

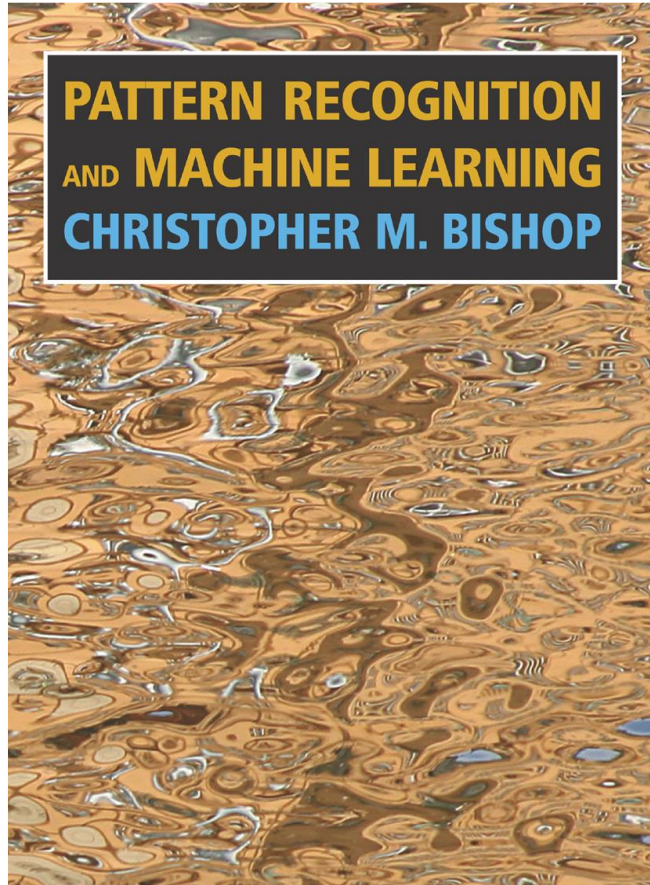
Gaussian Identities (cont.)

Song Liu (song.liu@bristol.ac.uk)

Office Hour: Wednesday 4pm-5pm

NOT LAB this week.

Reference



Today's class *roughly* follows
Chapter 2.3-2.34

Pattern Recognition and
Machine Learning

Christopher Bishop, 2006

Recap

- $N_x(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot \exp \left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}{2} \right]$
- MVNs are multi-dimensional generalizations of univariate normal distributions, in the sense that:
 - $\boldsymbol{\Sigma}^{-1} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$
 - Eigen-decomposition, \mathbf{D} is diagonal.
 - $\mathbf{U} \in \mathbb{R}^{d \times d}$, $\mathbf{U} \mathbf{U}^\top = \mathbf{U}^\top \mathbf{U} = \mathbf{I}$
 - $\mathbf{y} = \mathbf{U}^\top (\mathbf{x} - \boldsymbol{\mu})$.
 - $p(\mathbf{y}) = \prod_i N_{\mathbf{y}}(\mathbf{0}, \sigma_i^2)$, σ_i^2 is i -th eigenvalue of $\boldsymbol{\Sigma}$.
 - Use this to generate samples of MVN using uni-normal!

Recap

- **Mahalanobis distance**, $\sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$
 - Distance between a point \mathbf{x} to the center of $N_x(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
 - Distance between \mathbf{x} and $\boldsymbol{\mu}$ rotated by \mathbf{U} .
 - Can be used to define the **confidence region**.
- **Moments of MVN.**
 - $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$
 - Apply the transform $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$.
 - $\int_{-\infty}^0 \exp\left[-\frac{\mathbf{z}\boldsymbol{\Sigma}^{-1}\mathbf{z}}{2}\right] \mathbf{z} d\mathbf{z} = -\int_0^{\infty} \exp\left[-\frac{\mathbf{z}\boldsymbol{\Sigma}^{-1}\mathbf{z}}{2}\right] \mathbf{z} d\mathbf{z}$
 - $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\mu}^\top \boldsymbol{\mu} + \boldsymbol{\Sigma}$
 - Apply the transform $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$.
 - Use $\mathbf{z} = \mathbf{U}\mathbf{y}$ and $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$.

Partitioned MVNs

- **Given:**

- $p(\mathbf{x}_a, \mathbf{x}_b) = N_{\mathbf{x}_a, \mathbf{x}_b} \left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix} \right)$

- Represent $p(\mathbf{x}_a | \mathbf{x}_b)$ and $p(\mathbf{x}_a)$ using $\boldsymbol{\mu}_a$ and $\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab}$, $\boldsymbol{\mu}_b$ and $\boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb}$.

- Partitioned MVN formulas have huge applications in Bayesian regression, Gaussian graphical models etc.

- For simplicity, we let $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{aa} & \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba} & \boldsymbol{\Theta}_{bb} \end{bmatrix}$.

Proof Walkthrough

- You can prove by following the def. of conditional dist.
- **However, observe:**

$$\log N_{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = - \frac{\begin{matrix} \text{Quadratic term} \\ \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Theta}_{aa} & \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba} & \boldsymbol{\Theta}_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix} \end{matrix}}{2} + \text{const.}$$

- $\log N_{\mathbf{x}}$ is merely a quadratic function w.r.t $\mathbf{x} + \text{const.}$
- Expanding quad. term only leads to quad./linear terms.
 - w.r.t. $\mathbf{x}_a, \mathbf{x}_b$
- $\Rightarrow P(\mathbf{x}_a | \mathbf{x}_b)$ is an MVN (not rigorously speaking).

Proof Walkthrough

- If $p(\mathbf{t}) = N_{\mathbf{t}}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, then

$$\log p(\mathbf{t}) = -\frac{\mathbf{t}^\top \boldsymbol{\Sigma}^{-1} \mathbf{t}}{2} + \mathbf{t}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \text{const.}$$



- If we spot terms in $-\begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Theta}_{aa} & \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba} & \boldsymbol{\Theta}_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix} / 2$ with respect to \mathbf{x}_a which has the same form as those in $\phantom{-\begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Theta}_{aa} & \boldsymbol{\Theta}_{ab} \\ \boldsymbol{\Theta}_{ba} & \boldsymbol{\Theta}_{bb} \end{bmatrix} \begin{bmatrix} \mathbf{x}_a - \boldsymbol{\mu}_a \\ \mathbf{x}_b - \boldsymbol{\mu}_b \end{bmatrix} / 2}$, we can *directly identify* the covariance and mean for $p(\mathbf{x}_a | \mathbf{x}_b)$.

Proof Walkthrough

- The *quadratic term* w.r.t. \mathbf{x}_a after expansion:
- $-\mathbf{x}_a^\top \Theta_{aa} \mathbf{x}_a / 2 \Rightarrow \text{Cov}_{\mathbf{x}_a | \mathbf{x}_b} [\mathbf{x}_a] = [\Theta_{aa}]^{-1}$.
- The *linear terms* w.r.t. \mathbf{x}_a after expansion:
- $-\mathbf{x}_a^\top \Theta_{ab} \mathbf{x}_b + \mathbf{x}_a^\top \Theta_{ab} \boldsymbol{\mu}_b + \mathbf{x}_a^\top \Theta_{aa} \boldsymbol{\mu}_a$
- Collect terms: $\mathbf{x}_a^\top \Theta_{aa} (\boldsymbol{\mu}_a - \Theta_{aa}^{-1} \Theta_{ab} \mathbf{x}_b + \Theta_{aa}^{-1} \Theta_{ab} \boldsymbol{\mu}_b)$
- Knowing $\text{Cov}_{\mathbf{x}_a | \mathbf{x}_b} [\mathbf{x}_a] = [\Theta_{aa}]^{-1} \Rightarrow$

$$\mathbb{E}_{\mathbf{x}_a | \mathbf{x}_b} [\mathbf{x}_a] = \boldsymbol{\mu}_a - \Theta_{aa}^{-1} \Theta_{ab} \mathbf{x}_b + \Theta_{aa}^{-1} \Theta_{ab} \boldsymbol{\mu}_b$$

Conditional MVN formula

- $p(\mathbf{x}_a | \mathbf{x}_b) = N_{\mathbf{x}_a}(\boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \mathbf{x}_b + \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b, \boldsymbol{\Theta}_{aa}^{-1}).$
- You can use block matrix inversion formula to represent $\boldsymbol{\Theta}_{aa}, \boldsymbol{\Theta}_{ab}$ using $\boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab}$ and $\boldsymbol{\Sigma}_{bb}$.
- **See 2.76 in PRML**
- However, this formula is most easily expressed using block matrices of $\boldsymbol{\Theta}$.

Partitioned MVNs (Marginal)

- How to represent $p(\mathbf{x}_a)$ using $\begin{matrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{matrix}$ and $\begin{matrix} \boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb} \end{matrix}$?
- First, we marginalize $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$
- Write terms in $\log p(\mathbf{x}_a, \mathbf{x}_b)$ w.r.t. \mathbf{x}_b after expansion:
- $-\mathbf{x}_b^\top \boldsymbol{\Theta}_{bb} \mathbf{x}_b / 2 + \mathbf{x}_b^\top (\underbrace{\boldsymbol{\Theta}_{bb} \boldsymbol{\mu}_b - \boldsymbol{\Theta}_{ba} \mathbf{x}_a + \boldsymbol{\Theta}_{ba} \boldsymbol{\mu}_a}_{\mathbf{m}})$

$$= -\underbrace{(\mathbf{x}_b^\top - \boldsymbol{\Theta}_{bb}^{-1} \mathbf{m}) \boldsymbol{\Theta}_{bb} (\mathbf{x}_b - \boldsymbol{\Theta}_{bb}^{-1} \mathbf{m})}_{\text{Completing the square!}} / 2 + \mathbf{m}^\top \boldsymbol{\Theta}_{bb}^{-1} \mathbf{m} / 2,$$

Completing the square!

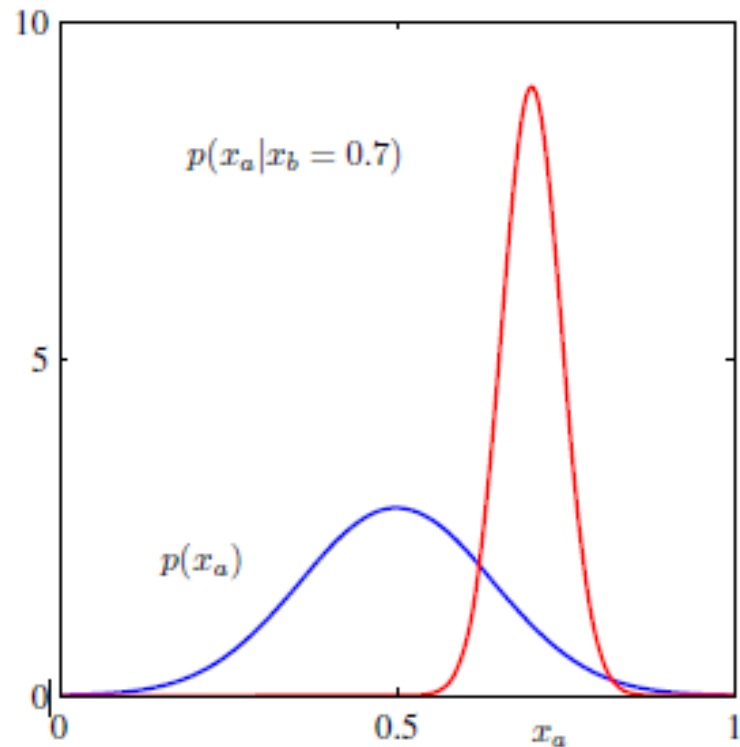
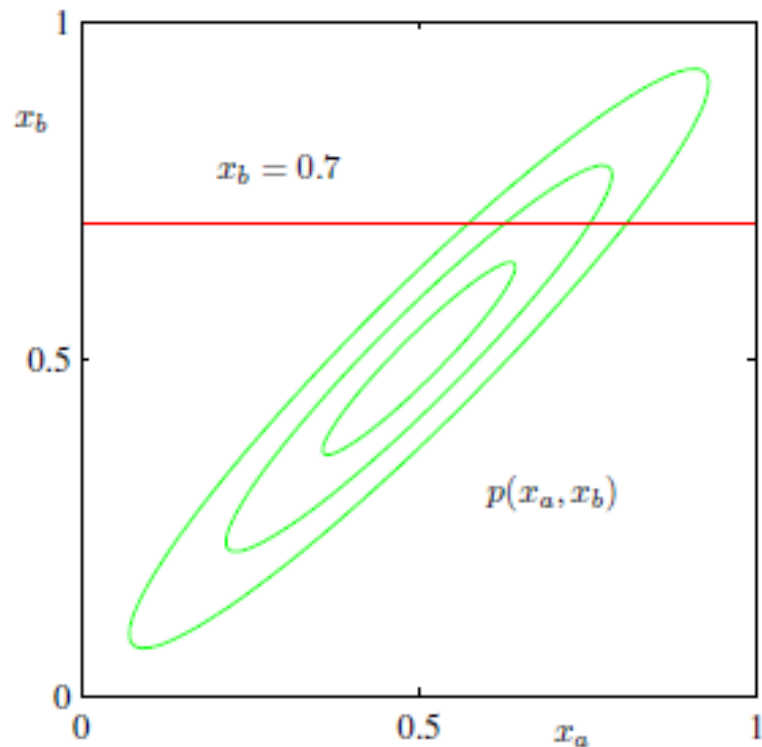
Proof Walkthrough

- Now we know
- $p(\mathbf{x}_a) =$
 $(\dots) \exp\left(\frac{\mathbf{m}^\top \Theta_{bb}^{-1} \mathbf{m}}{2}\right) \int \exp\left[-\frac{(\mathbf{x}_b^\top - \Theta_{bb}^{-1} \mathbf{m}) \Theta_{bb} (\mathbf{x}_b - \Theta_{bb}^{-1} \mathbf{m})}{2}\right] d\mathbf{x}_b$
- Inside integral, just a regular MVN w.r.t. \mathbf{x}_b without normalizing constant, so
- $p(\mathbf{x}_a) = (\dots) \exp\left(\frac{\mathbf{m}^\top \Theta_{bb}^{-1} \mathbf{m}}{2}\right) \cdot \text{const}$
- Now, let us find all terms w.r.t. \mathbf{x}_a in above expression.

Proof Walkthrough

- $\log p(\mathbf{x}_a) = - \frac{\mathbf{x}_a^\top (\Theta_{aa} - \Theta_{ab} \Theta_{bb}^{-1} \Theta_{ba}) \mathbf{x}_a}{2} + \mathbf{x}_a^\top (\Theta_{aa} - \Theta_{ab} \Theta_{bb}^{-1} \Theta_{ba}) \boldsymbol{\mu}_a + \text{const}$
- Using the block matrix inversion formula, $\Theta_{aa} - \Theta_{ab} \Theta_{bb}^{-1} \Theta_{ba} = \Sigma_{aa}^{-1}$.
- Therefore, $p(\mathbf{x}_a) = N_{\mathbf{x}_a}(\boldsymbol{\mu}_a, \Sigma_{aa})$
- The marginal of a joint MVN has mean and variance that is the same as the mean and variance of the partitioned MVN.

Visualization



- PRML 2.9

Gaussian Linear Model

- The prior: $p(\mathbf{x}) = N_{\mathbf{x}}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$
- The Likelihood: $p(\mathbf{y}|\mathbf{x}) = N_{\mathbf{y}}(\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1})$

Linear model



- The marginal: $p(\mathbf{y}) = N_{\mathbf{y}}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\top})$
- The posterior: $p(\mathbf{x}|\mathbf{y}) = N_{\mathbf{x}}(\boldsymbol{\Sigma}\{\mathbf{A}^{\top}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\top}\mathbf{L}\mathbf{A})^{-1}$

Proof: 1. Calculate the joint $p(\mathbf{y}, \mathbf{x})$, 2. Use formula we just derived to obtain marginal and conditional dist.

Read PRML, 2.3.3

Likelihood for MVN

- Given the dataset $D := \{\mathbf{x}_i\}_{i=1}^n$, the likelihood function of MVN density can be written as
- $$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D) = \sum_{i=1}^n \log N_{\mathbf{x}_i}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \text{const} - \frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{\text{tr}(\bar{\mathbf{X}} \bar{\mathbf{X}}^\top \boldsymbol{\Sigma}^{-1})}{2}$$
- where $\bar{\mathbf{X}} = [(\mathbf{x}_1 - \boldsymbol{\mu}) \dots (\mathbf{x}_n - \boldsymbol{\mu})] \in R^{d \times n}$ is the “centralized” dataset.
- tr is the trace operator.

Maximum Likelihood Estimator

- $\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D) = \max_{\boldsymbol{\Sigma}} \max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D).$
- First, solve the inner max by
 - $\frac{\partial L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D)}{\partial \boldsymbol{\mu}} = 0 \implies \boldsymbol{\mu}_{\text{MLE}} := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$
 - Then, plug in $\boldsymbol{\mu}_{\text{MLE}}$ and solve the outer max by
 - $\frac{\partial L(\boldsymbol{\mu}_{\text{MLE}}, \boldsymbol{\Sigma}, D)}{\partial \boldsymbol{\Sigma}} = 0 \implies$
 - $\boldsymbol{\Sigma}_{\text{MLE}} := \frac{1}{n} \bar{\mathbf{X}}_{\text{MLE}} \bar{\mathbf{X}}_{\text{MLE}}^{\top},$
 - where $\bar{\mathbf{X}}_{\text{MLE}} := [(\mathbf{x}_1 - \boldsymbol{\mu}_{\text{MLE}}) \dots (\mathbf{x}_n - \boldsymbol{\mu}_{\text{MLE}})]$

Bias-Variance Decomposition

Song Liu (song.liu@bristol.ac.uk)

