# Handout Probabilistic Graphical Models: Gaussian Computations

## 1 Definitions

The density of a multivariate Gaussian variable $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ($\boldsymbol{x} \in \mathbb{R}^n$):

$$P(\boldsymbol{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{1}$$

Here, $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}]$ , $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{x}]$

## 2 Closure Properties. How to Determine a Gaussian Result

A family of distributions is *closed* under a set of operations on distributions or random variables if whenever you apply an operation to a family member, the outcome lies in the family as well.

- Gaussians are closed under linear (affine) transformations:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ \boldsymbol{y} = \boldsymbol{E}\boldsymbol{x} + \boldsymbol{b} \quad \Rightarrow \quad \boldsymbol{y} \sim N(\boldsymbol{E}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{E}^T)$$

- Gaussians are closed under marginalization:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\ I \subset \{1, \dots, n\} \quad \Rightarrow \quad \boldsymbol{x}_I = (x_i)_{i \in I} \text{ Gaussian}$$

  In other words: the *sum* rule retains Gaussianity.

- Gaussians are closed under conditioning:

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\ I \subset \{1, \dots, n\},\ R = \{1, \dots, n\} \setminus I \quad \Rightarrow \quad P(\boldsymbol{x}_I | \boldsymbol{x}_R) \text{ Gaussian}$$

  In other words: the *product* rule retains Gaussianity.

### 3.2 Conditional Distribution by Sampling Argument

To get to $P(\boldsymbol{x}_I|\boldsymbol{x}_R)$ directly, we can use a sampling argument. Let's first get rid of means by transforming $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}$, adding it back in later. We know that $P(\boldsymbol{y}) = P(\boldsymbol{y}_I|\boldsymbol{y}_R)P(\boldsymbol{y}_R)$, which tells us how to sample $\boldsymbol{y}$:

1. Draw $\boldsymbol{y}_R \sim N(\mathbf{0}, \boldsymbol{\Sigma}_R)$
2. Draw $\boldsymbol{y}_I \sim P(\boldsymbol{y}_I|\boldsymbol{y}_R) = N(?, ?)$

And we know the outcome, namely $\boldsymbol{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. We use an *ansatz*, which means that we guess a form for the solution. The ansatz is that $\boldsymbol{y}_I = \boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R$, where $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{C})$ is independentof$\boldsymbol{y}_R$.

$$\begin{bmatrix} \boldsymbol{y}_I \\ \boldsymbol{y}_R \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y}_R \end{bmatrix}. \tag{3}$$

$$\mathrm{Cov}[\boldsymbol{y}] = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{B} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_R \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ \boldsymbol{B}^T & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} + \boldsymbol{B}\boldsymbol{\Sigma}_R\boldsymbol{B}^T & \boldsymbol{B}\boldsymbol{\Sigma}_R \\ \boldsymbol{B}^T\boldsymbol{\Sigma}_R & \boldsymbol{\Sigma}_R \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} \boldsymbol{\Sigma}_I & \boldsymbol{\Sigma}_{I,R} \\ \boldsymbol{\Sigma}_{R,I} & \boldsymbol{\Sigma}_R \end{bmatrix} \tag{4}$$

Note that the matrix in the middle is block-diagonal, because $\boldsymbol{u}$ and $\boldsymbol{y}_R$ are independent First, $\boldsymbol{B} = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R{}^{-1}$. Second,

$$\boldsymbol{C} = \boxed{\boldsymbol{\Sigma}_I - \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}\boldsymbol{\Sigma}_{R,I} =: \boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R.\ \textit{Schur complement.}} \tag{5}$$

==>

$$\mathrm{E}[\boldsymbol{y}_I|\boldsymbol{y}_R] = \mathrm{E}[\boldsymbol{u} + \boldsymbol{B}\boldsymbol{y}_R|\boldsymbol{y}_R] = \boldsymbol{B}\boldsymbol{y}_R = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}(\boldsymbol{x}_R - \boldsymbol{\mu}_R).$$

$$\mathrm{Cov}[\boldsymbol{x}_I|\boldsymbol{x}_R] = \mathrm{Cov}[\boldsymbol{u}] = \boldsymbol{C} = \boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R,$$

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ -\boldsymbol{B}^T & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} (\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_R)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_R^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{B} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \stackrel{!}{=} \begin{bmatrix} \boldsymbol{A}_I & \boldsymbol{A}_{I,R} \\ \boldsymbol{A}_{R,I} & \boldsymbol{A}_R \end{bmatrix},$$

where $\boldsymbol{B} = \boldsymbol{\Sigma}_{I,R}\boldsymbol{\Sigma}_R^{-1}$

# 4 Linear-Gaussian Model

the linear-Gaussian model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon}, \quad \boldsymbol{u} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Psi}).$$

This is a fundamental latent variable model: $\boldsymbol{u}$ is latent, $\boldsymbol{y}$ is observed. For modelling data, what matters is the marginal distribution of $\boldsymbol{y}$, whose structure is determined by the dimensionality of $\boldsymbol{u}$, its prior, the mapping $\boldsymbol{X}$, and the noise covariance $\boldsymbol{\Psi}$. We can use the *tower formulae* to compute the marginal distribution. Here is the proof for the covariance. Let $\boldsymbol{v} = \mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]$. Then,

$$\begin{aligned}\mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]] &= \mathrm{E}\left[\mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T|\boldsymbol{u}] - \boldsymbol{v}\boldsymbol{v}^T\right] \stackrel{*}{=} \mathrm{E}[\boldsymbol{y}\boldsymbol{y}^T] - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] = \mathrm{Cov}[\boldsymbol{y}] + \mathrm{E}[\boldsymbol{y}]\mathrm{E}[\boldsymbol{y}]^T - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] \\ &\stackrel{*}{=} \mathrm{Cov}[\boldsymbol{y}] + \mathrm{E}[\boldsymbol{v}]\mathrm{E}[\boldsymbol{v}]^T - \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^T] = \mathrm{Cov}[\boldsymbol{y}] - \mathrm{Cov}[\mathrm{E}[\boldsymbol{y}|\boldsymbol{u}]].\end{aligned}$$

Here , we used the tower formulae for expectation at each point "*" . For the linear -Gaussian model, $\mathrm{E}[\boldsymbol{y}]=\mathrm{E}[\boldsymbol{X}\boldsymbol{u}]=\boldsymbol{X}\boldsymbol{\mu}_0$, while

$$\mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{X}\boldsymbol{u}] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{y}|\boldsymbol{u}]] = \mathrm{Cov}[\boldsymbol{X}\boldsymbol{u}] + \mathrm{E}[\mathrm{Cov}[\boldsymbol{\varepsilon}]] = \boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi}.$$

Another way is to obtain the joint distribution by using

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{u} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\varepsilon} \end{bmatrix}$$

together with what we know about linear transforms of Gaussians. Note that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are independent. For example, $\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}] = \boldsymbol{\Sigma}_0\boldsymbol{X}^T$.

Let us compute the posterior $P(\boldsymbol{u}|\boldsymbol{y})$ for this model, along the two different ways we derived above. We already know it must be Gaussian. Moreover,

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}] &= \mathrm{Cov}[(\boldsymbol{u}, \boldsymbol{y})]/\mathrm{Cov}[\boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{u}] - \mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]\mathrm{Cov}[\boldsymbol{y}]^{-1}\mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]^T \\ &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi})^{-1}\boldsymbol{X}\boldsymbol{\Sigma}_0.\end{aligned}$$

==>

$$\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}] = \mathrm{E}[\boldsymbol{u}] + \mathrm{Cov}[\boldsymbol{u}, \boldsymbol{y}]\mathrm{Cov}[\boldsymbol{y}]^{-1}(\boldsymbol{y} - \mathrm{E}[\boldsymbol{y}]) = \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_0\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T + \boldsymbol{\Psi})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_0).$$

We can also match terms in the joint distribution $P(\boldsymbol{y}, \boldsymbol{u}) = P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})$. ==>

$$\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}]=(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}+{\boldsymbol{\Sigma}_0}^{-1})^{-1}, \mathrm{E}[\boldsymbol{u}|\boldsymbol{y}]=(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{X}+{\boldsymbol{\Sigma}_0}^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{\Psi}^{-1}\boldsymbol{y}+{\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_0).$$

An important take-home message for the linear-Gaussian model is as follows. Suppose that $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$. Then, you can always compute posterior quantities by doing expensive (superlinear) computations, such as inverses, in the *smaller* number only: $\min\{m, n\}$. The main vehicle to formaly get from one set of expression to the other is the Woodbury formula.