Lecture 03: Multivariate Probability

Introduction to Machine Learning

Sajjad Amini

Department of Electrical Engineering
Sharif University of Technology

Contents

References

Except explicitly cited, the reference for the material in slides is:

- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT press.

Section 1

Important Notation Definition

Notation Definition

Notation for Random Variable, Vector and Matrix

Throughout the course, we use the following notation to show random variable, random vector, random matrix and their corresponding outcomes:

X	Random variable (Upper-case letter)
x	Outcome of a random variable (lower-case letter)
$\mathbb{X}$	Random vector/matrix (Blackboard boldface letter)
$\boldsymbol{x}/\boldsymbol{X}$	Outcome of a random vector/matrix (Boldface letter)
Θ	Random variable/vector/matrix
θ	Outcome of random variable
$\boldsymbol{\theta}$	Outcome of random vector/matrix

Section 2

Basic Definitions

Basic Definitions

Covariance

- Suppose two random variables X and Y. The Covariance is defined as:

$$\mathrm{Cov}[X, Y] \triangleq \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

- Assume $\mathbb{X} = [X_1, X_2, \dots, X_D]^T$ is a D-dimensional random vector, then its covariance matrix is defined as:

$$\begin{aligned}\mathrm{Cov}[\mathbb{X}] &\triangleq \mathrm{E}[(\mathbb{X} - \mathrm{E}[\mathbb{X}])(\mathbb{X} - \mathrm{E}[\mathbb{X}])^T] = \boldsymbol{\Sigma} \\ &= \begin{bmatrix} \mathrm{Cov}[X_1, X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_D] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Cov}[X_2, X_2] & \cdots & \mathrm{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_D, X_1] & \mathrm{Cov}[X_D, X_2] & \cdots & \mathrm{Cov}[X_D, X_D] \end{bmatrix}\end{aligned}$$

- Cross-covariance: $\mathrm{Cov}[\mathbb{X}, \mathbb{Y}] = \mathrm{E}[(\mathbb{X} - \mathrm{E}[\mathbb{X}])(\mathbb{Y} - \mathrm{E}[\mathbb{Y}])^T]$

Covariance

- $\mathrm{E}[\mathbb{X}\mathbb{X}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T,\ \boldsymbol{\mu} \triangleq \mathrm{E}[\mathbb{X}]$
- $\mathrm{Cov}[\boldsymbol{A}\mathbb{X} + \boldsymbol{b}] = \boldsymbol{A}\,\mathrm{Cov}[\mathbb{X}]\boldsymbol{A}^T$

Basic Definitions

Correlation

- Suppose two random variables X and Y. The Correlation that measure the level of **Linear** relation between two variables is defined as:

$$\rho \triangleq \mathrm{Cor}[X, Y] \triangleq \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{V}[X]\,\mathrm{V}[Y]}}$$

- If $\mathbb{X}$ is a D-dimensional random vector, its correlation matrix is defined as:

$$\mathrm{Cor}[\mathbb{X}] \triangleq \begin{bmatrix} \mathrm{Cor}[X_1, X_1] = 1 & \mathrm{Cor}[X_1, X_2] & \cdots & \mathrm{Cor}[X_1, X_D] \\ \mathrm{Cor}[X_2, X_1] & \mathrm{Cor}[X_2, X_2] = 1 & \cdots & \mathrm{Cor}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cor}[X_D, X_1] & \mathrm{Cor}[X_D, X_2] & \cdots & \mathrm{Cor}[X_D, X_D] = 1 \end{bmatrix}$$

Correlation

- One can show that $-1 \leq \rho \leq 1$
- $|\,\mathrm{Cor}[X, Y]| = 1$ iff $Y = aX + b$

Correlation and Nonlinear Dependencies [1]

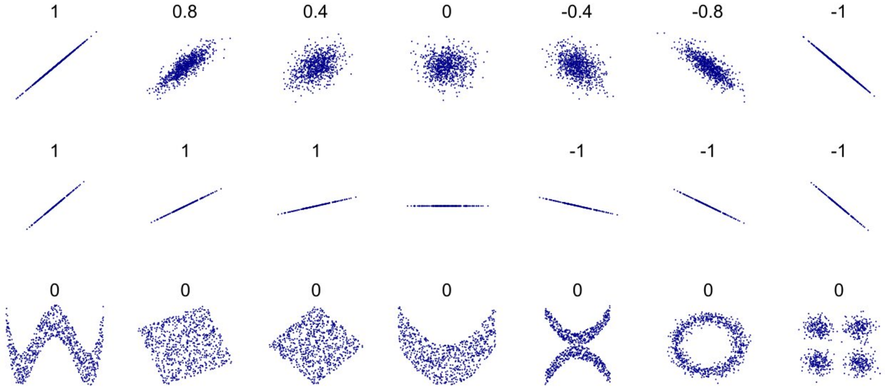


Figure: Visual interpretation of conditional probability

Uncorrelatedness vs. Independence

Independence implies Uncorrelatedness

$$\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[X]\,\mathrm{E}[Y] = 0$$
$$\Rightarrow \mathrm{Cor}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{V}[X]\,\mathrm{V}[Y]}} = 0$$

Uncorrelatedness Does NOT Imply Independence

Suppose: $\begin{cases} X \propto U(-1,1) \\ Y = X^2 \end{cases}$ Then: $\begin{cases} \mathrm{Cor}[X,Y] = 0 \ (\mathit{Uncorrelated}) \\ \mathrm{X} \not\perp\!\!\!\perp \mathrm{Y} \end{cases}$

Correlatedness vs. Causation

Causation Does NOT Imply Correlatedness

Suppose: $\begin{cases} X \propto U(-1,1) \\ Y = X^2 \end{cases}$ Then: $\begin{cases} \text{Cor}[X,Y] = 0 \ (\mathit{Uncorrelated}) \\ X \text{ clearly causes } Y. \end{cases}$

Correlatedness Does NOT Imply Causation

$\begin{cases} Z \propto U(-1,1) \\ X = Z^2 \\ Y = Z^2 \end{cases}$ Then: $\begin{cases} \text{Cor}[X,Y] = 1 \ (\mathit{Correlated}) \\ X \text{ and } Y \text{ don't have causal effect on each other.} \end{cases}$

Spurious Correlation [2]

Figure: Violent Crime Index vs Ice Cream Sales

Section 3

Sample Distributions

The Multivariate Gaussian (Normal) Distribution (MVN)

The Multivariate Gaussian (Normal) Distribution

Random vector $\mathbb{Y}$ is said to be multivariate normally distributed if every linear combination of its components has a univariate normal distribution.

Probability Density Function

The PDF for MVN with dimension D is defined as:

$$\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$$

where:

$$\boldsymbol{\mu} = \mathrm{E}[\mathbb{Y}] \in \mathbb{R}^D$$
$$\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbb{Y}] \in \mathbb{R}^{D\times D}$$

MVN Covariance Matrix Properties

Symmetric Matrix

Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric iff $\boldsymbol{A} = \boldsymbol{A}^T$ (We usually show this by $\boldsymbol{A} \in \mathbb{S}^n$)

Positive (Semi)Definite

Suppose $\boldsymbol{A} \in \mathbb{S}^n$. Then $\forall \boldsymbol{v} \in \mathbb{R}^n \setminus \{0\}$:

$$
\begin{array}{lcl}
\boldsymbol{A} \text{ is positive definite (PD), denoted } \boldsymbol{A} \succ 0 & \Leftrightarrow & \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} > 0 \\
\boldsymbol{A} \text{ is positive semidefinite (PSD), denoted } \boldsymbol{A} \succeq 0 & \Leftrightarrow & \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0 \\
\boldsymbol{A} \text{ is negative definite (ND), denoted } \boldsymbol{A} \prec 0 & \Leftrightarrow & \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} < 0 \\
\boldsymbol{A} \text{ is negative semidefinite (NSD), denoted } \boldsymbol{A} \preceq 0 & \Leftrightarrow & \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \leq 0
\end{array}
$$

$\boldsymbol{A}$ is indefinite iff it is none of the above.

MVN Covariance Matrix Properties

Covariance Matrix is PSD

Assume $\boldsymbol{\Sigma}$ to be the covariance matrix of $\mathbb{X}$ D-dimensional random vector. Then:

- $\boldsymbol{\Sigma} \in \mathbb{S}^D$ based on definition.
- $\boldsymbol{\Sigma} \succeq 0$ (PSD) because:

$$\boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v} = \mathbb{V}[\boldsymbol{v}^T \mathbb{X}] \geq 0, \ \forall \boldsymbol{v} \in \mathbb{R}^D$$

- If $\mathbb{X}$ is distributed normally, then $\boldsymbol{\Sigma} \succ 0$ (PD) because:

$$\exists \boldsymbol{v} \neq \mathbf{0} : \ \boldsymbol{v}^T \boldsymbol{\Sigma} \boldsymbol{v} = 0 \rightarrow \mathbb{V}[\boldsymbol{v}^T \mathbb{X}] = 0 \rightarrow \boldsymbol{v}^T \mathbb{X} \text{ is not normally distributed}$$

Bivariate Noraml (D=2)

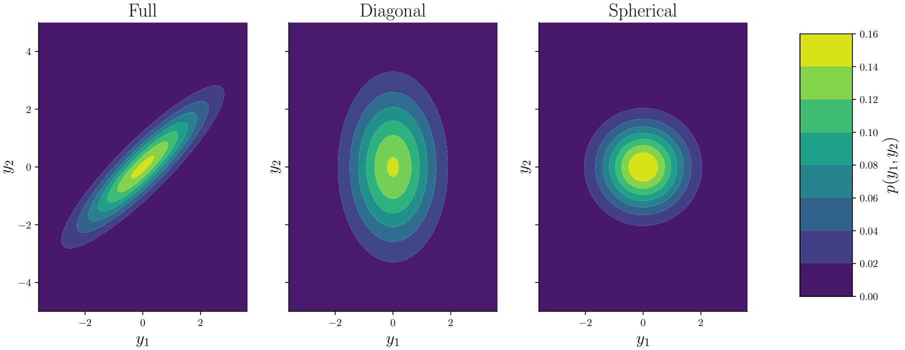


Figure: Level set of constant probability density

Mahalanobis Distance

Mahalanobis Distance

Mahalanobis Distance (Δ) is a metric to calculate the distance between point $\boldsymbol{y}$ and distribution p with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and is defined as:

$$\Delta^2 \triangleq (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$$

MVN and Mahalanobis Distance

The log probability of MVN at a specific point $\boldsymbol{y}$ is given by:

$$\log p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \overbrace{(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})}^{\Delta^2} + \text{constant}$$

Marginals and Conditionals of an MVN

Suppose $\mathbb{Y} = (\mathbb{Y}_1, \mathbb{Y}_2)$ where $\mathbb{Y}_1$ and $\mathbb{Y}_2$ have D_1 and D_2 dimension, respectively (thus $\mathbb{Y}$ is (D_1+D_2)-dimensional). Assume $\mathbb{Y}$ to be Gaussian with following parameters:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \; \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \; \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^{D_1}$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{D_2}$, $\boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{D_i \times D_j}$ and $\boldsymbol{\Lambda}_{ij} \in \mathbb{R}^{D_i \times D_j}$. Then the marginals and conditionals are given by:

$$\begin{aligned} p(\boldsymbol{y}_1) &= \mathcal{N}(\boldsymbol{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\boldsymbol{y}_2) &= \mathcal{N}(\boldsymbol{y}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \\ p(\boldsymbol{y}_1 | \boldsymbol{y}_2) &= \mathcal{N}(\boldsymbol{y}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \end{aligned}$$

where:

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2) \text{ (Affine function of observed vector } \boldsymbol{y}_2\text{)}$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \text{ (Independent of observed vector } \boldsymbol{y}_2\text{)}$$

Using MVN Marginals

Imputing Missing Values

Consider the following scenario:

- Select D movies
- Ask N people to give them scores ($\mathbb{Y} \in \mathbb{R}^D$)
- Some people have not scored all movies.
- You know that the scoring vector comes from $\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

How to fill missing scores by MVN marginals?

Solution

We can fill person n scoring vector as:

- Compute $p(\boldsymbol{y}_{n,\boldsymbol{h}}|\boldsymbol{y}_{n,\boldsymbol{v}}, \boldsymbol{\theta})$ where: $\begin{cases} \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \text{Parameters} \\ \boldsymbol{h} : \text{missing (hidden) score indices} \\ \boldsymbol{v} : \text{submitted (visible) score indices} \end{cases}$
- Impute missing values by: $\begin{cases} \bar{\boldsymbol{y}}_{n,\boldsymbol{h}} = \mathrm{E}[\mathbb{Y}_{n,\boldsymbol{h}}|\boldsymbol{y}_{n,\boldsymbol{v}}, \boldsymbol{\theta}] : \text{Posterior mean} \\ \text{Posterior sampling} \end{cases}$

Imputing Missing Values

How to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$? *Solution:* By using *Expectation Maximization.*

Section 4

Linear Gaussian Systems

Linear Gaussian Systems

Assume the following items:

- $\mathbb{Z} \in \mathbb{R}^L$: Unknown vector
- $\mathbb{Y} \in \mathbb{R}^D$: Noisy measurements
- The following distributions hold:
  - $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$
  - $p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y)$, $\boldsymbol{W} \in \mathbb{R}^{D\times L}$, $\boldsymbol{b} \in \mathbb{R}^D$

then:

- Joint distribution $p(\boldsymbol{z}, \boldsymbol{y}) = p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})$ is a $L + D$ dimensional Gaussian with the following parameters:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_z \\ \boldsymbol{W}\boldsymbol{\mu}_z + \boldsymbol{b} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_z\boldsymbol{W}^T \\ \boldsymbol{W}\boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_y + \boldsymbol{W}\boldsymbol{\Sigma}_z\boldsymbol{W}^T \end{bmatrix},$$

- Using Bayes rule, the posterior $p(\boldsymbol{z}|\boldsymbol{y})$ is also L dimensional Gaussian with the following parameters:

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{W}$$

$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y}\left[\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\mu}_z\right]$$

Conjugate Priors

Conjugate Priors

Assume $\mathcal{F}$ as a family of distribution functions (*e.g.* Gaussian). We say that a prior $p(\boldsymbol{z}) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(\boldsymbol{y}|\boldsymbol{z})$ if the posterior is in the same family of distribution, i.e., $p(\boldsymbol{z}|\boldsymbol{y}) \in \mathcal{F}$.

Conjugate Priors

Based on slide 22, Gaussian prior is a conjugate prior for the Gaussian likelihood.

Linear Gaussian System

Inferring an Unknown Scalar

Suppose:

- *Prior:* We want to estimate unknown quantity Z where $p(z) = \mathcal{N}(z|\mu_0, \lambda_0^{-1})$
- *Likelihood* We have N independent noisy measurements y_i distributed as $p(y_i|z) = \mathcal{N}(y_i|z, \lambda_y^{-1})$

compute the posterior $p(z|y_1, \ldots, y_N)$.

Solution

We start by defining $\mathbb{Y} = (y_1, \ldots, y_N)$. Then we can easily show that the problem is linear Gaussian system with $\boldsymbol{W} = \mathbf{1}_N$ and $\boldsymbol{\Sigma}_y^{-1} = \text{diag}(\lambda_y \boldsymbol{I})$. Thus:

$$p(z|\boldsymbol{y}) = \mathcal{N}(z|\mu_N, \lambda_N^{-1})$$

where:

$$\boldsymbol{\Sigma}_{z|\boldsymbol{y}}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \boldsymbol{W}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} \boldsymbol{W} \Rightarrow \lambda_{z|y} = \lambda_0 + \mathbf{1}^T \text{diag}(\lambda_y \boldsymbol{I}) \mathbf{1} = \lambda_0 + N\lambda_y$$

$$\boldsymbol{\mu}_{z|\boldsymbol{y}} = \boldsymbol{\Sigma}_{z|\boldsymbol{y}} \left[\boldsymbol{W}^T \boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} (\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z \right] \Rightarrow \mu_{z|\boldsymbol{y}} = \lambda_{z|\boldsymbol{y}}^{-1} \left[\mathbf{1}^T \text{diag}(\lambda_y \boldsymbol{I}) (\boldsymbol{y} - \mathbf{0}) + \lambda_0 \mu_0 \right]$$

$$\Rightarrow \mu_{z|\boldsymbol{y}} = \frac{N\lambda_y \bar{y} + \lambda_0 \mu_0}{\lambda_{z|\boldsymbol{y}}} = \frac{N\lambda_y}{N\lambda_y + \lambda_0} \bar{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0} \mu_0$$

Linear Gaussian System

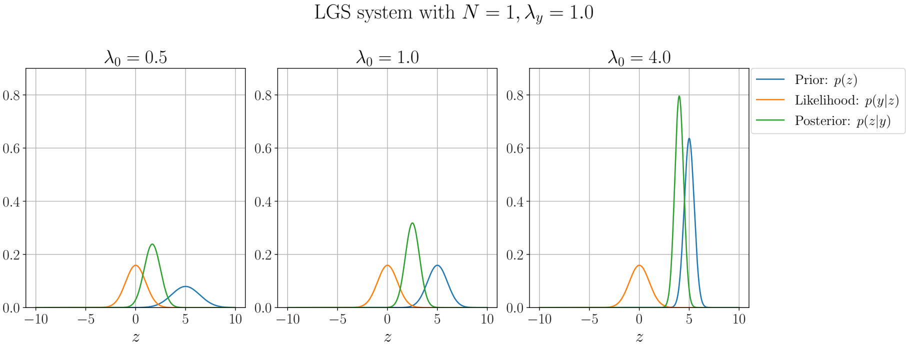


Figure: Prior precision (λ_0) effect

Linear Gaussian System

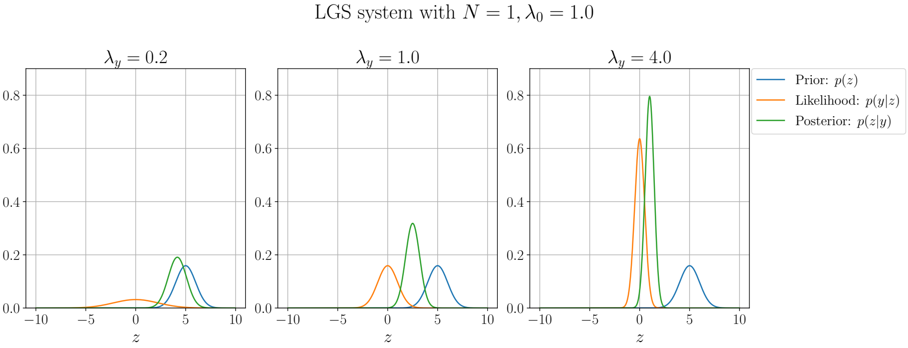


Figure: Likelihood precision (λ_y) effect

Linear Gaussian System

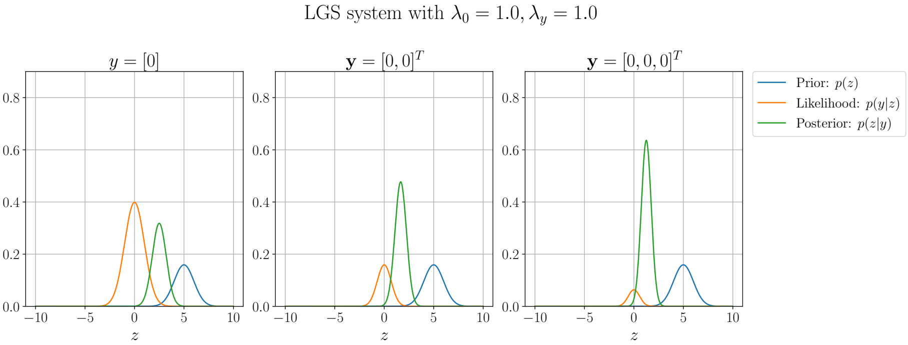


Figure: Number of measurements (N) effect

Linear Gaussian System

Sensor Fusion

Suppose:

- *Prior:* We want to estimate unknown vector $\mathbb{Z}$ where $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mu_0, \boldsymbol{\Sigma}_0)$
- *Likelihood:* We have 2 sensors and 1 measurements of each sensor, denoted $\mathbb{Y}_1$ and $\mathbb{Y}_2$, distributes as $\mathcal{N}(\boldsymbol{y}_i|\boldsymbol{z}, \boldsymbol{\Sigma}_i)$ ($\boldsymbol{\Sigma}_i$ demonstrates the reliability for i-th sensor).

compute the posterior $p(\boldsymbol{z}|\boldsymbol{y}_1, \boldsymbol{y}_2)$.

Solution

We start by defining $\mathbb{Y} = (\mathbb{Y}_1, \mathbb{Y}_2)$. Then we can easily show that the problem is linear Gaussian system with $\boldsymbol{W} = [\boldsymbol{I}; \boldsymbol{I}]$ and $\boldsymbol{\Sigma}_y = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_2 \end{bmatrix}$. Thus the posterior $p(\boldsymbol{z}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y})$ where $\boldsymbol{\mu}_{z|y}$ and $\boldsymbol{\Sigma}_{z|y}$ can be calculated using formulas in Slide 22.

Sensor Fusion

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\boldsymbol{\mu}_0 = [0;0],\ \boldsymbol{\Sigma}_0 = 1000\boldsymbol{I},\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 0.01\boldsymbol{I}$$

and assume $\boldsymbol{y}_1 = (0,-1)$ and $\boldsymbol{y}_2 = (1,0)$. Visualize th measurements and posterior $p(\boldsymbol{z}|\boldsymbol{y})$.

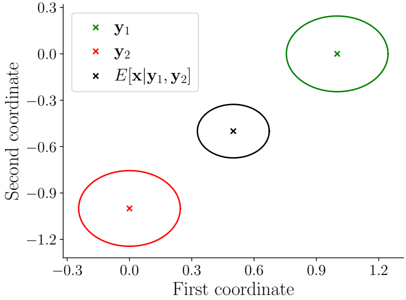


Figure: Sensor fusion result

Sensor Fusion

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\boldsymbol{\mu}_0 = [0; 0], \ \boldsymbol{\Sigma}_0 = 1000\boldsymbol{I}, \ \boldsymbol{\Sigma}_1 = 0.01\boldsymbol{I}, \ \boldsymbol{\Sigma}_2 = 0.05\boldsymbol{I}$$

and assume $\boldsymbol{y}_1 = (0, -1)$ and $\boldsymbol{y}_2 = (1, 0)$. Visualize th measurements and posterior $p(\boldsymbol{z}|\boldsymbol{y})$.

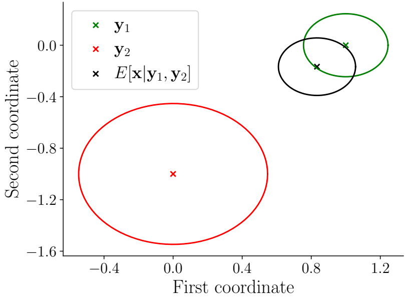


Figure: Sensor fusion result

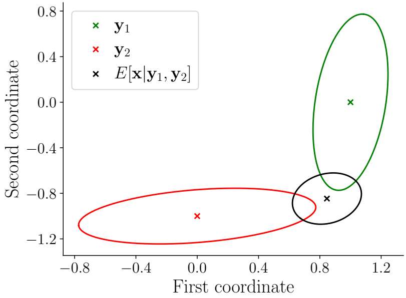
Sensor Fusion

Sensor Fusion

Suppose the sensor fusion example in Slide 28, with the following parameters:

$$\boldsymbol{\mu}_0 = [0;0],\ \boldsymbol{\Sigma}_0 = 1000\boldsymbol{I},\ \boldsymbol{\Sigma}_1 = 0.01\begin{bmatrix} 10 & 1 \\ 1 & 1 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = 0.01\begin{bmatrix} 1 & 1 \\ 1 & 10 \end{bmatrix}$$

and assume $\boldsymbol{y}_1 = (0,-1)$ and $\boldsymbol{y}_2 = (1,0)$. Visualize th measurements and posterior $p(\boldsymbol{z}|\boldsymbol{y})$.



Section 5

Mixture Models

Mixture Models

Mixture Models

One way to create more complex probability models is to take a convex combination of simple distributions. This is called a mixture model. This has the form $p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\boldsymbol{y})$ where:

- p_k is the k-th mixture component
- $\{\pi_k\}_{k=1}^{K}$ are mixture weights with the following constraints:
  - $0 \leq \pi_k \leq 1, k = 1, \ldots, K$
  - $\sum_{k=1}^{K} \pi_k = 1$

Mixture Models - Generative Story

Suppose latent variable Z to be a categorical RV and distributed as $p(z|\boldsymbol{\theta}) = Cat(z|\boldsymbol{\pi})$ and conditional $p(\boldsymbol{y}|z = k, \boldsymbol{\theta}) = p_k(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}_k)$. We can interpret mixture models as follows:

- We sample a specific component.
- We generate $\boldsymbol{y}$ using sampled value of z.

Using the above procedure, we have:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z = k|\boldsymbol{\theta}) p(\boldsymbol{y}|z = k, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{y}|\boldsymbol{\theta}_k)$$

Gaussian Mixture Model

Gaussian Mixture Model (GMM) or Mixture of Gaussian (MoG) is defined as:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

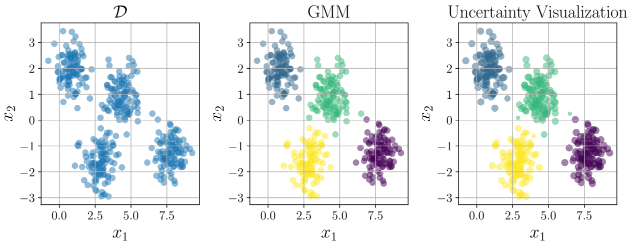


Figure: Sample GMM distribution and its application for clustering

References I

"Pearson correlation coefficient," https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

"The logic of causal conclusions: How we know that fire burns, fertilizer helps plants grow, and vaccines prevent disease," http://icbseverywhere.com/blog/2014/10/the-logic-of-causal-conclusions/.