Probability Distributions

Spring 2024

Instructor: Shandian Zhe
zhe@cs.utah.edu
School of Computing

THE UNIVERSITY OF UTAH

Outline

- Maximum likelihood estimation (MLE), Maximum A posterior estimation (MAP)
- Probability distributions
  - Binomial, multinomial
  - Beta, Dirichlet
  - Gaussian, student t
  - (inverse) Gamma, (inverse) Wishart

Outline

- Maximum likelihood estimation (MLE), Maximum A posterior estimation (MAP)
- Probability distributions
  - Binomial, multinomial
  - Beta, Dirichlet
  - Gaussian, student t
  - (inverse) Gamma, (inverse) Wishart

Maximum likelihood estimation (MLE)

Suppose we have a distribution $p(\mathbf{x}|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$

We have observed a set of Independent and identically distributed (IID) random variables from $p(\mathbf{x}|\boldsymbol{\theta})$

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$$ observations

How do we estimate $\boldsymbol{\theta}$ from $\mathcal{D}$?

Maximum likelihood estimation (MLE)

The probability density (or mass) evaluated at each observation is called the "likelihood" of the observation

We want to find $\boldsymbol{\theta}$ that maximizes the likelihood of all the observations

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta})$$

Log-likelihood

Maximum a posterior estimation (MAP)

- What is the problem of MLE?

We are in the Bayesian world! We always have some prior knowledge about $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \boxed{p(\boldsymbol{\theta})} \cdot \prod_{i=1}^{n} p(\mathbf{x}_i | \boldsymbol{\theta})$$

prior

$$\boldsymbol{\theta}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \boxed{\log p(\boldsymbol{\theta})} + \sum_{i=1}^{n} \log p(\mathbf{x}_i | \boldsymbol{\theta})$$

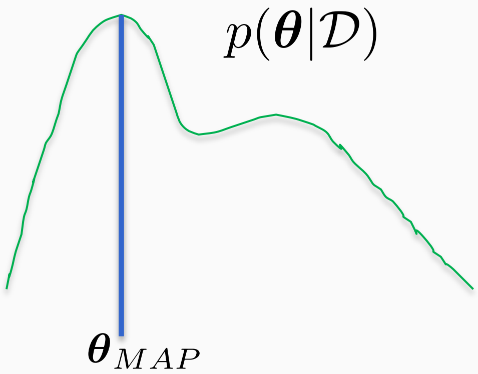
Corresponds to the regularizer in non-Bayesian view

Be aware

- Although MAP looks a good way to incorporate the prior knowledge, it is not ideal in Bayesian (probabilistic) perspective

Goal:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$



$\boldsymbol{\theta}_{MAP}$ is just the **mode** of the posterior distribution

Outline

- Maximum likelihood estimation (MLE), Maximum A posterior estimation (MAP)
- Probability distributions
  - Binomial, multinomial
  - Beta, Dirichlet
  - Gaussian, student t
  - (inverse) Gamma, (inverse) Wishart

Let’ s review commonly used probability distributions

- They are used everywhere – all kinds of statistical (Bayesian or non-Bayesian) applications
- They are building blocks to construct more complex probabilistic models

Like 1+1=2, you should be very familiar with them!

Binary variables

- Consider a binary random variable $x \in \{0, 1\}$

  e.g., toss a coin, buy or not buy

Bernoulli distribution: $p(x = 1) = \mu$

$$p(x) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$
$$\mathrm{var}[x] = \mu(1 - \mu)$$

Binary variables - MLE

- Suppose we have *N* IID observations $\mathcal{D} = \{x_1, \ldots, x_N\}$,

  what is the MLE of μ ?

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\Downarrow$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

$$\Downarrow$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Ratio of 1s

Binary variables

- Binomial distribution: suppose I toss a coin for N times, what is the number of heads?

Repeat Bernoulli experiments N times

If $x \sim \mathrm{Bin}(N, \mu)$, $x \in \{0, 1, 2, \ldots, N\}$

$$p(x) = \left(\begin{array}{c} N \\ x \end{array} \right) \mu^x (1-\mu)^{N-x}$$

$$\left(\begin{array}{c} N \\ x \end{array} \right) = \frac{N!}{(N-x)!x!}$$

Binary variables

- Binomial distribution: how to compute the expectation and variance?

$$\mathbb{E}[x] = N\mu$$

$$\mathrm{var}[x] = N\mu(1-\mu)$$

Trick: represent x as a summation of Bernoulli variables!

Categorical variables

- Suppose a random variable can take *K* values (K>= 2). We call it a categorical (or discrete) variable.
- We use a K-dimensional vector with only **one nonzero** entry (i.e., 1) to represent a sample of categorical variable.

$$\mathbf{x} = [x_1, \ldots, x_K]^\top$$

only one entry can be 1, others=0

- e.g., K = 4, the variable observed as category 2

$$\mathbf{x} = [0, 1, 0, 0]^\top$$

Also called one-hot encoding

Categorical variables

- The distribution of a categorical variable is

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$$

Note each x_k is either 0 or 1

Only one x_k is 1

Note: we have constraints on the parameter $\boldsymbol{\mu}$

$$\mu_k \geq 0 \qquad \sum_{k=1}^{K} \mu_k = 1$$

Categorical variables - MLE

- Consider we have N IID observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k} \qquad m_k = \sum_n x_{nk}$$

Log likelihood

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

Lagrange multiplier: why?

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

Ratio of each category

Categorical variables

- Multinomial distribution: the distribution of the counts of the *K* categories in *N* IID observations:

$$\mathbf{m} = [m_1, \ldots, m_K]^\top \sim \mathrm{Mult}(N, \boldsymbol{\mu})$$

$$p(\mathbf{m}|N, \boldsymbol{\mu}) = \left(\begin{array}{c} N \\ m_1 m_2 \ldots m_K \end{array} \right) \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\sum_{k=1}^{K} m_k = N \qquad \qquad \binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$$

Link categorical variables to ML models (we will discuss them later)

- Key: how to model the parameters μ or $\boldsymbol{\mu}$

  in terms of features $\boldsymbol{\alpha}$
  - Logistic regression

$$\mu = 1/\big(1 + \exp(-\mathbf{w}^\top \boldsymbol{\alpha})\big)$$

  - Probit regression

$$\mu = \mathrm{GaussianCDF}(\mathbf{w}^\top \boldsymbol{\alpha})$$

  - Multi-class classification

$$\mu_k = \frac{\exp(\mathbf{w}_k^\top \boldsymbol{\alpha})}{\sum_j \exp(\mathbf{w}_j^\top \alpha)}$$

  - Ordinal regression

$$\mu_k = \int_{b_{k-1}}^{b_k} \mathcal{N}(t|\mathbf{w}^\top \boldsymbol{\alpha}, 1)\mathrm{d}t$$

Distribution of discrete distributions

- A Bernoulli distribution is determined by $\mu \in [0, 1]$

$$p(x) = \mu^x (1-\mu)^{1-x}$$

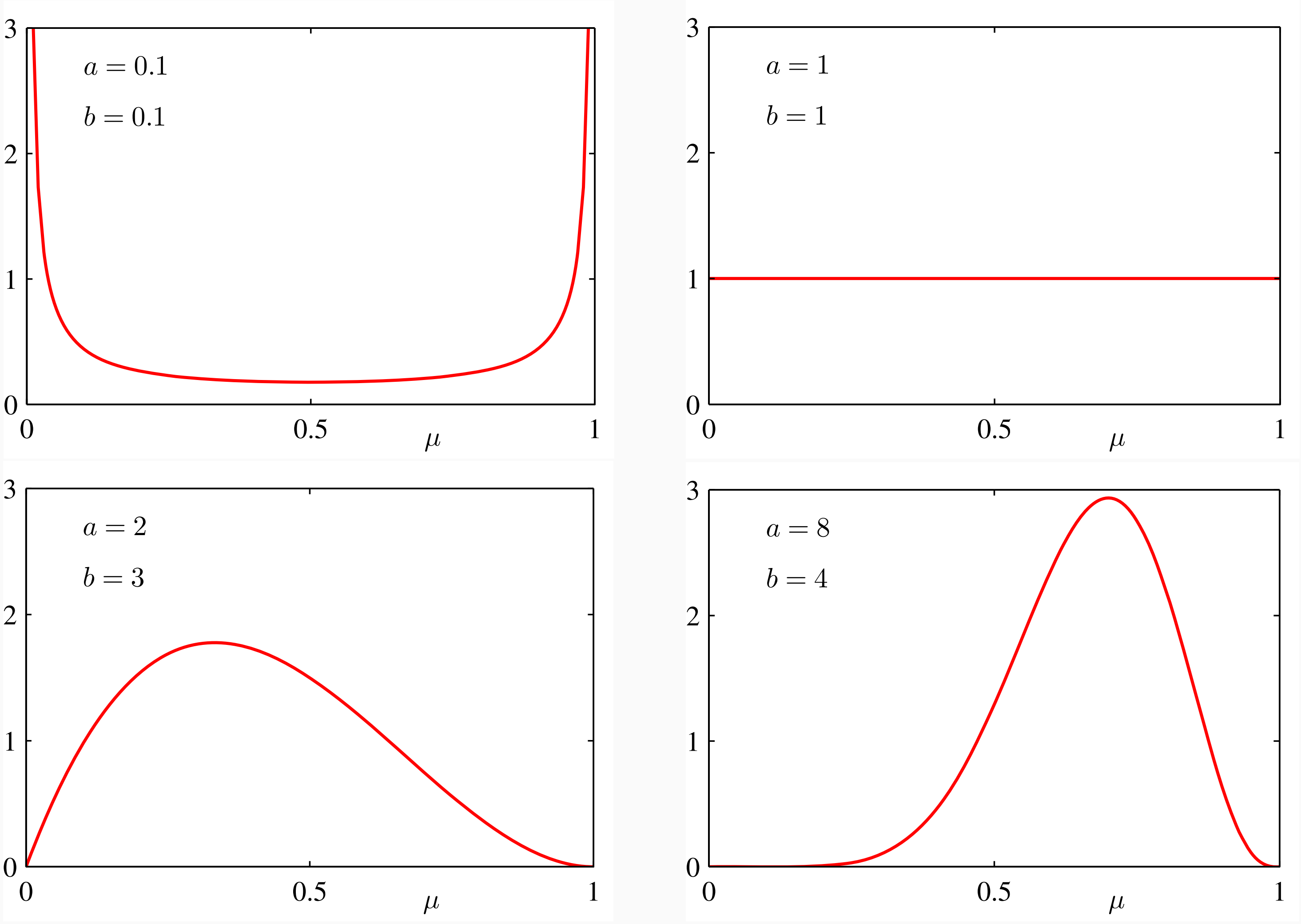
- Can we have a distribution over μ ? Beta distribution

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

$\Gamma(a)$: The general version of $(a-1)!$, a can be continuous

$$\Gamma(1) = 1 \qquad \Gamma(a) = (a-1)\Gamma(a-1)$$

Beta distribution with different a,b



Beta distribution

$$
\begin{aligned}
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

Beta distribution is a conjugate prior to the Bernoulli likelihood. We will discuss it later.

Distribution of discrete distributions

- A Categorical distribution is determined by

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$$

$$\mu_k \geq 0 \qquad \sum_{k=1}^{K} \mu_k = 1$$

- Can we have a distribution over $\boldsymbol{\mu}$? Dirichlet distribution

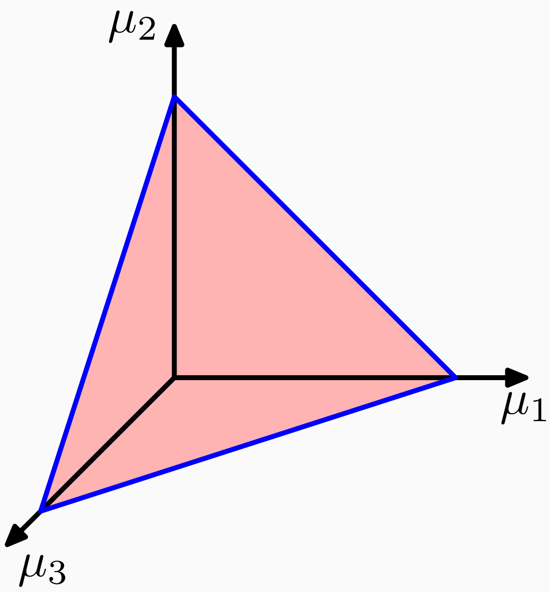
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]^\top$ are called concentration parameters

Each $\alpha_k > 0$

Dirichlet distribution: distribution over simplexes

The Dirichlet distribution over three variables μ_1, μ_2, μ_3 is confined to a simplex (a bounded linear manifold) of the form shown, as a consequence of the constraints $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$.



Beta dist. is a special case of Dirichlet dist. when K=2

Dirichlet distribution

$$\mathbb{E}[\mu_k] = \frac{a_k}{\sum_{j=1}^{K} a_j}$$

$$\mathbb{E}[\log \mu_k] = \psi(\alpha_k) - \psi(\sum_{j=1}^{K} \alpha_j)$$

Dirichlet distribution

$$\mathbb{E}[\mu_k] = \frac{a_k}{\sum_{j=1}^{K} a_j}$$

$$\mathbb{E}[\log \mu_k] = \psi(\alpha_k) - \psi(\sum_{j=1}^{K} \alpha_j)$$

digamma function $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log\left(\Gamma(x)\right) = \frac{\Gamma'(x)}{\Gamma(x)}$

Dirichlet distribution is a conjugate prior to the categorical likelihood. We will discuss it later.

Latent Dirichlet allocation (LDA)

[Blei et. al. 03]

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Continuous variables

- Gaussian distribution

Everybody knows the single-variable case

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Multivariate Gaussian distribution

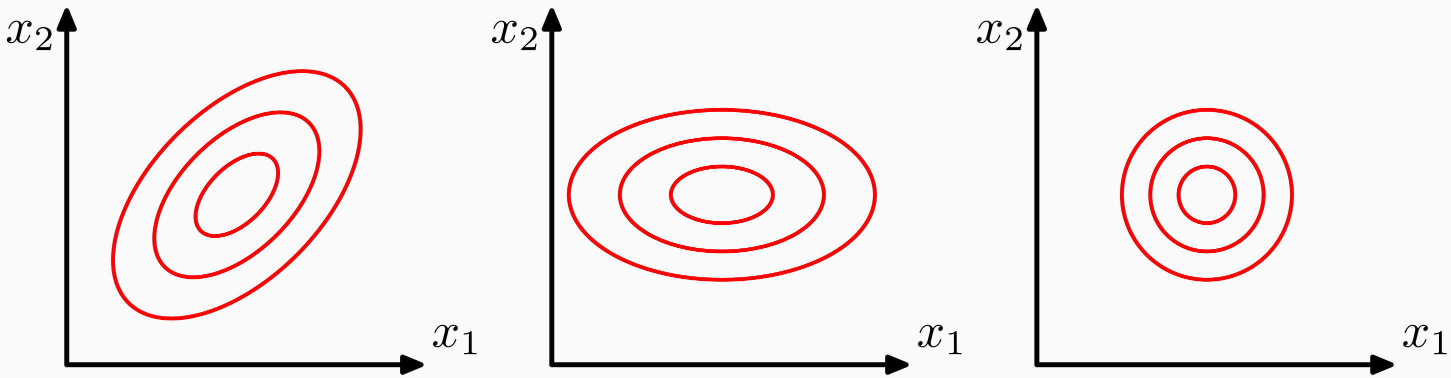
- We need to be familiar the multivariate (general) case

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(-\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}_{\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}\right)}\big)$$

$\boldsymbol{\mu}$:mean $\qquad \boldsymbol{\Sigma} \succ 0$:covariance matrix

Sometimes we use $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$, which is called precision matrix

Contours of 2-D Gaussian



(a) (b) (c)

covariance general diagonal identity

Multivariate Gaussian distribution - MLE

- The key fact $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$
- Given IID observations $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

  The variable is *d* dimensional

$$\log\left(p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Sufficient statistics $\displaystyle\sum_{n=1}^{N}\mathbf{x}_n,$ $\displaystyle\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}.$

Multivariate Gaussian distribution - MLE

$$\log\big(p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

set $$\frac{\partial \log\big(p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\big)}{\partial \boldsymbol{\mu}} = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = \mathbf{0}$$

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

Multivariate Gaussian distribution - MLE

$$\log\big(p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma})\big) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

$$\frac{\partial\log\big(p(\mathcal{D}|\boldsymbol{\mu}_{\mathrm{ML}},\boldsymbol{\Sigma})\big)}{\partial\boldsymbol{\Sigma}} = -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^\top\boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

It is semi-positive definite

Multivariate Gaussian distribution - MLE

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\
\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}
\end{aligned}
$$

Why?

Multivariate Gaussian distribution - MLE

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \boxed{\frac{N-1}{N}\boldsymbol{\Sigma}}$$

Biased estimate

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

Unbiased estimate

Partitioned Gaussian

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Question 1: What is $p(\mathbf{x}_a|\mathbf{x}_b)$?

Conditional Gaussian distribution

- We need to use the "completing the square" trick

The exponent of a general Gaussian distribution is

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\underbrace{\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}}_{\text{Quadratic term}} + \underbrace{\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}_{\text{Linear term}} + \text{const}$$

Conditional Gaussian distribution

- Let us expand the partitioned variables

$$
\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
&-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
&-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a)-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(\mathbf{x}_a|\mathbf{x}_b)$

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
&\qquad -\frac{1}{2}\boxed{(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)} - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
&\qquad -\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

Quadratic term $$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a$$

Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(\mathbf{x}_a|\mathbf{x}_b)$

$$
\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
-\frac{1}{2}\boxed{(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a)} - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

Quadratic term $-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a \Rightarrow \boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$

Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(\mathbf{x}_a|\mathbf{x}_b)$

$$
\begin{aligned}
-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \\
-\frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b) \\
-\frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{ba}(\mathbf{x}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{x}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Lambda}_{bb}(\mathbf{x}_b-\boldsymbol{\mu}_b).
\end{aligned}
$$

Linear term: $\mathbf{x}_a^{\mathrm{T}}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b)\right\}$

Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(\mathbf{x}_a|\mathbf{x}_b)$

Linear term: $\mathbf{x}_a^{\mathrm{T}} \boxed{\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\}}$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boxed{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} + \text{const}$$

Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(\mathbf{x}_a|\mathbf{x}_b)$

Linear term: $\mathbf{x}_a^{\mathrm{T}} \boxed{\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\}}$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boxed{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} + \text{const}$$

$$\begin{aligned}
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

Conditional Gaussian distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

Conditional Gaussian distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\end{aligned}$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Conditional Gaussian distribution

- Block matrix inverse

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$$

Conditional Gaussian distribution

- Block matrix inverse

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}$$

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix} \Rightarrow \begin{aligned} \boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1} \end{aligned}$$

Conditional Gaussian distribution

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

$$\begin{aligned}\boldsymbol{\Lambda}_{aa} &= (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba})^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\end{aligned}$$

$$\boxed{\begin{aligned}\boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}.\end{aligned}}$$

Marginal Gaussian distribution

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

Question 2: What is $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b$?

Marginal Gaussian distribution

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

Use the same trick, we can derive that

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$$

Leave it as your exercise

Gamma distribution

A scalar Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Gamma distribution

A scalar Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Do we have a distribution over the precision? $\lambda = 1/\sigma^2$ $\lambda > 0$

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

Gamma distribution

A scalar Gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Do we have a distribution over the precision? $\lambda = 1/\sigma^2 \qquad \lambda > 0$

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \qquad a > 0, b > 0$$

$$\begin{aligned} \mathbb{E}[\lambda] &= \frac{a}{b} \\ \mathrm{var}[\lambda] &= \frac{a}{b^2} \end{aligned}$$

Gamma distribution

A scalar Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

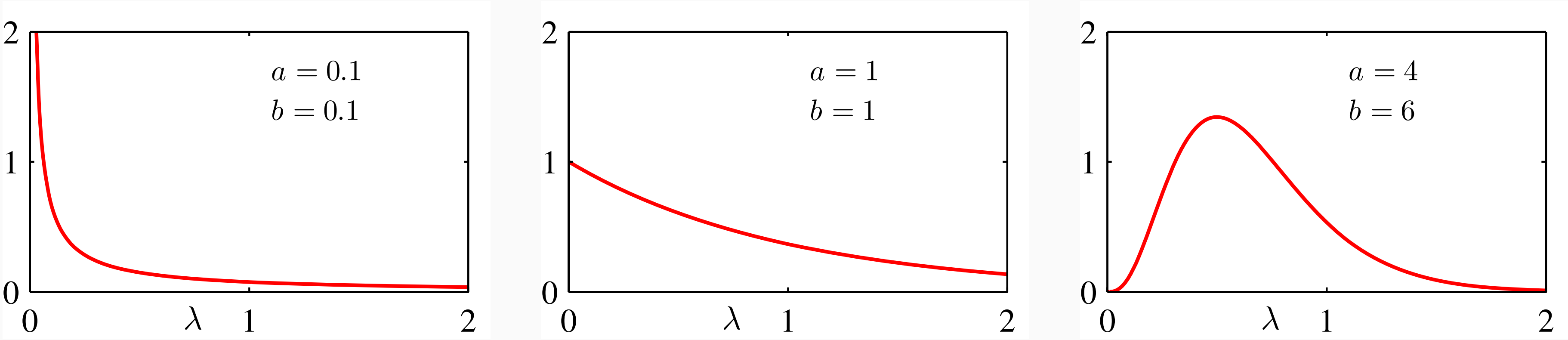
Do we have a distribution over the precision? $\lambda = 1/\sigma^2$ $\lambda > 0$

$$\mathrm{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \quad a > 0, b > 0$$

$$\begin{aligned} \mathbb{E}[\lambda] &= \frac{a}{b} \\ \mathrm{var}[\lambda] &= \frac{a}{b^2} \end{aligned}$$

$$\mathbb{E}[\log(\lambda)] = \psi(\mathbf{a}) - \log(b)$$ [digamma function]

Gamma distribution



Inverse Gamma distribution

$$\lambda \sim \mathrm{Gamma}(\lambda|a,b)$$

$$\lambda^{-1} \sim \mathrm{InvGamma}(\lambda|a,b)$$

Inverse Gamma distribution

$$\lambda \sim \mathrm{Gamma}(\lambda|a, b)$$

$$\lambda^{-1} \sim \mathrm{InvGamma}(\lambda|a, b)$$

Inverse Gamma distribution is often used as a prior distribution over the Gaussian variance

$$\mathcal{N}(x|\mu, \boxed{\sigma^2}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

Wishart Distribution

- Now let us switch to multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = |2\pi\mathbf{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x})$$

Do we have a distribution over the **precision matrix** $\mathbf{\Lambda} \equiv \mathbf{\Sigma}^{-1}$?

Wishart Distribution

- Now let us switch to multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x})$$

Do we have a distribution over the **precision matrix** $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$?

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = \frac{|\boldsymbol{\Lambda}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\nu/2}\Gamma_d(\frac{\nu}{2})}$$

$\mathbf{W} \succ \mathbf{0}$ $\quad$ $\nu > d-1$

degree of freedom

Wishart Distribution

- Now let us switch to multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x})$$

Do we have a distribution over the **precision matrix** $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$?

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = \frac{|\boldsymbol{\Lambda}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)}{2^{\frac{d\nu}{2}} |\mathbf{W}|^{\nu/2} \underbrace{\Gamma_d(\frac{\nu}{2})}_{\text{multivariate gamma function}}}$$

$\mathbf{W} \succ \mathbf{0}$ $\quad \nu > d - 1$

degree of freedom

Wishart Distribution

- Now let us switch to multivariate Gaussian distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}\mathbf{x})$$

Do we have a distribution over the **precision matrix** $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$?

$$\mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, \nu) = \frac{|\boldsymbol{\Lambda}|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)}{2^{\frac{d\nu}{2}}|\mathbf{W}|^{\nu/2}\underbrace{\Gamma_d(\frac{\nu}{2})}_{\text{multivariate gamma function}}}$$

$\mathbf{W} \succ \mathbf{0}$ $\quad \nu > d - 1$

degree of freedom

Multi-dimensional version of Gamma distribution!

Inverse Wishart Distribution

$$\mathbf{\Lambda} \sim \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu)$$

$$\mathbf{\Lambda}^{-1} \sim \mathcal{W}^{-1}(\mathbf{\Lambda}|\mathbf{W}^{-1}, \nu)$$

Inverse Wishart Distribution

$$\mathbf{\Lambda} \sim \mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu)$$

$$\mathbf{\Lambda}^{-1} \sim \mathcal{W}^{-1}(\mathbf{\Lambda}|\mathbf{W}^{-1}, \nu)$$

Inverse Wishart distribution is often used as a prior distribution over the covariance matrixs of the multivariate Gaussian dist.

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boxed{\mathbf{\Sigma}}) = |2\pi\mathbf{\Sigma}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x})$$

Student t's distribution

- Infinite mixture of Gaussian distribution

Suppose we have a Gaussian random variable $p(x|\mu,\tau) = \mathcal{N}(x|\mu,\tau^{-1})$

If we place a Gamma prior distribution over the precision τ

$$p(\tau|a,b) = \mathrm{Gamma}(\tau|a,b)$$

What is the marginal distribution of x ?

$$p(x|\mu,a,b) = \int_0^\infty p(x|\mu,\tau)p(\tau|a,b)\mathrm{d}\tau$$

Student t's distribution

$$
\begin{aligned}
p(x|\mu, a, b) &= \int_0^{\infty} \mathcal{N}(x|\mu, \tau^{-1}) \mathrm{Gam}(\tau|a, b)\, \mathrm{d}\tau \\
&= \int_0^{\infty} \frac{b^a e^{(-b\tau)} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} \mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a + 1/2)
\end{aligned}
$$

Student t's distribution

$$
\begin{aligned}
p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)\,\mathrm{d}\tau \\
&= \int_0^\infty \frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} \mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a+1/2)
\end{aligned}
$$

$$\nu = 2a \qquad \lambda = a/b.$$

Student t's distribution

Infinite weighted sum of Gaussians!

$$
\begin{aligned}
p(x|\mu, a, b) &= \boxed{\int_0^\infty \mathcal{N}(x|\mu, \tau^{-1})\mathrm{Gam}(\tau|a, b)\,\mathrm{d}\tau} \\
&= \int_0^\infty \frac{b^a e^{(-b\tau)}\tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\}\,\mathrm{d}\tau \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a+1/2)
\end{aligned}
$$

$$\nu = 2a \qquad \lambda = a/b.$$

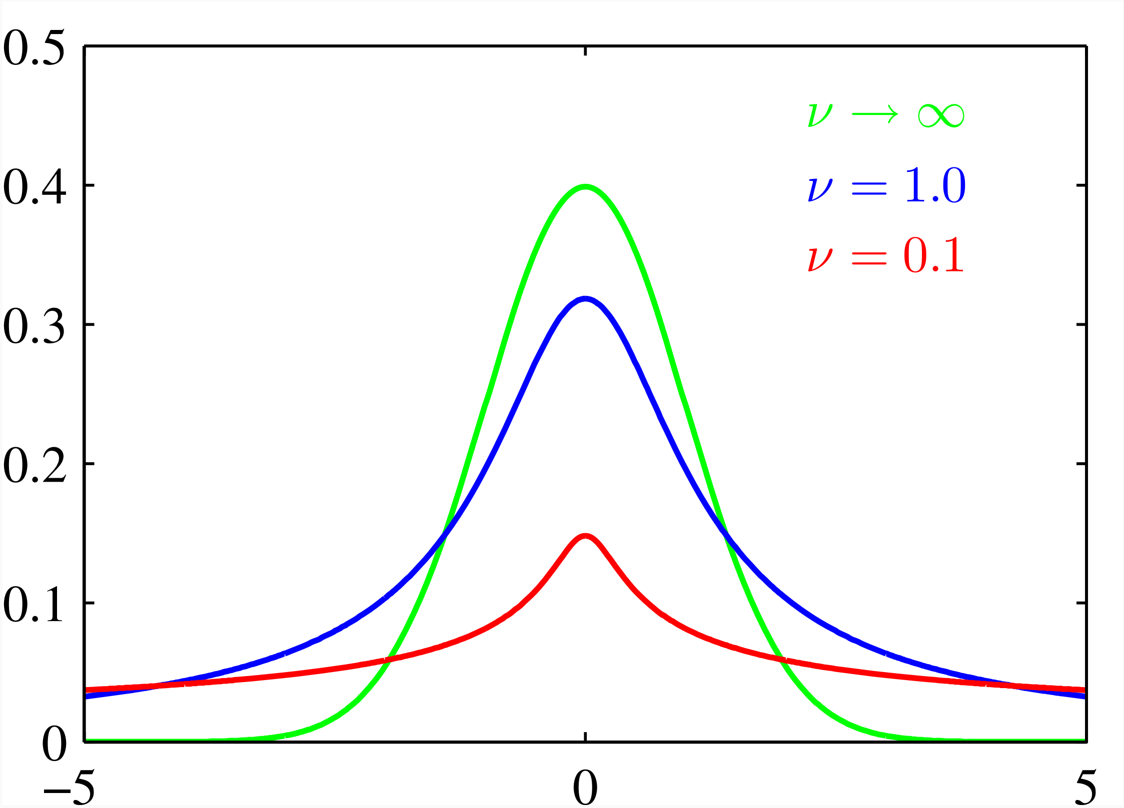
$$\mathrm{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

mean

precision

degree of freedom $\nu > 0$

Student t's distribution – heavy tail



$\nu \to \infty$ ⇒ $\mathrm{St}(x|\boldsymbol{\mu}, \lambda, \nu) \to \mathcal{N}(x|\boldsymbol{\mu}, \lambda^{-1})$

Student t's distribution - robustness

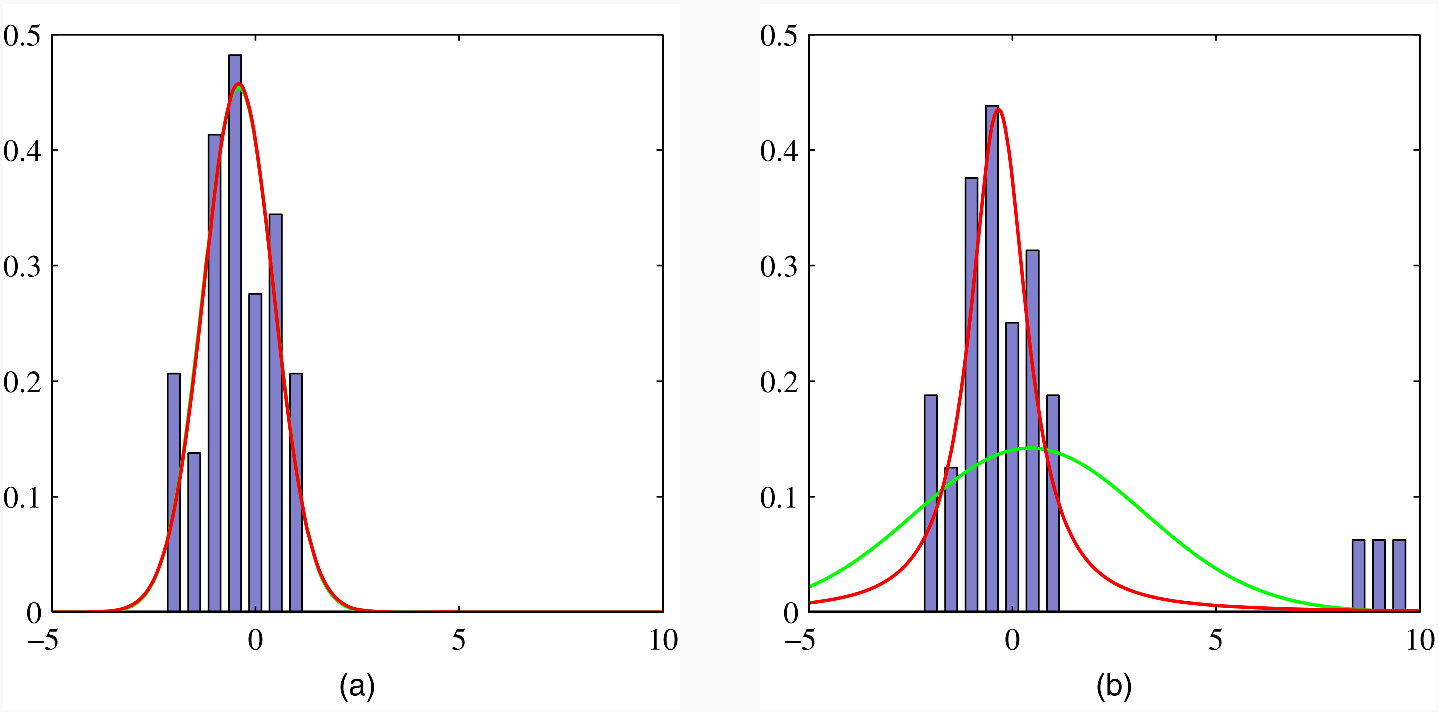


Figure 2.16 Illustration of the robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden by the red curve). Because the t-distribution contains the Gaussian as a special case it gives almost the same solution as the Gaussian. (b) The same data set but with three additional outlying data points showing how the Gaussian (green curve) is strongly distorted by the outliers, whereas the t-distribution (red curve) is relatively unaffected.

Student t's distribution

$$p(x|\mu, a, b) = \int_0^\infty p(x|\mu, \tau)p(\tau|a, b)\mathrm{d}\tau$$

Student t's distribution

$$p(x|\mu, a, b) = \int_0^{\infty} p(x|\mu, \tau)p(\tau|a, b)\mathrm{d}\tau$$

$$\nu = 2a, \lambda = a/b, \eta = \tau b/a$$

Student t's distribution

$$p(x|\mu, a, b) = \int_0^\infty p(x|\mu, \tau) p(\tau|a, b) \mathrm{d}\tau$$

$$\nu = 2a, \lambda = a/b, \eta = \tau b/a$$

$$\boxed{\mathrm{St}(x|\mu, \lambda, \nu) = \int_0^\infty \mathcal{N}\left(x|\mu, (\eta\lambda)^{-1}\right) \mathrm{Gam}(\eta|\nu/2, \nu/2)\, \mathrm{d}\eta}$$

Multivariate student-t distribution

$$\mathrm{St}(x|\mu, \lambda, \nu) = \int_0^\infty \mathcal{N}\left(x|\mu, (\eta\lambda)^{-1}\right) \mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\eta$$

$$\downarrow$$

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, (\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2, \nu/2)\,\mathrm{d}\eta$$

Multivariate student-t distribution

$$\mathrm{St}(x|\mu,\lambda,\nu) = \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta$$

$$\downarrow$$

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},(\eta\boldsymbol{\Lambda})^{-1})\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta$$

$$\downarrow$$

$$\mathrm{St}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda},\nu) = \frac{\Gamma(d/2+\nu/2)}{\Gamma(\nu/2)}\frac{|\boldsymbol{\Lambda}|^{1/2}}{(\pi\nu)^{d/2}}[1+\frac{1}{\nu}(\mathbf{x}-\mu)^\top\boldsymbol{\Lambda}(\mathbf{x}-\mu)]^{-d/2-\nu/2}$$

Multivariate student-t distribution

$$\mathbf{x} \sim \mathrm{St}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu)$$

$$\begin{aligned} \mathbb{E}[\mathbf{x}] &= \boldsymbol{\mu}, && \text{if} \quad \nu > 1 \\ \mathrm{cov}[\mathbf{x}] &= \frac{\nu}{(\nu-2)}\boldsymbol{\Lambda}^{-1}, && \text{if} \quad \nu > 2 \\ \mathrm{mode}[\mathbf{x}] &= \boldsymbol{\mu} \end{aligned}$$

Ding, Peng. "On the conditional distribution of the multivariate t distribution." *The American Statistician* 70.3 (2016): 293-295.

Conditional distribution

Shah, Amar, Andrew Wilson, and Zoubin Ghahramani. "Student-t processes as alternatives to Gaussian processes." *Artificial intelligence and statistics*. 2014.

What you need to know

- The commonly used distributions for binary, categorical, continuous random variables
- For multi-variate Gaussian distribution, know how to derive the conditional distribution and marginal distribution
- The commonly used prior distribution of the distribution parameters (Gamma, Beta, Dirichlet…)
- Know how the student t distribution is derived and its heavy tail property.