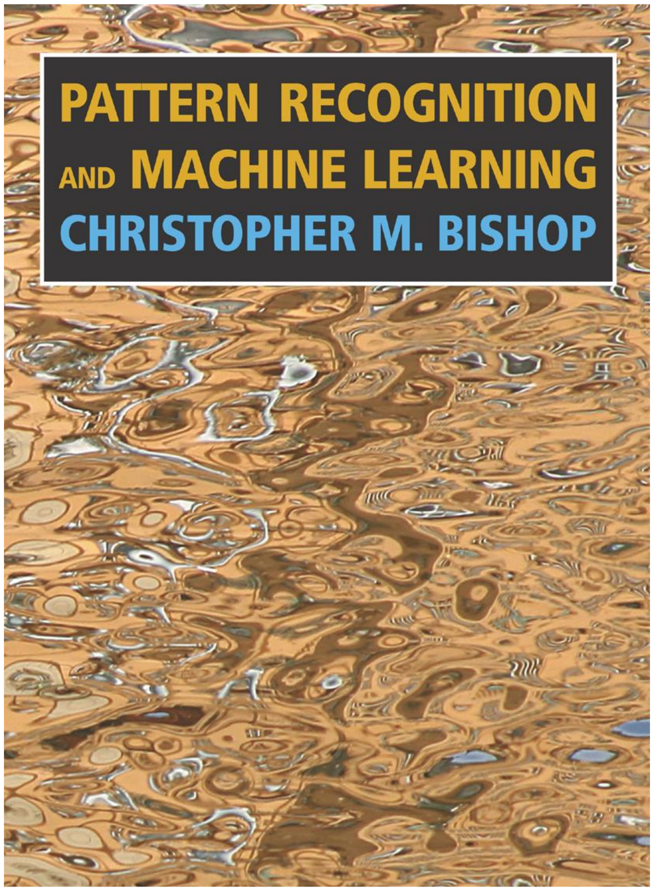
Gaussian Identities (cont.)

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

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

Recap

- **Mahalanobis distance,** $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$
    - Distance between a point $\boldsymbol{x}$ to the center of $N_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
    - Distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$ rotated by $\boldsymbol{U}$.
    - Can be used to define the **confidence region.**

- **Moments of MVN.**
    - $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$
        - Apply the transform $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$.
        - $\int_{-\infty}^{\mathbf{0}} \exp\left[-\frac{\boldsymbol{z}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}{2}\right] \boldsymbol{z} d\boldsymbol{z} = -\int_{\mathbf{0}}^{\infty} \exp\left[-\frac{\boldsymbol{z}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}}{2}\right] \boldsymbol{z}\, d\boldsymbol{z}$
    - $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\mu}^\top\boldsymbol{\mu} + \boldsymbol{\Sigma}$
        - Apply the transform $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{\mu}$.
        - Use $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{y}$ and $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$.

Partitioned MVNs

- **Given:**
- $p(\boldsymbol{x}_a, \boldsymbol{x}_b) = N_{\boldsymbol{x}_a, \boldsymbol{x}_b}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb} \end{bmatrix}\right)$
- Represent $p(\boldsymbol{x}_a | \boldsymbol{x}_b)$ and $p(\boldsymbol{x}_a)$ using $\begin{matrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{matrix}$ and $\begin{matrix} \boldsymbol{\Sigma}_{aa}, \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba}, \boldsymbol{\Sigma}_{bb} \end{matrix}$.

- 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:**

Quadratic term

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

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

Proof Walkthrough

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

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

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

Proof Walkthrough

- The *quadratic term* w.r.t. $\boldsymbol{x}_a$ after expansion:
- $-\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} \boldsymbol{x}_a / 2 \Longrightarrow \boxed{\mathrm{Cov}_{\boldsymbol{x}_a|\boldsymbol{x}_b}[\boldsymbol{x}_a] = [\boldsymbol{\Theta}_{aa}]^{-1}}$.
- The *linear terms* w.r.t. $\boldsymbol{x}_\mathrm{a}$ after expansion:
- $-\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{ab} \boldsymbol{x}_b + \boldsymbol{x}_a^\top \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b + \boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} \boldsymbol{\mu}_a$

- Collect terms: $\boldsymbol{x}_a^\top \boldsymbol{\Theta}_{aa} (\boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{x}_b + \boldsymbol{\Theta}_{aa}^{-1} \boldsymbol{\Theta}_{ab} \boldsymbol{\mu}_b)$
- Knowing $\mathrm{Cov}_{\boldsymbol{x}_a|\boldsymbol{x}_b}[\boldsymbol{x}_a] = [\boldsymbol{\Theta}_{aa}]^{-1} \Longrightarrow$

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

Conditional MVN formula

- $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \boxed{N_{\boldsymbol{x}_a}(\boldsymbol{\mu}_a - \boldsymbol{\Theta}_{aa}^{-1}\boldsymbol{\Theta}_{ab}\boldsymbol{x}_b + \boldsymbol{\Theta}_{aa}^{-1}\boldsymbol{\Theta}_{ab}\boldsymbol{\mu}_b, \boldsymbol{\Theta}_{\boldsymbol{aa}}^{-1}).}$

- 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 $p(\boldsymbol{x}_a)$ 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 $p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_b$
- Write terms in $\log p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ w.r.t. $\boldsymbol{x}_b$ after expansion:
- $-\boldsymbol{x}_b^\top \boldsymbol{\Theta}_{bb} \boldsymbol{x}_b / 2 + \boldsymbol{x}_b^\top \underbrace{(\boldsymbol{\Theta}_{bb} \boldsymbol{\mu}_b - \boldsymbol{\Theta}_{ba} \boldsymbol{x}_a + \boldsymbol{\Theta}_{ba} \boldsymbol{\mu}_a)}_{\boldsymbol{m}}$

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

Proof Walkthrough

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

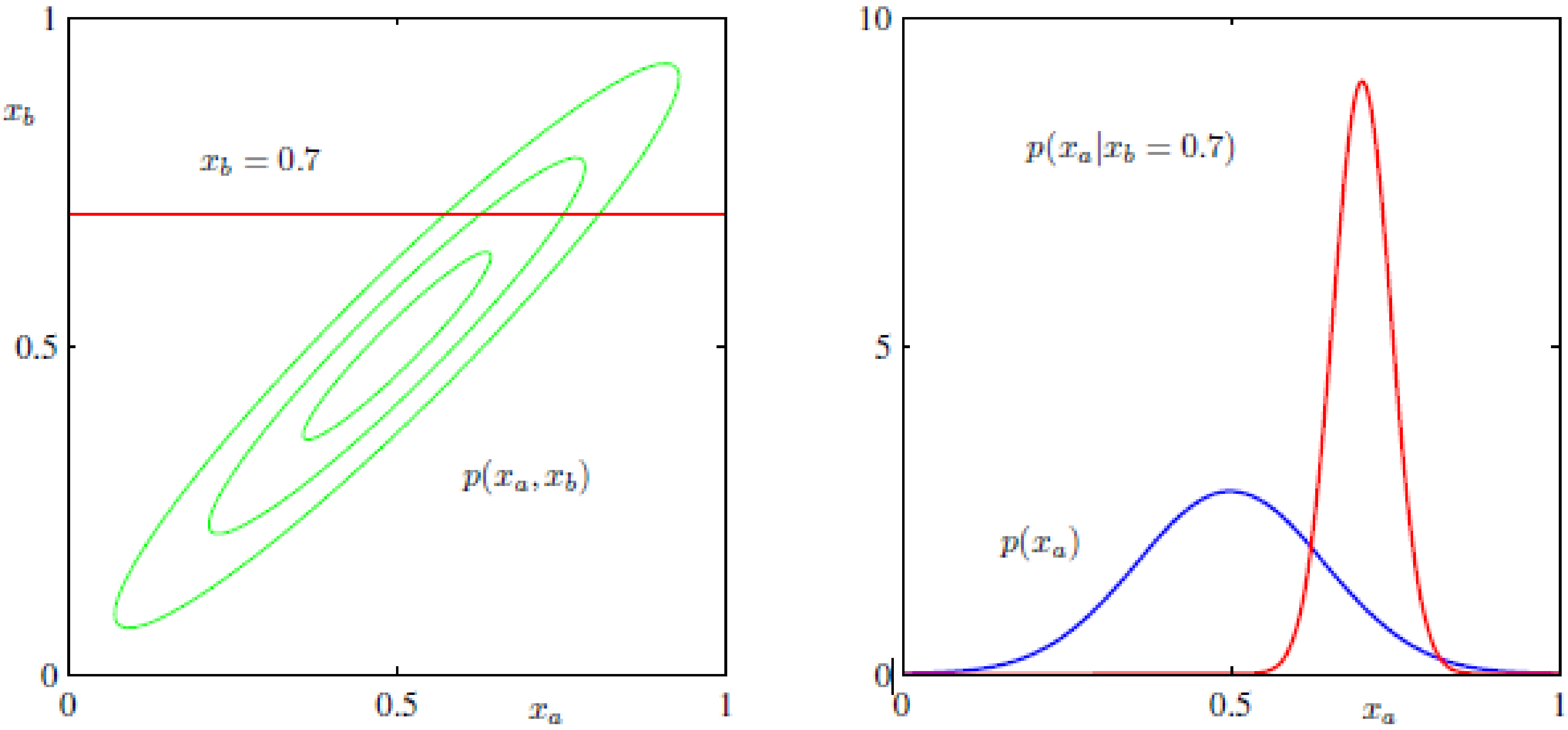
Proof Walkthrough

- $\log \boldsymbol{p}(\boldsymbol{x}_a) = -\frac{\boldsymbol{x}_a^\top \boxed{(\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba})} \boldsymbol{x}_a}{2} + \boldsymbol{x}_a^\top (\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba}) \boxed{\boldsymbol{\mu}_a} + \text{const}$

- Using the block matrix inversion formula, $\boldsymbol{\Theta}_{aa} - \boldsymbol{\Theta}_{ab}\boldsymbol{\Theta}_{bb}^{-1}\boldsymbol{\Theta}_{ba} = \boldsymbol{\Sigma}_{aa}^{-1}$.
- Therefore, $\boxed{\boldsymbol{p}(\boldsymbol{x}_a) = N_{\boldsymbol{x}_a}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})}$

- 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: $p(\boldsymbol{x}) = N_x(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$
- The Likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = N_y(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$

Linear model

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

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

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

Read PRML, 2.3.3

Likelihood for MVN

- Given the dataset $D := \{\boldsymbol{x}_i\}_{i=1}^{n}$, the likelihood function of MVN density can be written as

- $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, D) = \sum_{i=1} \log N_{\boldsymbol{x}_i}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$= \text{const} - \frac{n}{2}\log|\Sigma| - \frac{\text{tr}\left(\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^{\top}\boldsymbol{\Sigma}^{-1}\right)}{2}$$

- where $\overline{\boldsymbol{X}} = [(\boldsymbol{x}_1 - \boldsymbol{\mu}) \dots (\boldsymbol{x}_n - \boldsymbol{\mu})] \in R^{d\times n}$ 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
- $\frac{\partial L(\boldsymbol{\mu},\boldsymbol{\Sigma},D)}{\partial \boldsymbol{\mu}} = 0 \implies \boldsymbol{\mu}_{\mathrm{MLE}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$
- Then, plug in $\boldsymbol{\mu}_{\mathrm{MLE}}$ and solve the outer max by
- $\frac{\partial L(\boldsymbol{\mu}_{\mathrm{MLE}},\boldsymbol{\Sigma},D)}{\partial \boldsymbol{\Sigma}} = 0 \implies$
- $\boldsymbol{\Sigma}_{\mathrm{MLE}} \coloneqq \frac{1}{n}\overline{\boldsymbol{X}}_{\mathrm{MLE}}\overline{\boldsymbol{X}}_{\mathrm{MLE}}^{\top}$,
- where $\overline{\boldsymbol{X}}_{\mathrm{MLE}} := [(\boldsymbol{x}_1 - \boldsymbol{\mu}_{\mathrm{MLE}}) \ldots (\boldsymbol{x}_n - \boldsymbol{\mu}_{\mathrm{MLE}})]$