Gaussian Models

Congyuan Duan

School of Mathematics
Sun Yat-sen University

- Introduction
- Why Gaussian Assumption?
- MLE for Gaussian

1. Introduction

If $X \sim N(\boldsymbol{\mu}, \Sigma)$, the probability density function (pdf) of X is defined as follow:

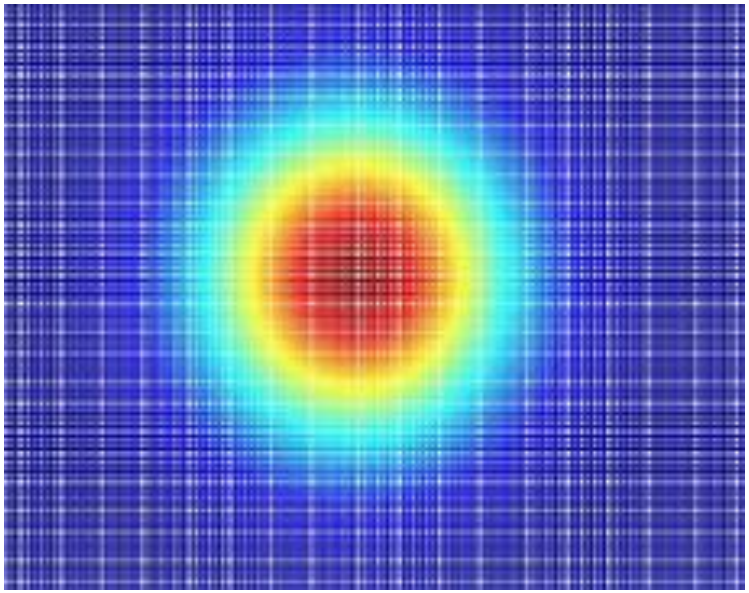
$$\mathbf{p}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\},$$

where $\boldsymbol{\mu}$ is the mean vector of X, Σ is the covariance matrix of X.

1. Introduction

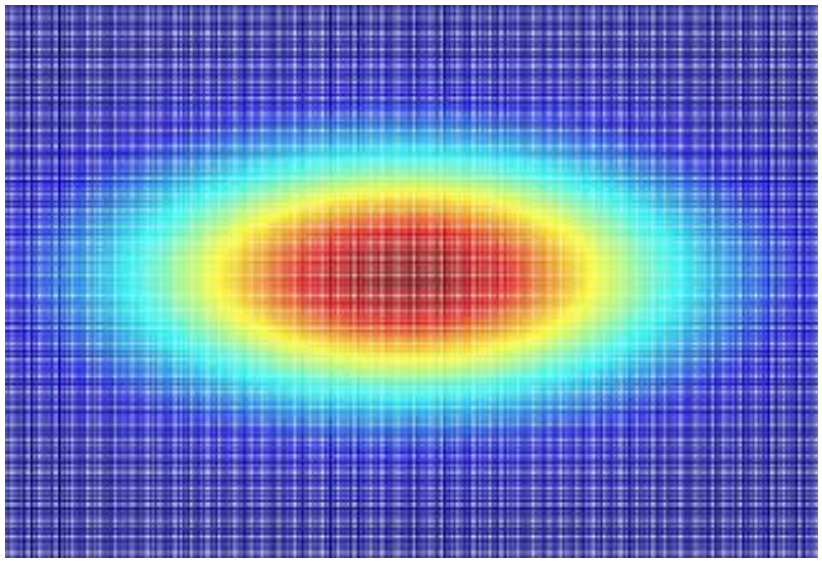
Independent case:

$$X = (X_1, \ldots, X_n)^T, X_i \overset{iid}{\sim} N(0, \sigma^2).$$



Heterogeneous independent case:

$$X = (X_1, \ldots, X_n)^T, X_i \overset{iid}{\sim} N(0, \sigma_i^2).$$



1. Introduction

In general, Σ is a positive definite matrix. According to eigen decomposition:

$$\Sigma = \Gamma^T \Lambda \Gamma ,$$

where $\Lambda = diag(\lambda_1, \dots \lambda_n), \Gamma^T \Gamma = I_n$. Define the square root of Σ as:

$$\Sigma^{\frac{1}{2}} = \Gamma^T \Lambda^{\frac{1}{2}} \Gamma.$$

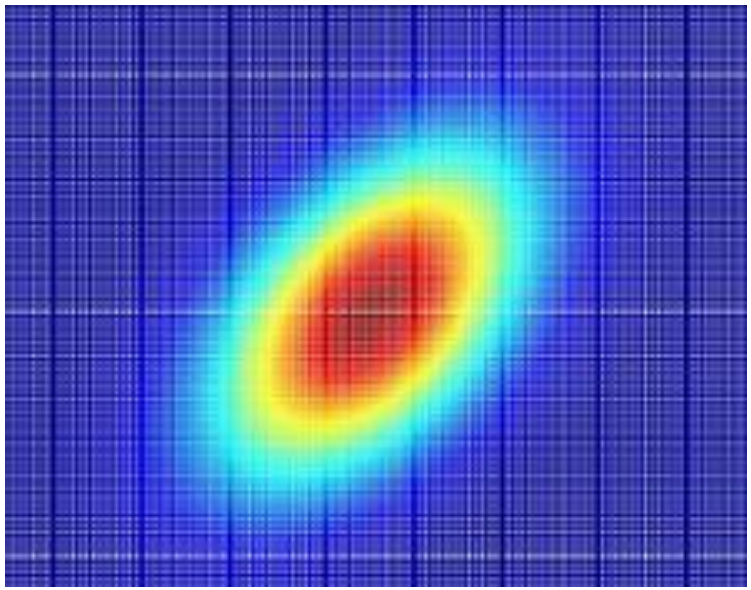
$\Sigma^{\frac{1}{2}}$ is also positive definite.

1. Introduction

Let $\boldsymbol{Z} \sim N(0, I_n)$, define:

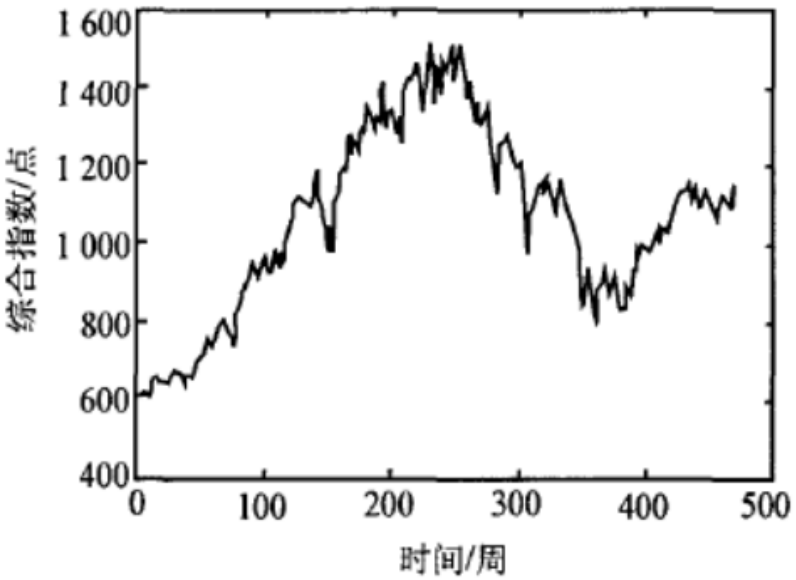
$$\boldsymbol{X} = \Sigma^{\frac{1}{2}} \boldsymbol{Z} + \boldsymbol{\mu},$$

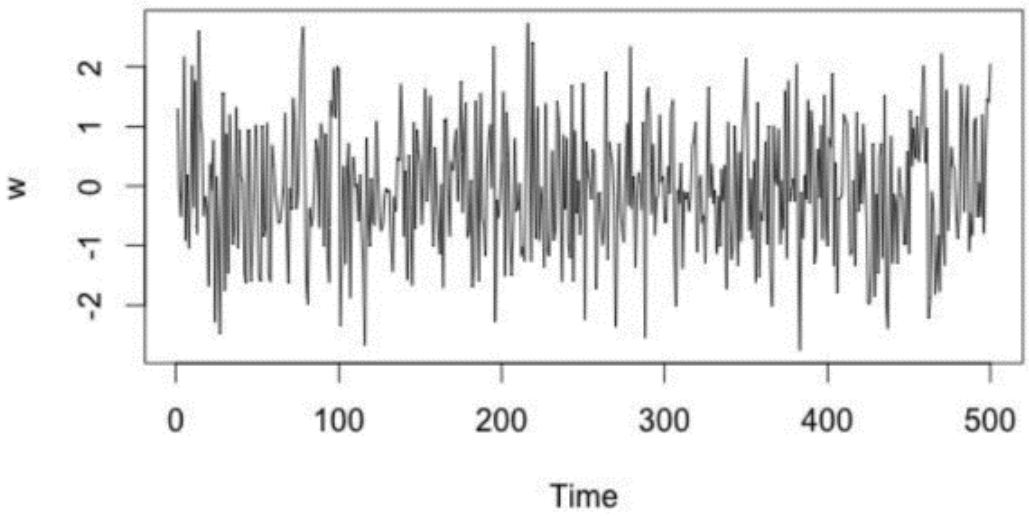
then $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $Cov(\boldsymbol{X}) = \Sigma$.



1. Introduction

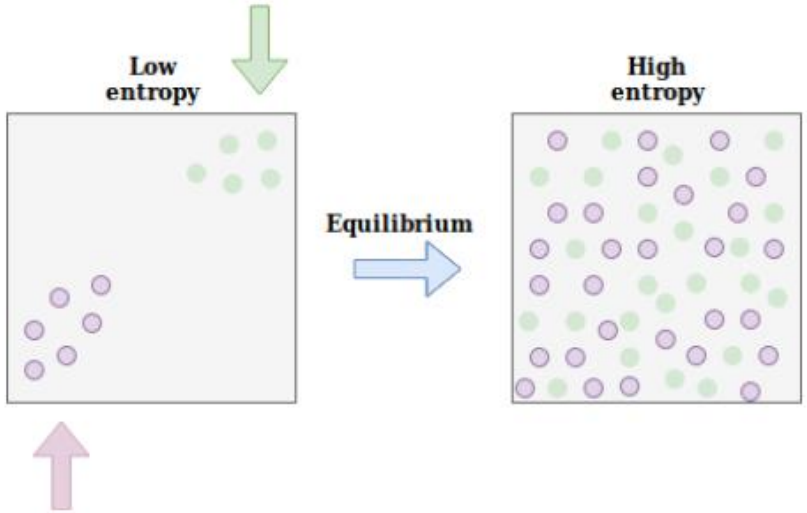
Example:





2. Why Gaussian Assumption?

Shannon Entropy: Measure the level of chaos in a system.



2. Why Gaussian Assumption?

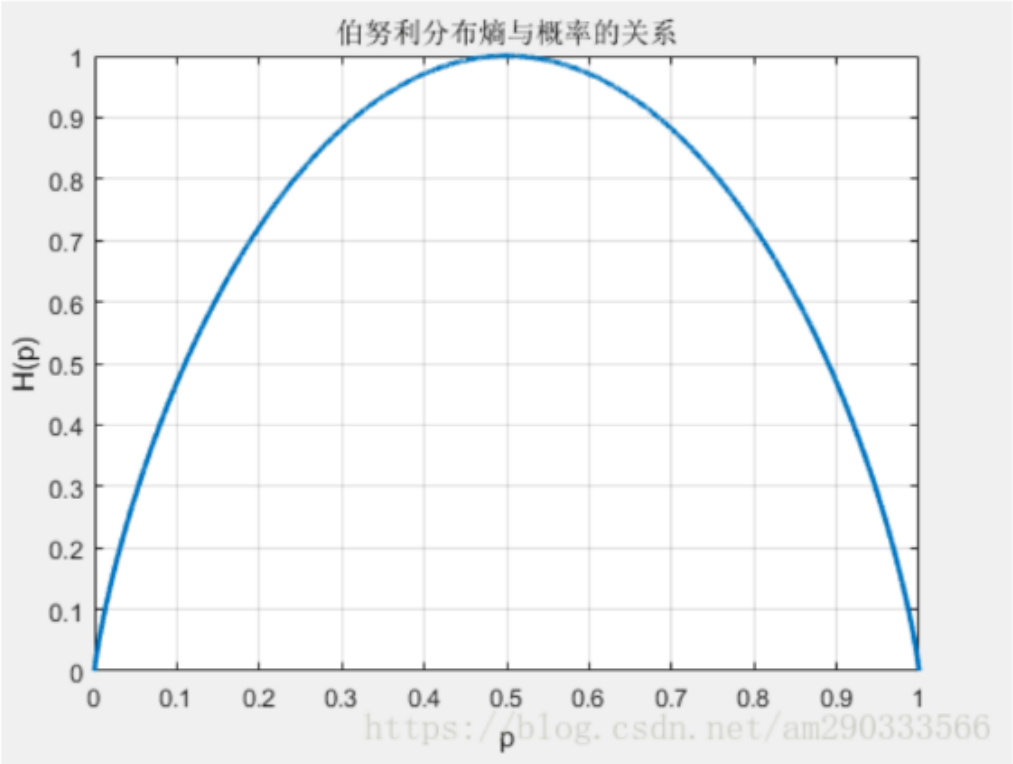
Given a random element X with pdf: $p(x)$, the entropy of X is defined as:

$$H(p) = -\int p(x) \log p(x)\, dx,$$

which is quantified the randomness of X.

2. Why Gaussian Assumption?

Example:



2. Why Gaussian Assumption?

Theorem 4.1.2

Let q($\boldsymbol{x}$) be the pdf of random vector $\boldsymbol{X}$, which has

$$E(\boldsymbol{X}) = \boldsymbol{\mu}, Cov(\boldsymbol{X}) = \Sigma,$$

then $H(q) \leq H(p)$, where $p \sim N(\boldsymbol{\mu}, \Sigma)$,

Proof:

$$\begin{aligned}
0 \leq KL(q\|p) &= \int_{-\infty}^{+\infty} q(x) \log \frac{q(x)}{p(x)} dx \\
&= -h(q) - \int q(x) \log p(x) dx \\
&= -h(q) - \int q(x) \left(\log \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} - \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right) dx \\
&= -h(q) + \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\Sigma| + \frac{1}{2} \int q(x) (x-\mu)^T \Sigma^{-1} (x-\mu) dx \\
&= -h(q) + h(p) \\
\therefore\ & h(p) \geq h(q)
\end{aligned}$$

2. Why Gaussian Assumption?

There are mainly two reason for the choice of Gaussian noise:

- Maximize the randomness of noise if the second moment of noise exists.
- Lindeberg-Feller Central limit theorem.

3. MLE for Gaussian

Theorem 4.2

$X_i \sim N(\boldsymbol{\mu}, \Sigma)$ $i = 1, \ldots, n$ then the MLE for $\boldsymbol{\mu}, \Sigma$ is given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum X_i \,,\; \hat{\Sigma} = \frac{1}{n}\sum (X_i - \overline{X})(X_i - \overline{X})^T$$

Proof: The Likelihood function:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathrm{tr}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{x}\mathbf{x}^T\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}^T)$$

$$L(\mu, V) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left(\sum_{i=1}^{n} -\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathrm{tr}\left(\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left(\mathrm{tr}\left[-\frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)(x_i-\bar{x}+\bar{x}-\mu)^T\right]\right)$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left[\mathrm{tr}\left(-\frac{1}{2}\Sigma^{-1}\left(A + n(\bar{x}-\mu)(\bar{x}-\mu)^T\right)\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{mn}{2}}|\Sigma|^{\frac{n}{2}}} \exp\left[-\mathrm{tr}\left(\frac{1}{2}\Sigma^{-1}A\right) - \frac{n}{2}(\bar{x}-\mu)^T\Sigma^{-1}(x-\mu)^T\right]$$

$$\Rightarrow \hat{\mu} = \bar{X}$$

3. MLE for Gaussian

Lemma: Let B be $m \times m$ positive definite matrix, then

$$tr(B) - \ln|B| \geq m$$

the equality holds iff $B = I_m$

证明

因为$B>0$，所以B的全部特征值$\lambda_1,\ldots,\lambda_m>0$,且

$|B|=\lambda_1\cdots\lambda_m$

利用不等式$\ln(1+x)\leq x$(当$x+1>0$),可得

$$\ln|B| = \sum_{i=1}^{m} \ln \lambda_i = \sum_{i=1}^{m} \ln(1+\lambda_i-1)$$

$$\leq \sum_{i=1}^{m} (\lambda_i - 1) = \mathrm{tr}(B) - m$$

所以 $\mathrm{tr}B - \ln|B| \geq m$

当且仅当λ_i均等于1时等号成立，即$B=I_p$

3. MLE for Gaussian

$$
\begin{aligned}
\ln L(\bar{x}, \Sigma) &= -\frac{nm}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A) \\
&= -\frac{nm}{2}\ln(2\pi) - \frac{n}{2}\left[\ln|\Sigma| + \mathrm{tr}\left(\Sigma^{-1}\frac{A}{n}\right)\right] \\
&= C - \frac{n}{2}\left[\mathrm{tr}\left(\Sigma^{-1}\frac{A}{n}\right) - \ln\left|\Sigma^{-1}\frac{A}{n}\right| + \ln\left|\frac{A}{n}\right|\right] \\
&= C - \frac{n}{2}\left[\mathrm{tr}\left(\Sigma^{-\frac{1}{2}}\frac{A}{n}\Sigma^{-\frac{1}{2}}\right) - \ln\left|\Sigma^{-\frac{1}{2}}\frac{A}{n}\Sigma^{-\frac{1}{2}}\right| + \ln\left|\frac{A}{n}\right|\right] \\
\Rightarrow \hat{\Sigma} &= \frac{A}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T
\end{aligned}
$$

Appendix: Information theory

KL divergence:

$$KL(q||p) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

Conditional Entropy:

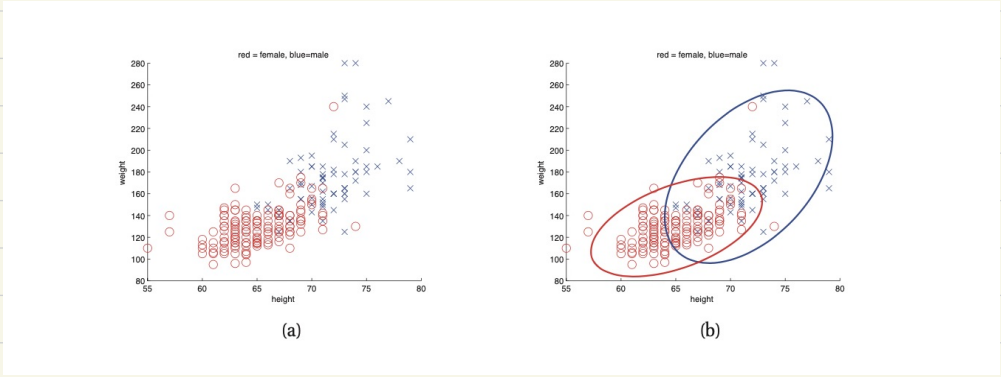
$$H(Y|X) = -\iint p(x,y) \log p(y|x) dxdy\,.$$

Joint Entropy:

$$H(X,Y) = -\iint p(x,y) \log p(x,y) dxdy.$$

Part II: Gaussian discriminant analysis

Classification problem: Given X, we want to know $Y=0$ or $Y=1$.



discriminative learning v.s. generative learning

Find $P(Y|X)$ — Find joint $P(X,Y)$, $P(X=0|Y)$, $P(X=1|Y)$

$f: X \to \{0,1\}$ — priors $P(Y) \Rightarrow P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

Remark: $\frac{P(Y=0|X)}{P(Y=1|X)} = \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}$ Bayes classifier.

- Linear discriminant analysis

Assumption: $X|Y=0 \sim N(\mu_0, \Sigma)$ same covariance matrix

$X|Y=1 \sim N(\mu_1, \Sigma)$ $Y \sim Ber(p)$

$$P(Y=c|X) \propto \pi_c \exp\left[\mu_c^T \Sigma^{-1} X - \frac{1}{2} X^T \Sigma^{-1} X - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c\right]$$

$$\propto \exp\left[\mu_c^T \Sigma^{-1} X - \frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c\right] \exp\left[-\frac{1}{2} X^T \Sigma^{-1} X\right]$$

Let $\beta_c = \Sigma^{-1} \mu_c$ $\gamma_c = -\frac{1}{2} \mu_c^T \Sigma^{-1} \mu_c + \log \pi_c$

$$P(Y=c|X) = \frac{\exp(\beta_1^T X + \gamma_1)}{\exp(\beta_0^T X + \gamma_0) + \exp(\beta_1^T X + \gamma_1)} = sigm((\beta_1 - \beta_0)^T X + (\gamma_1 - \gamma_0))$$

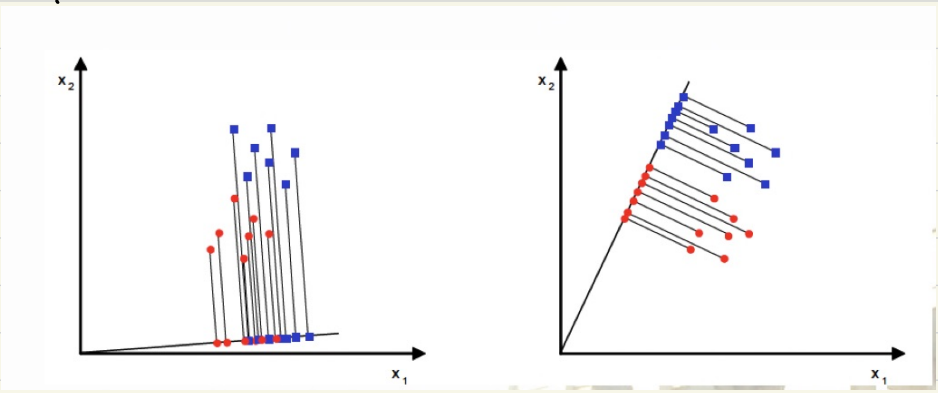
Let $w = \beta_1 - \beta_0$, $x_0 = \frac{1}{2}(\mu_1 + \mu_0) - (\mu_1 - \mu_0) \frac{\log(\pi_1/\pi_0)}{(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0)}$

$P(Y=c|X) = sigm(w^T(x - x_0))$, if $w^T(x - x_0) \Rightarrow Y=1$

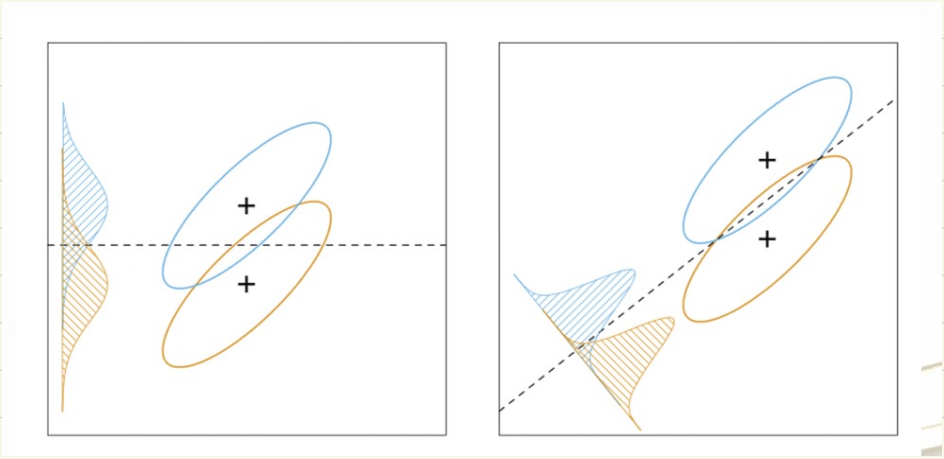
Remark: $w^T(x - x_0) > 0 \iff \log \frac{P(X|Y=1)}{P(X|Y=0)} > 0 \iff \frac{P(X|Y=1)}{P(X|Y=0)} > 1$

Projection and LDA

- separable



- inseparable



LDA is equal to:

$$\max_a \frac{(E(a^T X_1) - E(a^T X_2))^2}{\mathrm{Var}(a^T X_1 - a^T X_2)} \Rightarrow a \propto \Sigma^{-1}(\mu_1 - \mu_2)$$

- Quadratic discriminant analysis

Assumption: $X|Y=0 \sim N(\mu_0, \Sigma_0)$ $\quad \Sigma_0 \neq \Sigma_1$

$X|Y=1 \sim N(\mu_1, \Sigma_1)$

$$P(Y=c|X) = \frac{\pi_c |2\pi\Sigma_c|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)]}{\sum_{c=0}^{1} \pi_c |2\pi\Sigma_c|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)]}$$

$$\log P(Y=c|X) \propto -\frac{1}{2}\log|\Sigma_c| - \frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c) + \log \pi_c$$ (Discriminant function)

$$\log \frac{P(Y=1|X)}{P(Y=0|X)} = \underbrace{\{\frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|} + \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0\}}_{\beta_0} + \underbrace{\{\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0\}^T}_{\beta_1} X - \frac{1}{2}X^T\underbrace{(\Sigma_1^{-1} - \Sigma_0^{-1})}_{\Omega}X$$

$$= \beta_0 + \beta_1^T X + X^T \Omega X$$

If $\beta_0 + \beta_1^T X + X^T \Omega X > 0 \Longrightarrow Y=1$

Remark: In LDA, discriminant function is linear function.
In QDA, discriminant function is quadratic function.

- How to estimate parameters in discriminant analysis?

Solution: use MLE

$$\begin{cases} \hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} x_i \\ \hat{\Sigma}_c = \frac{1}{N_c} \sum_{i:y_i=c} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T \end{cases}$$

In high dimensions, the MLE may badly overfit. ($N_c < D$)

Solution: ① Assume Σ_c is diagonal $\Longrightarrow$ naive Bayes
② Assume $\Sigma_c = \Sigma$.
③ Projection.
④ MAP estimation.

- Diagonal LDA

In RDA, let $\lambda = 1 \Rightarrow$ diagonal LDA

$$\log P(Y=C|X) = -\sum_{j=1}^{p} \frac{(x_j - \mu_{cj})^2}{2\sigma_j^2} + \log \pi_c$$

$$\hat{\mu}_{cj} = \bar{x}_{cj}, \quad \hat{\sigma}_j^2 = \frac{\sum_{c=1}^{C} \sum_{i:y_i=c} (x_{ij} - \bar{x}_{cj})^2}{N-C}$$

Remark: diagonal LDA is a special case of naive Bayes.

Naive Bayes classifier

Assumption: Give a class $Y=j$, features X_k are independent

$$P(X|Y=j) = \prod_{k=1}^{p} P(X_k|Y=j)$$

$$\log \frac{P(Y=j|X)}{P(Y=J|X)} = \log \frac{\pi_j P(X|Y=j)}{\pi_J P(X|Y=J)}$$

$$= \log \frac{\pi_j \prod_{k=1}^{p} P(X_k|Y=j)}{\pi_J \prod_{k=1}^{p} P(X_k|Y=J)}$$

$$= \log \frac{\pi_j}{\pi_J} + \sum_{k=1}^{p} \log \left(\frac{P(X_k|Y=j)}{P(X_k|Y=j)} \right)$$

$$= \alpha_j + \sum_{k=1}^{p} \alpha_{jk}(X_k)$$ generalized additive model

- Nearest shrunken centroids classifier (NSCC)

Motivation: diagonal LDA depends on all of the features.

In high-dimension problem. we only want to a subset.

We let some $\bar{x}_{cj} = \bar{x}_j$, then drop gene j in classification.

centroid

Method:

① Shrinking the classwise mean toward the overall mean.

$$d_{cj} = \frac{\bar{x}_{cj} - \bar{x}_j}{m_c (s_j + s_0)}$$

where $m_c^2 = \frac{1}{N_c} - \frac{1}{N}$, $s_0 > 0$ is a small constant(median(s_j))

② use soft thresholding

$$d'_{cj} = \text{sign}(d_{cj})(|d_{cj}| - \Delta)_+$$

Δ is a parameter to be determined (may use CV)

Remark: hard thresholding: $d'_{cj} = d_{cj} I(|d_{cj}| > \Delta)$

③ obtain the shrunken version of $\bar{x}_{cj}'$

$$\bar{x}_{cj}' = \bar{x}_j + m_k(s_j + s_0)d_{cj}'$$

Remark: $\bar{x}_{cj}'$ is also a lasso-style estimator.

④ use $\bar{x}_{cj}'$ in discriminant score

$$\log P(Y=c|X) = -\sum_{j=1}^{p} \frac{(x_j - \mu_{cj})^2}{2\sigma_j^2} + \log \pi_c$$

$$\hat{\mu}_{cj} = \bar{x}_{cj}', \quad \hat{\sigma}_j^2 = \frac{\sum_{c=1}^{C}\sum_{i:y_i=c}(x_{ij} - \bar{x}_{cj}')^2}{N-C}$$

Remark: If $d_{cj}' = 0$, then $\bar{x}_{cj}' = \bar{x}_j$, gene j don't play a role in classification.

- LDA vs logistic regression

LDA: $\log \frac{P(Y=1|X)}{P(Y=0|X)} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) + X^T\Sigma^{-1}(\mu_1-\mu_0)$

$= \alpha_{10} + \alpha_1^T X$

Logistic: $\log \frac{P(Y=1|X)}{P(Y=0|X)} = \beta_{10} + \beta_1^T X$

① LDA need specify joint distribution, but logistic not.

② $P(X, Y=K) = P(X)\,P(Y=K|X)$

Both LDA and logistic have the logit-linear form.

③ logistic treats $P(X)$ as an arbitrary density function.

LDA $P(X, Y=K) = N(X|\mu_K, \Sigma)\pi_K$

$P(X) = \sum_{K=0}^{1} N(X|\mu_K, \Sigma)\pi_K$

④ logistic fits $P(Y|X)$ by maximizing conditional likelihood.

LDA fits $P(Y|X)$ by MLE.

⑤ logistic: outlier are down-weighted.

LDA: no robust.

Part III : Evaluation in jointly Gaussian distributions

Given a joint distribution $P(X_1, X_2)$, we consider marginals $P(X_1)$ and conditionals $P(X_1|X_2)$.

Thm 4.3.1 Suppose $x = (X_1, X_2)$ is jointly Gaussian with parameters

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

Then,

$$P(X_1) = \mathcal{N}(X_1 | \mu_1, \Sigma_{11})$$
$$P(X_2) = \mathcal{N}(X_2 | \mu_2, \Sigma_{22})$$
$$P(X_1|X_2) = \mathcal{N}(X_1 | \mu_{1|2}, \Sigma_{1|2})$$

where $\mu_{1|2} = \Sigma_{1|2}(\Lambda_{11}\mu_1 - \Lambda_{12}(X_2 - \mu_2))$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}$$

Proof : ① Linear transform

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} I_p & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_q \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} = BX$$

Z is MVN.

$$E(Z) = BE(X) = \begin{pmatrix} \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mu_2 \end{pmatrix}$$

$$V(Z) = BV(X)B^T = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

② $Z \sim \mathcal{N}(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \mathcal{N}(\mu_2, \Sigma_{22})$

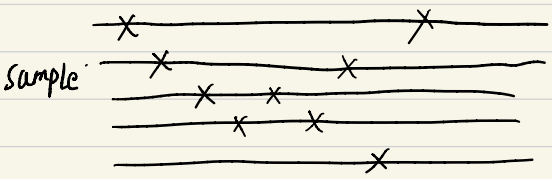
$$\begin{cases} Z_1 = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ Z_2 = X_2 \end{cases}$$

$$X_1|X_2 \sim \mathcal{N}(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Some example

- Data imputation

Aim: Predict the missing entries



① Compute $P(X_{hi}|X_{vi}, \theta)$, h_i: hidden, v_i: visible.

② compute the marginal : $P(X_{hij}|X_{vi}, \theta)$

③ $\hat{X_{ij}} = E[X_j|X_{vi}, \theta]$

Part IV : Linear Gaussian systems

$X \in \mathbb{R}^{D_x}$ hidden variable , $Y \in \mathbb{R}^{D_y}$ noise observation of X

Assumption: $P(X) = \mathcal{N}(X \mid \mu_x, \Sigma_x)$

$P(Y|X) = \mathcal{N}(Y \mid AX+b, \Sigma_y)$

Aim : infer X from Y.

Thm 4.4.1 $P(X|Y) = \mathcal{N}(X \mid \mu_{x|y}, \Sigma_{x|y})$, $\Sigma_{x|y}^{-1} = \Sigma_x^{-1} + A^T \Sigma_y^{-1} A$

$\mu_{x|y} = \Sigma_{x|y}[A^T \Sigma_y^{-1}(y-b) + \Sigma_x^{-1}\mu_x)$

$P(Y) = \mathcal{N}(Y \mid A\mu_x + b, \Sigma_y + A\Sigma_x A^T)$

Proof: ① compute $P(X|Y) = P(X)P(Y|X)$

$\log P(Y|X) \propto -\frac{1}{2}(X-\mu_x)^T \Sigma_x^{-1}(x-\mu_x) - \frac{1}{2}(y-Ax-b)^T \Sigma_y^{-1}(y-Ax-b) \Rightarrow$ gaussian

Ignore linear and constant terms:

$$Q = -\frac{1}{2}x^T \Sigma_x^{-1} x - \frac{1}{2} y^T \Sigma_y^{-1} y - \frac{1}{2}(Ax)^T \Sigma_y^{-1}(Ax) + y^T \Sigma_y^{-1} Ax$$

$$= -\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Sigma_x^{-1} + A^T \Sigma_y^{-1} A & -A^T \Sigma_y^{-1} \\ -\Sigma_y^{-1} A & \Sigma_y^{-1} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

$$= -\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} x \\ y \end{pmatrix}$$

② use Thm 4.3.1 , let $\Sigma^{-1} = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{yx} & \Lambda_{yy} \end{pmatrix}$

$P(X|Y) = \mathcal{N}(\mu_{x|y}, \Sigma_{x|y})$

$\Sigma_{x|y} = \Lambda_{xx}^{-1} = (\Sigma_x^{-1} + A^T \Sigma_y^{-1} A)^{-1}$

$\mu_{x|y} = \Sigma_{x|y}(\Sigma_x^{-1}\mu + A^T \Sigma_y^{-1}(y-b))$

Example: $\mu \to X \to Y$

- Inferring an unknown scalar from noisy measurements.

Assumption : $P(y_i|X) = \mathcal{N}(y_i \mid X, \lambda_y^{-1})$, $\lambda_y = \frac{1}{\sigma^2}$

$P(X) = \mathcal{N}(X \mid \mu_0, \lambda_0^{-1})$

$P(X|Y) = \mathcal{N}(X \mid \mu_N, \lambda_N^{-1})$

where $\lambda_N = \lambda_0 + N\lambda_y$ $\mu_N = \frac{N\lambda_y \bar{y} + \lambda_0 \mu_0}{\lambda_N} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\bar{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0$ (convex combination)

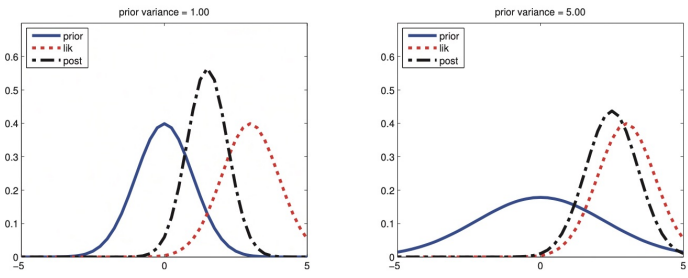


Figure 4.12 Inference about x given a noisy observation $y = 3$. (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0. (a) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE. Figure generated by `gaussInferParamsMean1d`.

$$p(x|D, \sigma^2) = \mathcal{N}(x|\mu_N, \tau_N^2)$$

$$\tau_N^2 = \frac{1}{\lambda_N} = \frac{1}{\lambda_0 + N\lambda_y} = \frac{1}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}} = \frac{\sigma^2\tau_0^2}{N\tau_0^2 + \sigma^2}, \text{ where } \tau_0^2 = \frac{1}{\lambda_0}$$

$$\mu_N = \frac{\sigma^2}{N\tau_0^2 + \sigma^2}\mu_0 + \frac{N\tau_0^2}{N\tau_0^2 + \sigma^2}\bar{y} \qquad \tau_N^2 = \frac{1}{\lambda_N}$$

Special case $N=1$: (define $\Sigma_y = \sigma^2$, $\Sigma_0 = \tau_0^2$, $\Sigma_1 = \tau_1^2$)

$$p(x|y) = \mathcal{N}(x|\mu_1, \Sigma_1)$$

$$\Sigma_1 = \frac{\Sigma_y\Sigma_0}{\Sigma_0 + \Sigma_y} = \left(\frac{1}{\Sigma_0} + \frac{1}{\Sigma_y}\right)^{-1}$$

$$\mu_1 = \Sigma_1\left(\frac{\mu_0}{\Sigma_0} + \frac{y}{\Sigma_y}\right)$$

$$= \frac{\Sigma_y}{\Sigma_y + \Sigma_0}\mu_0 + \frac{\Sigma_0}{\Sigma_y + \Sigma_0}y \quad \text{convex combination}$$

$$= \mu_0 + (y - \mu_0)\frac{\Sigma_0}{\Sigma_y + \Sigma_0} \quad \text{prior mean adjusted toward data}$$

$$= y - (y - \mu_0)\frac{\Sigma_y}{\Sigma_y + \Sigma_0} \quad \text{shrinkage}$$

signal-to-noise ratio $\frac{Ex^2}{E\varepsilon^2} = \frac{\Sigma_0 + \mu_0^2}{\Sigma_y}$

- Inferring an unknown vector from noisy measurements

observation: $y_i \sim \mathcal{N}(x, \Sigma_y)$, Gaussian prior: $x \sim \mathcal{N}(\mu_0, \Sigma_0)$

Aim: Find x. (y may has multiple observations with different Σ)

$$p(x|y_1, \cdots, y_N) = \mathcal{N}(x|\mu_N, \Sigma_N)$$

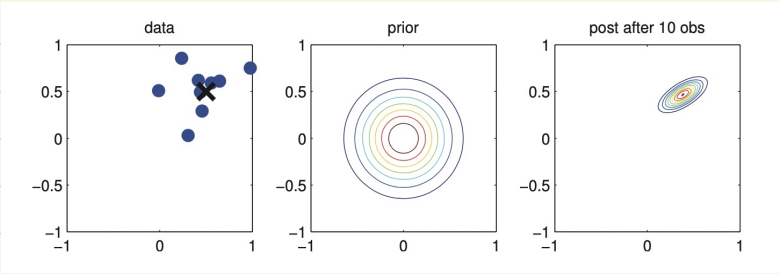
$$\mu_N = \Sigma_N(\Sigma_y^{-1}(N\bar{y}) + \Sigma_0^{-1}\mu_0)$$

$$\Sigma_N^{-1} = \Sigma_0^{-1} + N\Sigma_y^{-1}$$

Simple case:

$y_i \sim N(X, \Sigma_y)$

Prior: $N(X|0, 0.1I)$



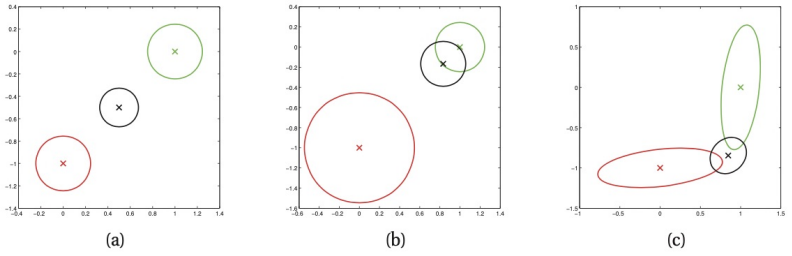
$y_1 = (0, -1)$, $y_2 = (1, 0)$

(a) $\Sigma_{y_1} = \Sigma_{y_2} = 0.01 I_2$

(b) $\Sigma_{y_1} = 0.05 I_2$ $\Sigma_{y_2} = 0.01 I_2$

(c) $\Sigma_{y_1} = 0.01 \begin{pmatrix} 10 \\ 1 \end{pmatrix}$

$\Sigma_{y_2} = 0.01 \begin{pmatrix} 1 \\ 110 \end{pmatrix}$

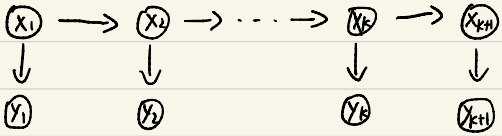


- An example: Kalman Filter

① Model:

$X_k = F X_{k-1} + B u_{k-1} + w_{k-1}$

$Y_k = H X_k + v_k$



where F is the state transition matrix,

u_k is control vector,

$w_{k-1} \sim N(0, Q)$ noise

H is measurement matrix.

$v_k \sim N(0, R)$ noise

We need to estimate X_k given $X_0, Y_1, \cdots, Y_k$

② Algorithm

$E(X_n | Y_1, \cdots, Y_n) \xrightarrow{\text{Prediction}} E(X_{n+1} | Y_1, \cdots, Y_n) \xrightarrow{\text{correction}} E(X_{n+1} | Y_1, \cdots, Y_{n+1})$

Prediction:

(a) state estimate $\hat{X}_k^- = F \hat{X}_{k-1}^+ + B u_{k-1}$

(b) error covariance $P_k^- = F P_{k-1}^+ F^T + Q$

update:

(a) measurement residual: $\tilde{y}_k = Y_k - H \hat{X}_k^-$

(b) Kalman gain: $K_k = P_k^- H^T (R + H P_k^- H^T)^{-1}$

(c) state estimate: $\hat{X}_k^+ = \hat{X}_k^- + K_k \tilde{y}_k$

(d) error covariance: $P_k^+ = (I - K_k H) P_k^-$

(-: predicted (prior) estimate ; +: updated (posterior) estimate)

③ Application

A robot.

$X_k = [P, V]$ P: position, V: velocity

$F = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}$ $P_k = P_{k-1} + \Delta t V_{k-1}$

$V_k = V_{k-1}$

u_k: acceleration

$B = \begin{bmatrix} \frac{\Delta t^2}{2} \\ \Delta t \end{bmatrix}$ $P_k = P_{k-1} + \Delta t V_{k-1} + \frac{1}{2} a \Delta t^2$

$V_k = V_{k-1} + a \Delta t$

$P_k = F_k P_{k-1} F_k^T + Q_k$