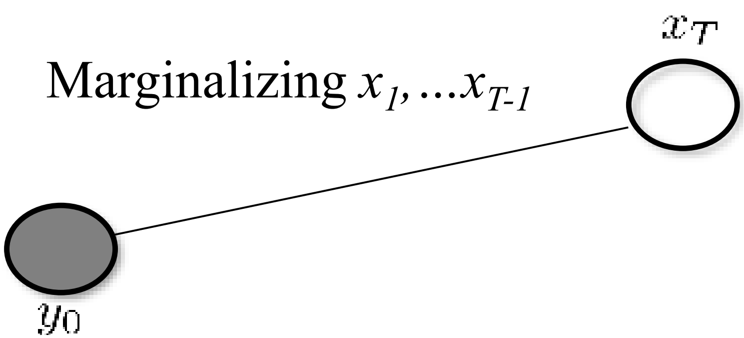
Σ_{y_0} and Σ_1 are from unconstrained distribution. What matters is Σ_{1y_0} .

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 $x_1, \dots x_{T-1}$



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

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

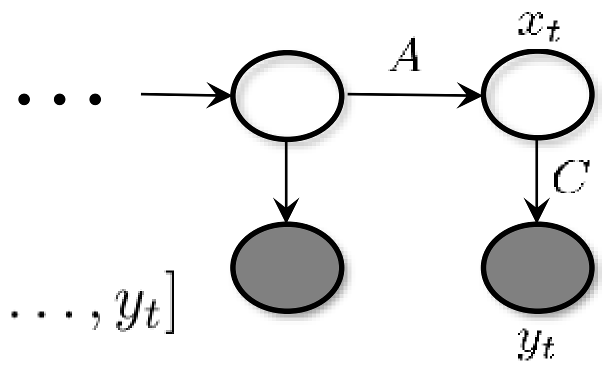
$$\Rightarrow \quad \mu_{T|0} = \mu_T + \Sigma_{Ty_0}\Sigma_{y_0}^{-1}(y_0 - \mu_{y_0})$$

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

What matters is how we can find the covariance of joint density function!!

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

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

- Measurement update

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

Time update

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

Similar to the unconstrained distribution calculation!

Kalman Filter

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

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

Also,

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

Joint:

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

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,

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

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

- Alternative form of K_{t+1}

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

The Kalman Filter? A Tool We Use Everyday

Posted on **April 3, 2012** by **ekrayer**

Almost all modern control systems, both military and commercial, use the Kalman filter. It guided the Apollo 11 lunar module to the moon's surface and is used in phased-array radars to track missiles, inertial guidance systems in aircraft, submarines, missile autopilots, the Global Positioning System, the Space Shuttle and rockets.

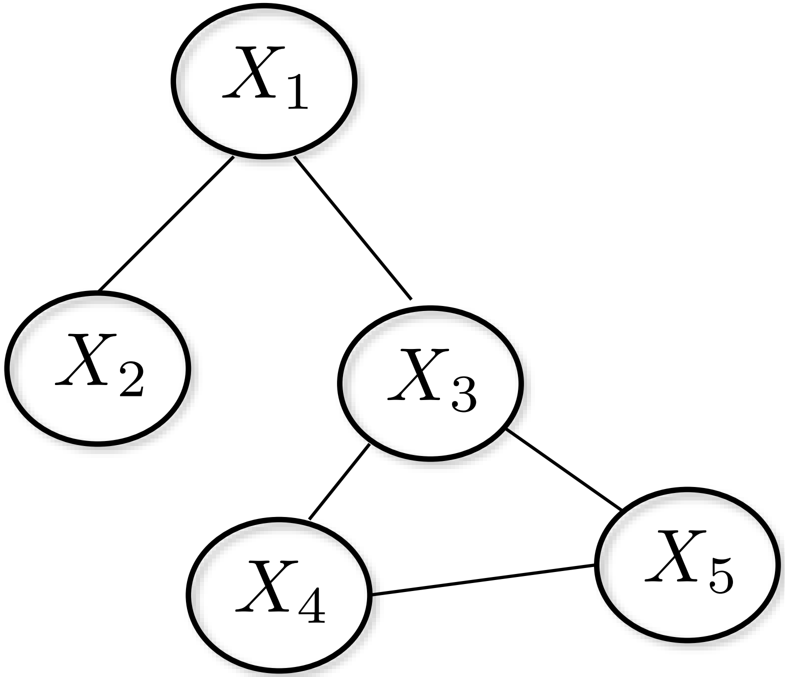
AFOSR initiated support for Dr. Rudolph E. Kalman and Dr. Richard Bucy in 1958 to investigate the use of modern mathematical statistical methods in estimation. At the time, AFOSR program managers saw an opportunity in science for the creation of new mathematical techniques that could alter control applications. With AFOSR support, Kalman and Bucy wrote several papers that revolutionized the area of estimation.

This research ultimately led to the development of what is now known as the Kalman filter, which revolutionized the field of estimation, and had an enormous impact on the design and development of precise navigation systems. The Kalman and Bucy technique of combining and filtering information from multiple sensor sources achieved accuracies that clearly constituted a major breakthrough in guidance technology.

http://afrl.dodlive.mil/2012/04/03/the-kalman-filter-a-tool-we-use-everyday/
http://www.wpafb.af.mil/News/Article-Display/Article/401306/computer-mouse-kalman-filter-trace-origins-to-air-force-basic-research-funding/

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:

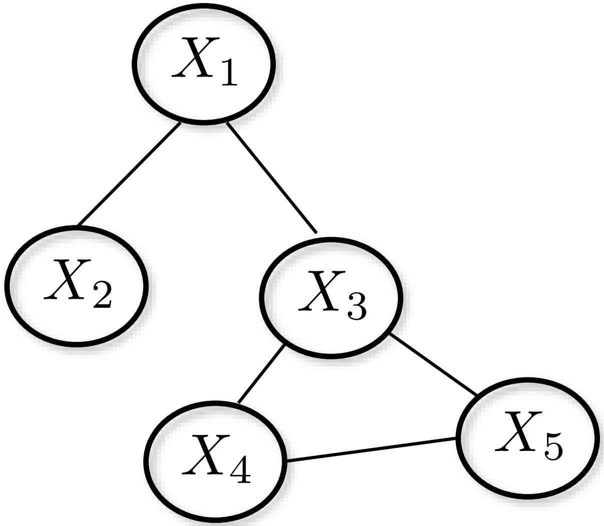
$$X_i \perp\!\!\!\perp X_j | X_{\setminus i,j}$$

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

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

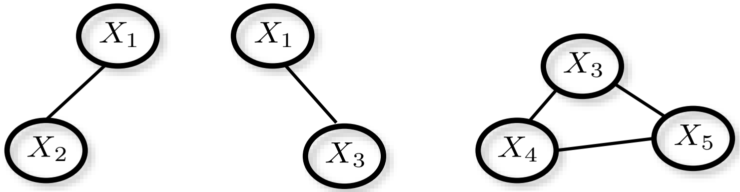
Joint Distribution

- Product of functions on cliques



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

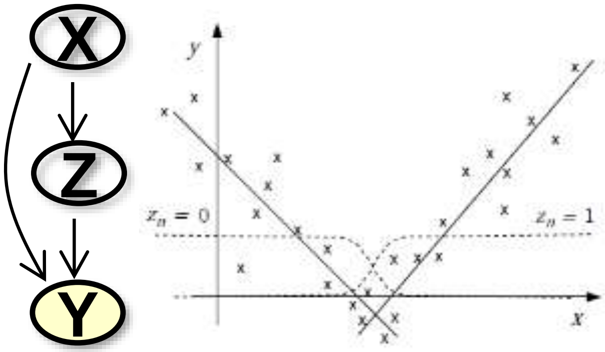
Cliques:

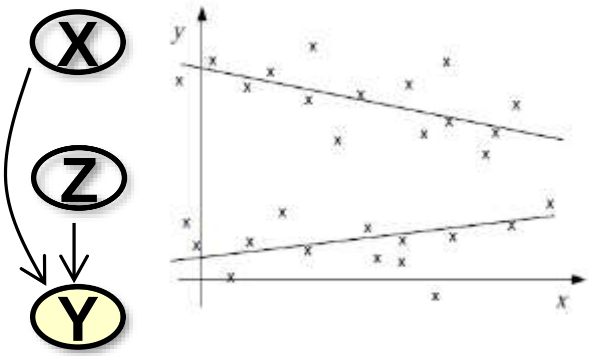


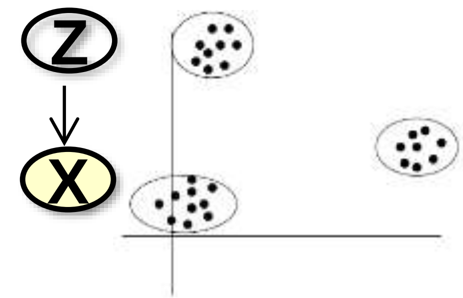
The set of distributions satisfying MRF conditions (Markov random field)
= The set of distributions decomposed by cliques (Gibbs random field)
(Hammersley-Clifford Theorem)

More Fancy Models

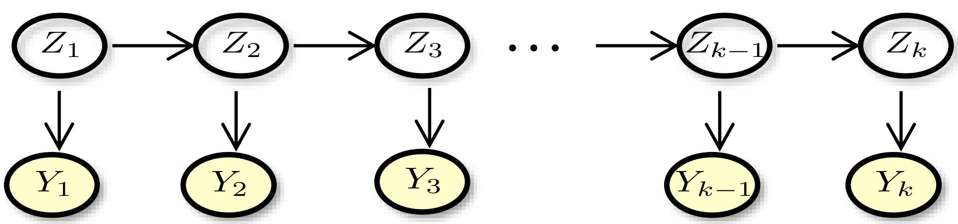
- Latent Variable Model



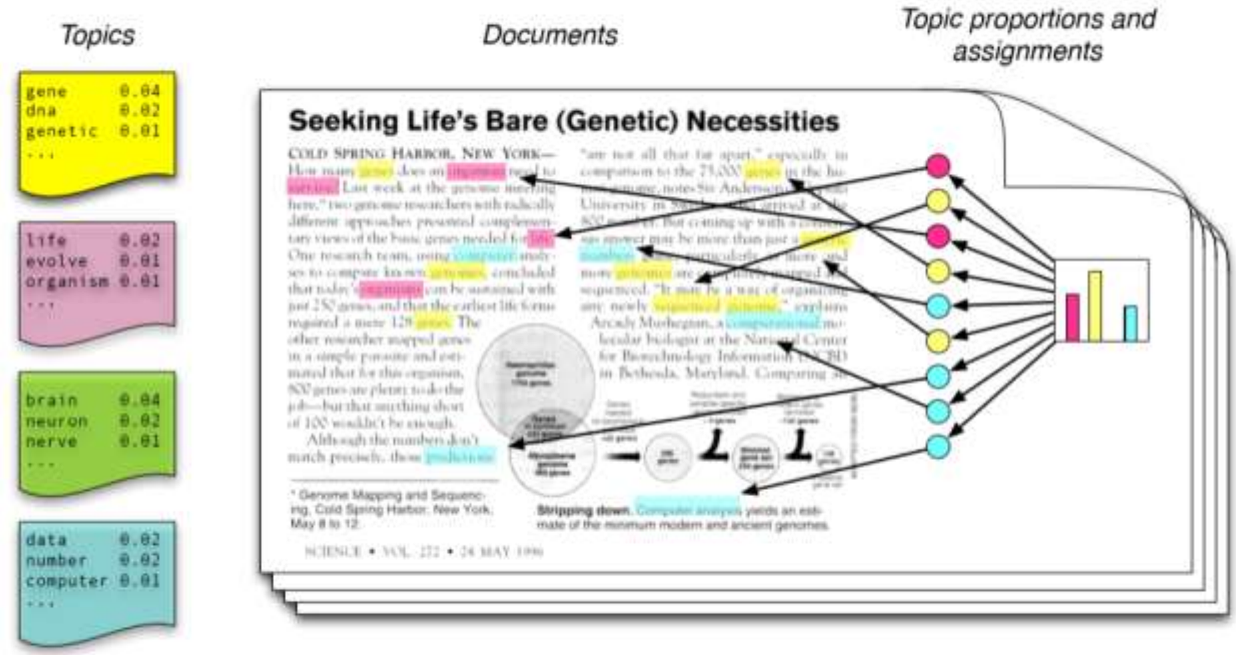


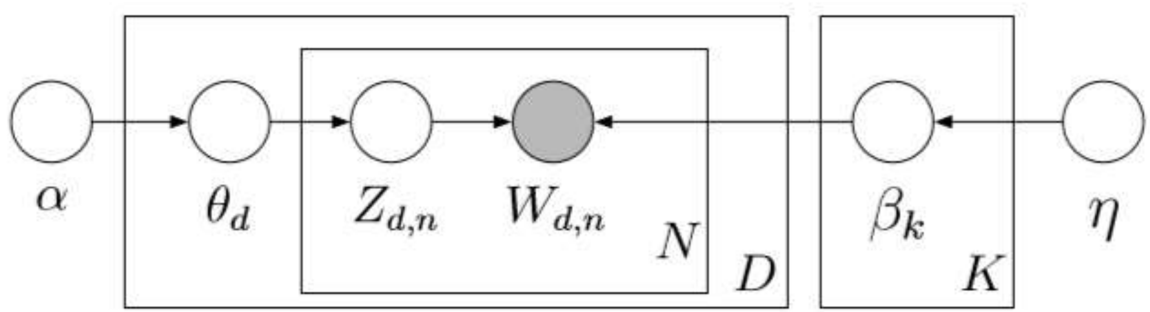


- Filtering



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

http://machinelearning.snu.ac.kr/NIPS2015/NIPS2015_Accepted/nipsnice.html
http://cs.stanford.edu/people/karpathy/nipspreview/

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

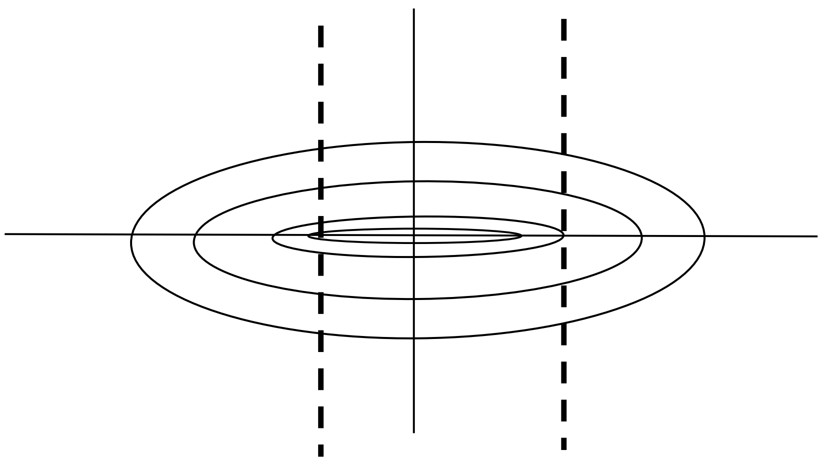
[pdf] [bibtex] [supplementary]
[rank by tf-idf similarity to this]
[abstract]

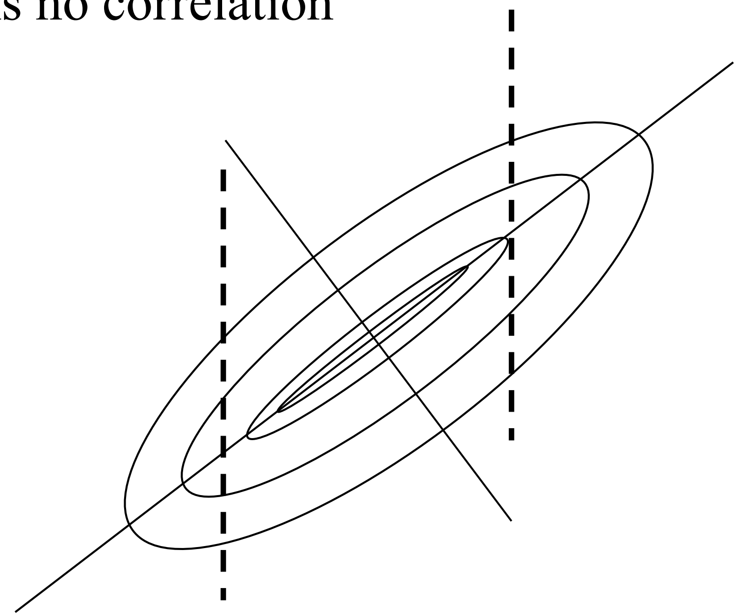
Independency

- Correlation and Independency

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

Independency in Gaussian means no correlation





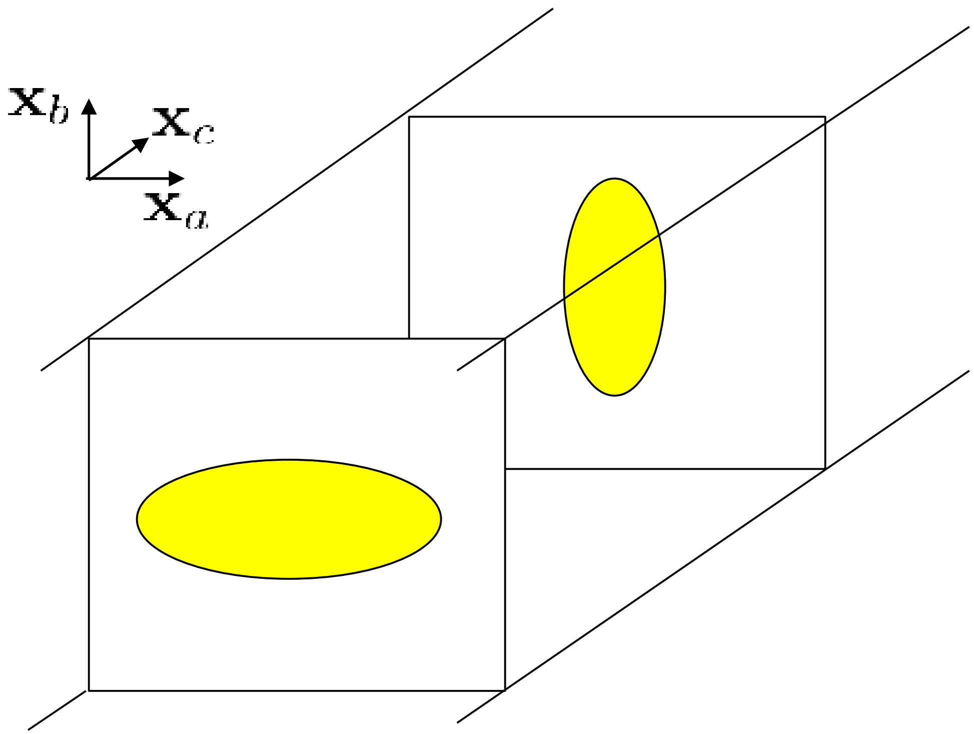
Naïve Bayes?
Mixture of Gaussian?

Conditional Independency

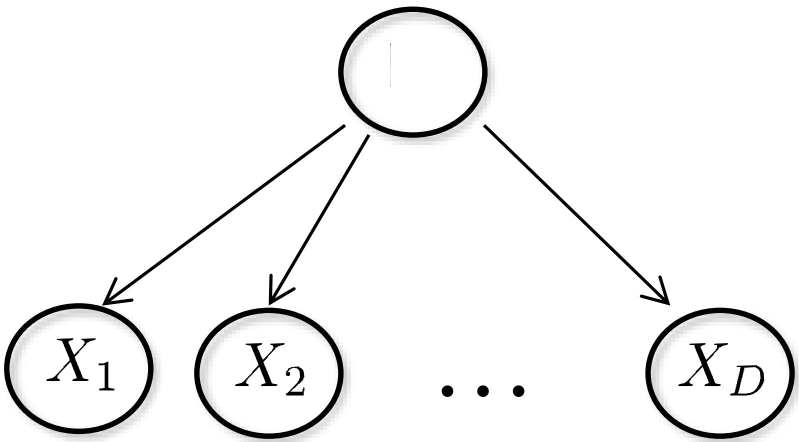
$$p(\mathbf{x}_a, \mathbf{x}_b) = p(\mathbf{x}_a)p(\mathbf{x}_b)$$

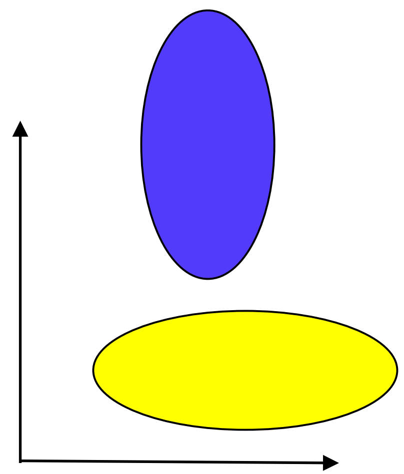
vs.

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



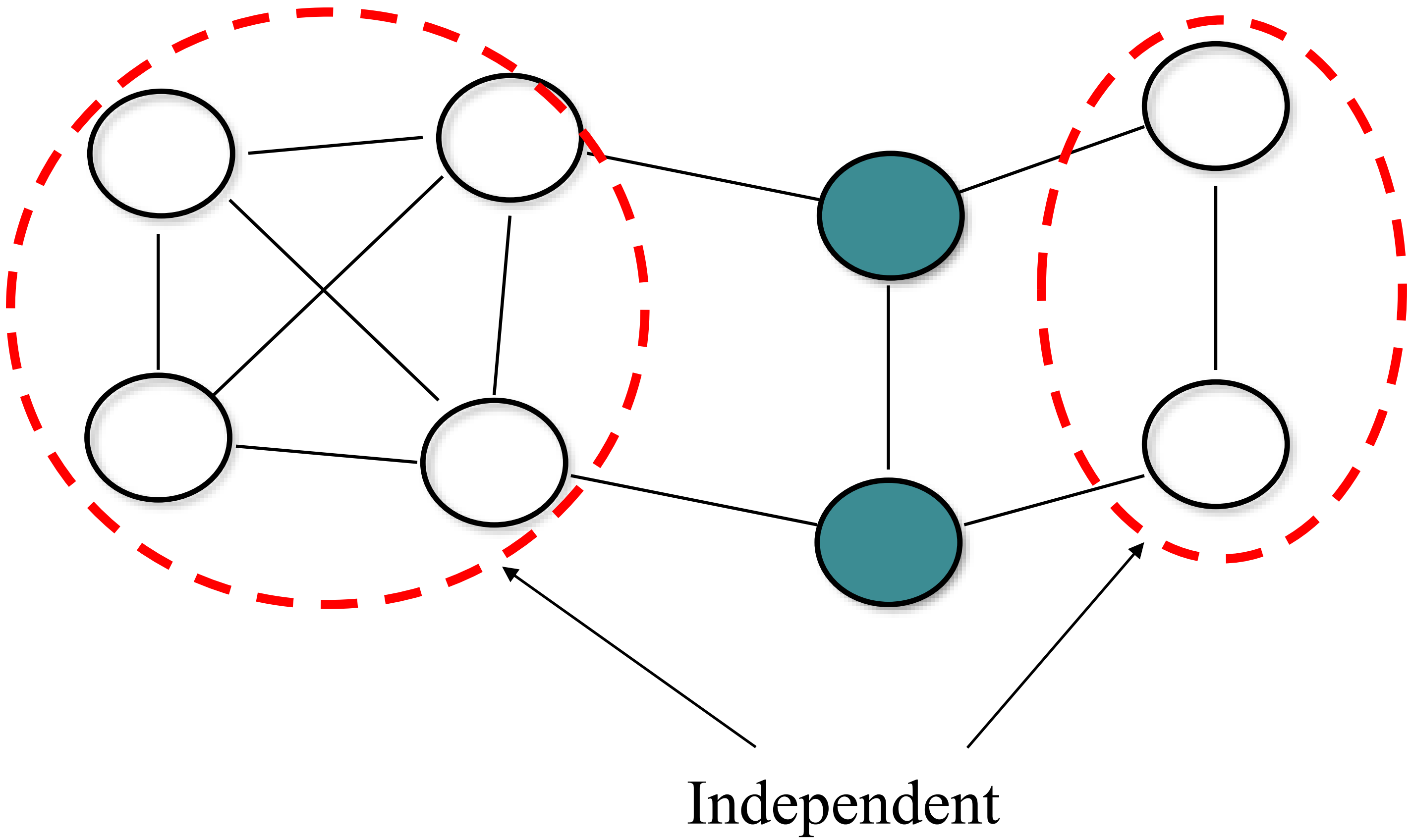
Naïve Bayes for Classification





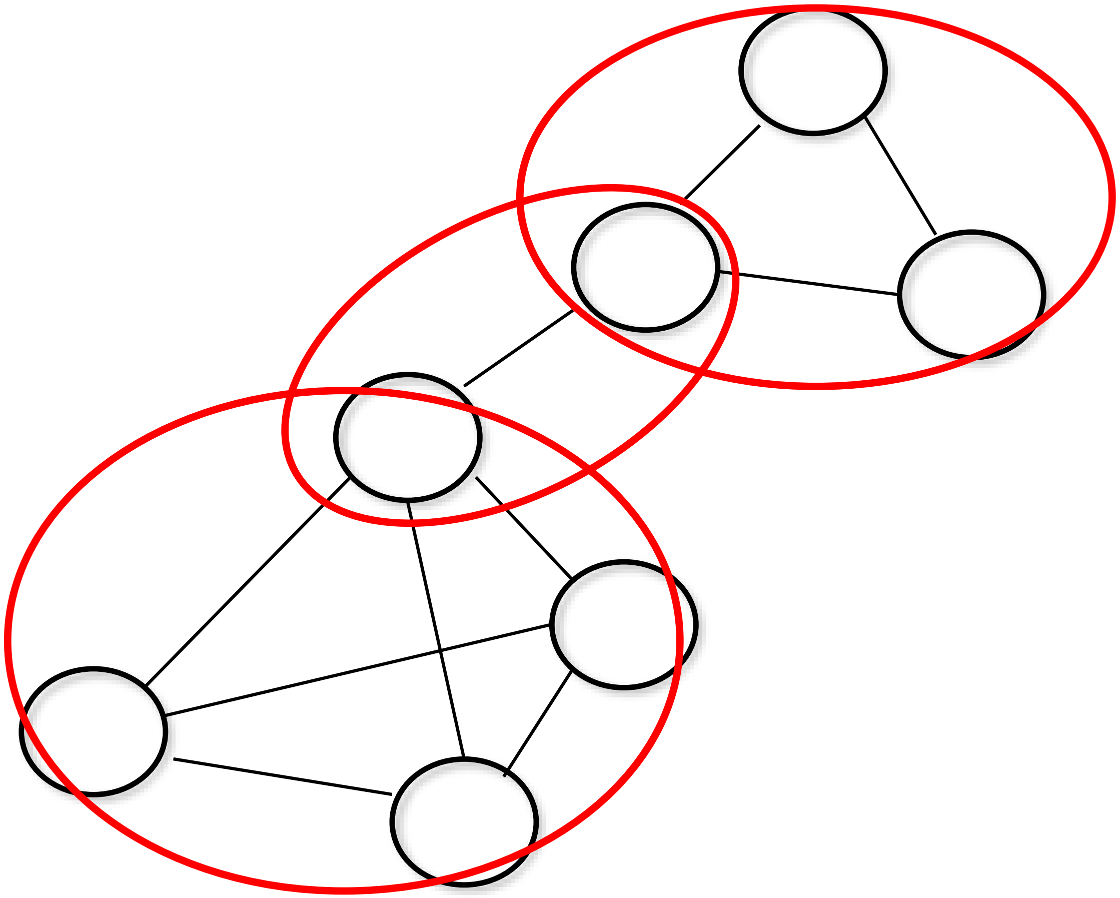
$$P(X,Y) = P(Y)P(X_1|Y)\ldots P(X_D|Y)$$

Undirected Graphical Models



Undirected Graphical Models

- Find all maximal cliques:



Undirected Graphical Models

- Potential functions on cliques

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

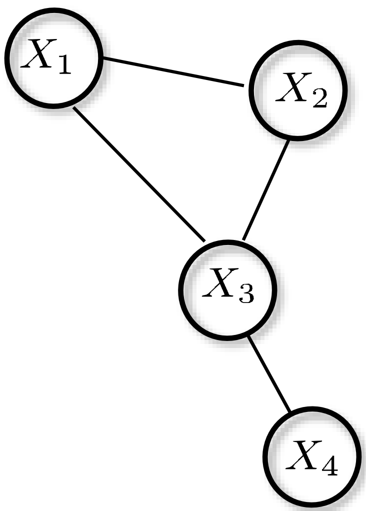
$$P(X) = \frac{1}{Z}\Psi_1(X_1)\Psi_2(X_2)\cdots\Psi_C(X_C)$$

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

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

1	1	1
1	0	0
0	1	0
0	0	3

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

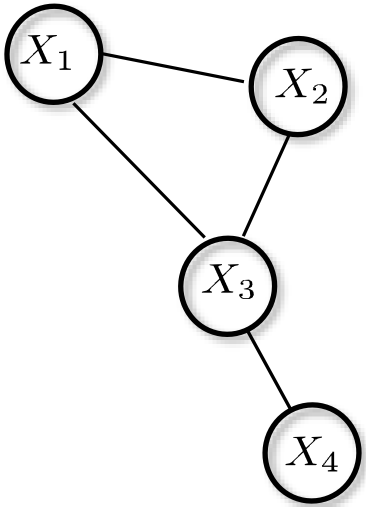
$$= \Psi_{X_1,X_2,X_3}(1,1,1)\Phi_{X_3,X_4}(1,1) + \Psi_{X_1,X_2,X_3}(1,1,1)\Phi_{X_3,X_4}(1,0) + \dots$$

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

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



$$
\begin{array}{ccccl}
X_1 & X_2 & X_3 & & \\
1 & 1 & 1 & = & \Psi_{1,1,1} \\
1 & 1 & 0 & = & \Psi_{1,1,0} \\
1 & 0 & 1 & & \vdots \\
1 & 0 & 0 & & \\
0 & 1 & 1 & & \\
0 & 1 & 0 & & \\
0 & 0 & 1 & & \\
0 & 0 & 0 & &
\end{array}
\qquad
\begin{array}{cccl}
X_3 & X_4 & & \\
1 & 1 & = & \Phi_{1,1} \\
1 & 0 & = & \Phi_{1,0} \\
0 & 1 & & \vdots \\
0 & 0 & &
\end{array}
$$

12 parameters

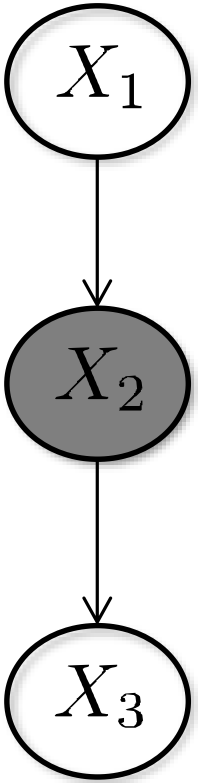
Ψ_{X_1,X_2,X_3} : *8*

Φ_{X_3,X_4} : *4*

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

$$
\begin{array}{ccccccl}
X_1 & X_2 & X_3 & X_4 & & \\
1 & 1 & 1 & 1 & = & P(1,1,1,1) \\
1 & 1 & 1 & 0 & = & P(1,1,1,0) \\
1 & 1 & 0 & 1 & & \vdots \\
 & \cdots & & & &
\end{array}
$$

Conditional Independency



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

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

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

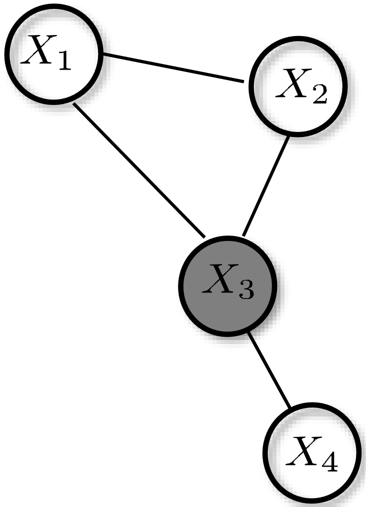
$$= \frac{P(X_1 = 0, X_2 = 1, X_3 = 1)}{P(X_1 = 0, X_2 = 1, X_3 = 1) + P(X_1 = 0, X_2 = 1, X_3 = 0)}$$

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

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

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	$=\Psi_{0,0}\Phi_0$

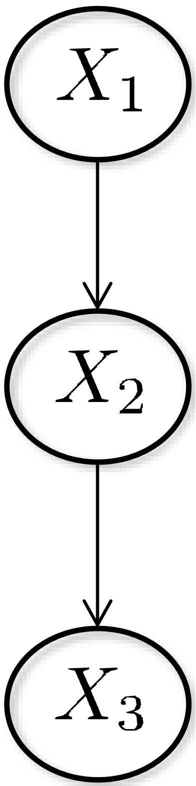
$$P(X_4=1|X_1=0,X_2=1)=?$$

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

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

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

Marginalization

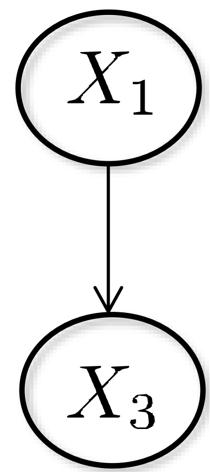


$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2)$$

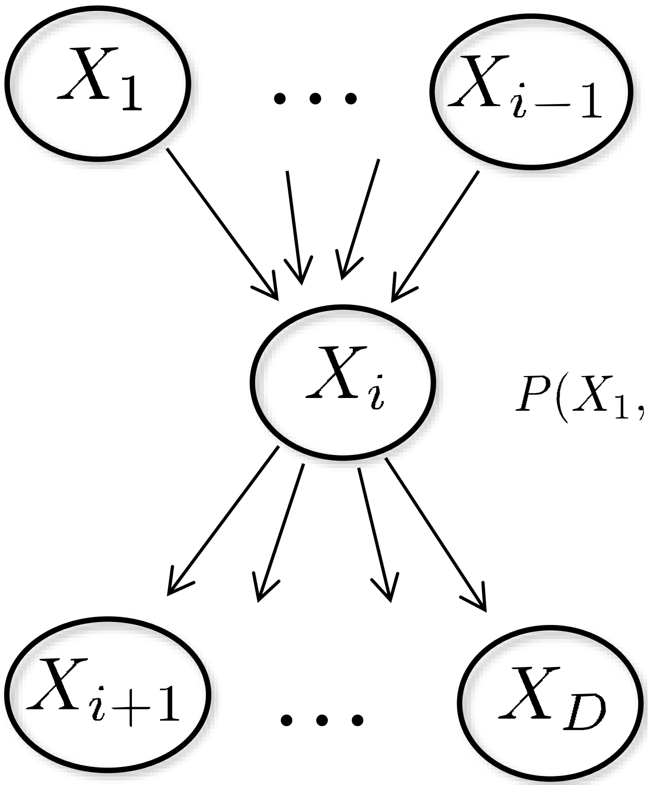
$$P(X_1, X_3) = \int P(X_1, X_2, X_3)dX_2$$

Any good property like

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



Marginalization



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

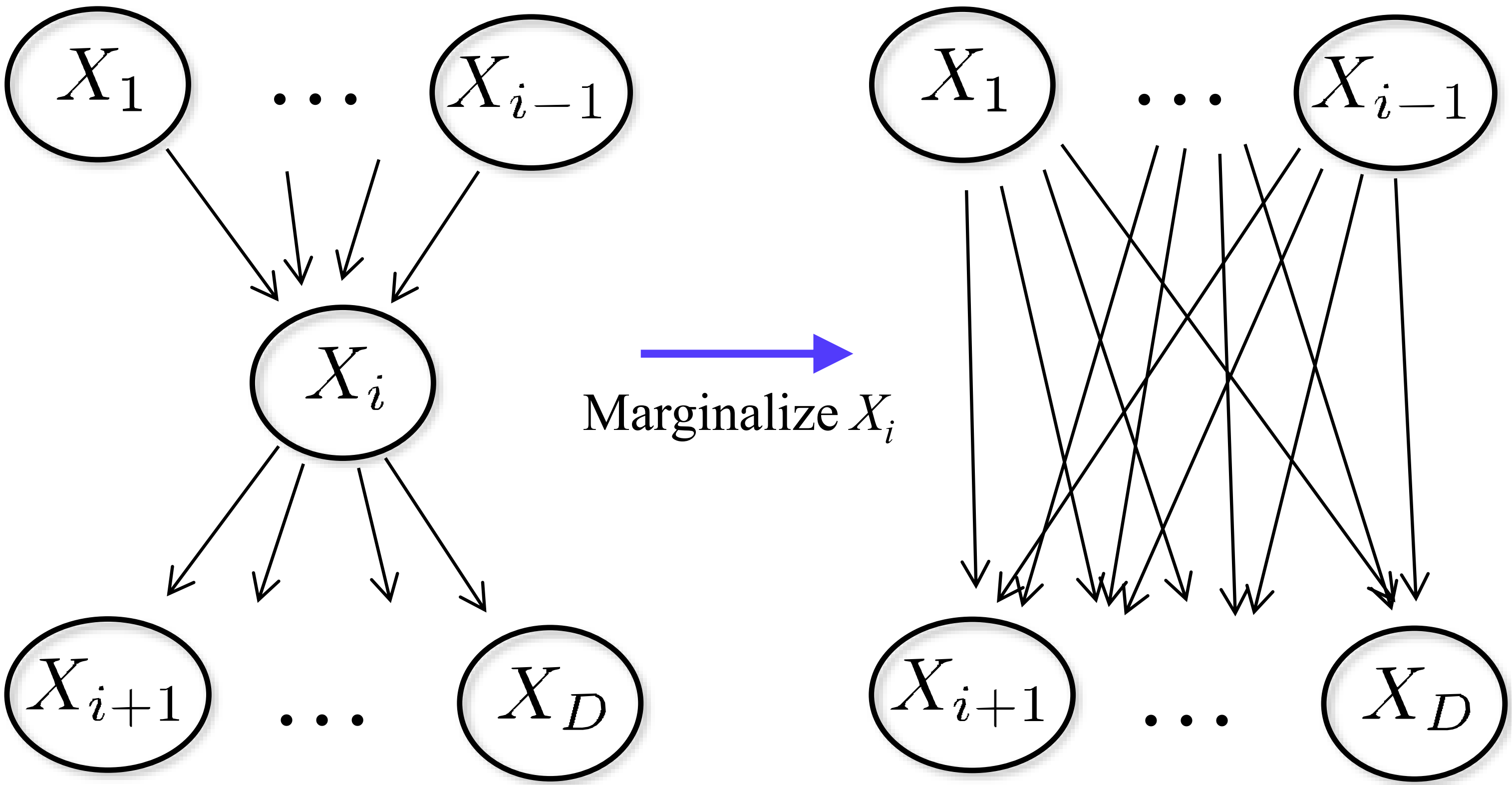
$$P(X_1,\dots,X_{i-1},X_{i+1},\dots,X_D) = \int P(X_1,\dots,X_D)dX_i$$

Any decomposition with

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

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

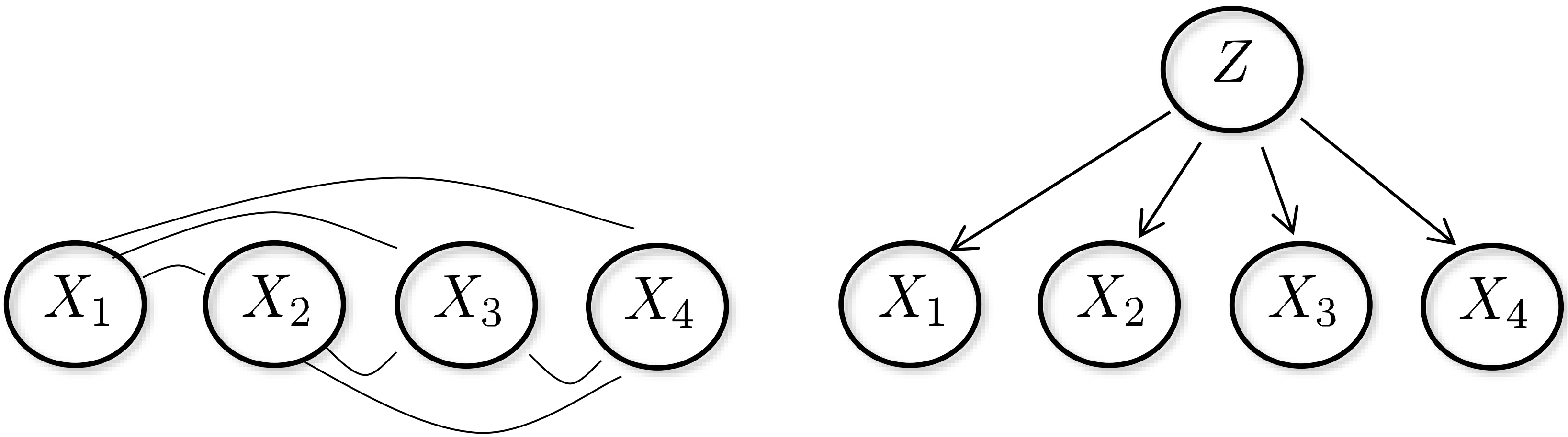
Marginalization



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

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

Introducing Latent Variables



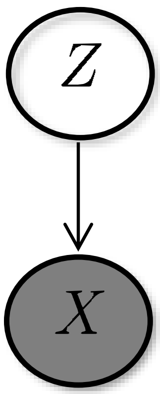
- Issue: how can a model be simplified as much as possible, while the flexibility is kept enough to incorporate the true dependency.

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

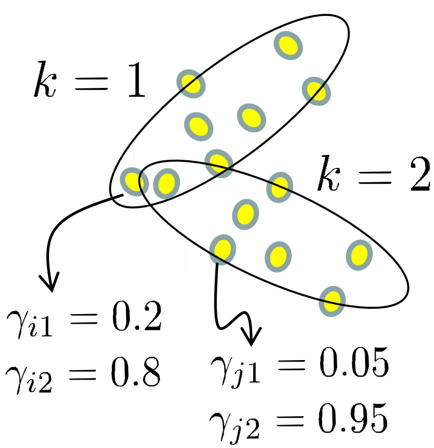
- Gaussian mixture model



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

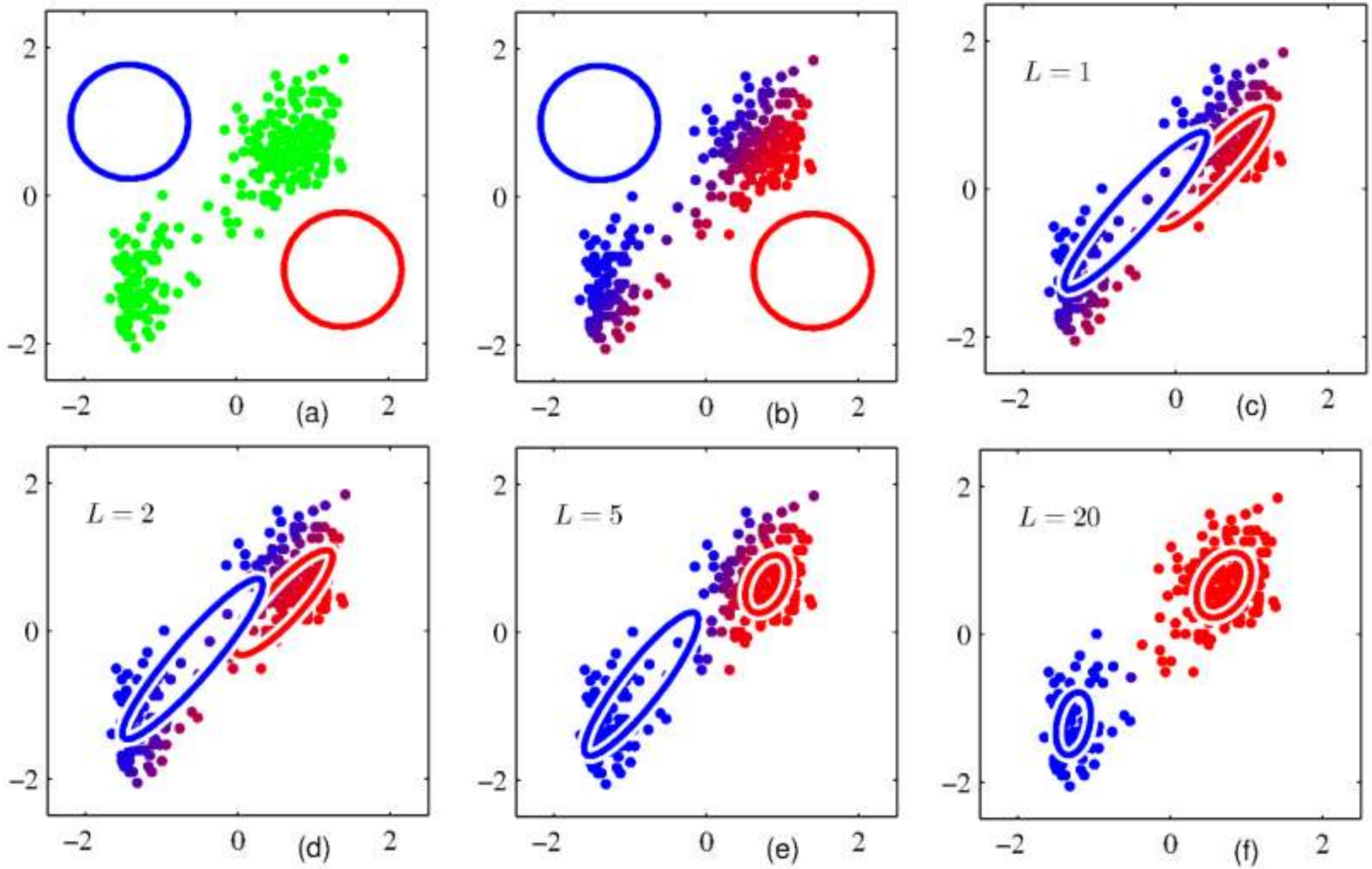
M-step: Estimate parameters.

$$
\left\{
\begin{aligned}
\mu_k &= \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik})\ \mathbf{x}_i \\
\Sigma_k &= \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik})(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top \\
\pi_k &= \frac{N_k}{N}
\end{aligned}
\right.
$$

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)

ANY QUESTIONS?

THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr