

Lecture 3

Probability Distributions

Instructor: Shibo Li

shiboli@cs.fsu.edu



Department of Computer Science
Florida State University

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

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 $p(\mathbf{x}|\boldsymbol{\theta})$

$$\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \quad \text{observations}$$

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

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

We want to find θ that maximizes the likelihood of all the observations

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

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

- What is the problem of MLE?

We are in the Bayesian world! We always have some prior knowledge about θ

$$\theta_{MAP} = \operatorname{argmax}_{\theta} p(\theta) \cdot \prod_{i=1}^n p(\mathbf{x}_i | \theta) \quad \text{prior}$$

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



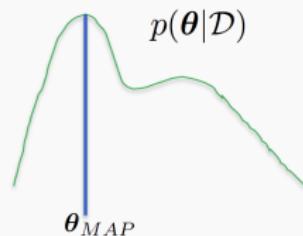
Corresponds to the regularizer in non-Bayesian view

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

Goal:

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

θ_{MAP} is just the **mode** of the posterior distribution



- 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!

- Consider a binary random variable $x \in \{0, 1\}$
e.g., toss a coin, buy or not buy

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

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

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

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

- Suppose we have N IID observations $\mathcal{D} = \{x_1, \dots, x_N\}$
what is the MLE of μ ?

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



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



$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad \text{Ratio of 1s}$$

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

Repeat Bernoulli experiments N times

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

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

$$\binom{N}{x} = \frac{N!}{(N - x)!x!}$$

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

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

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

Trick: represent x as a summation of Bernoulli variables!

- Suppose a random variable can take K values ($K \geq 2$). 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.

$$\mathbf{x} = [x_1, \dots, x_K]^\top \quad \text{only one entry can be 1, others=0}$$

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

$$\mathbf{x} = [0, 1, 0, 0]^\top \quad \text{Also called one-hot encoding}$$

- The distribution of a categorical variable is

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

Note each x_k is either 0 or 1

Only one x_k is 1

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

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

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

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

Log likelihood

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

Lagrange multiplier: why?

$$\mu_k^{\text{ML}} = \frac{m_k}{N} \quad \text{Ratio of each category}$$

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

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

$$p(\mathbf{m}|N, \boldsymbol{\mu}) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}$$

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

- Key: how to model the parameters μ or α

in terms of features α

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

- Probit regression

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

- Multi-class classification

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

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

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

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

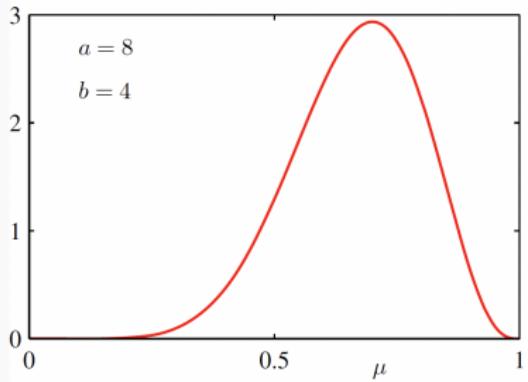
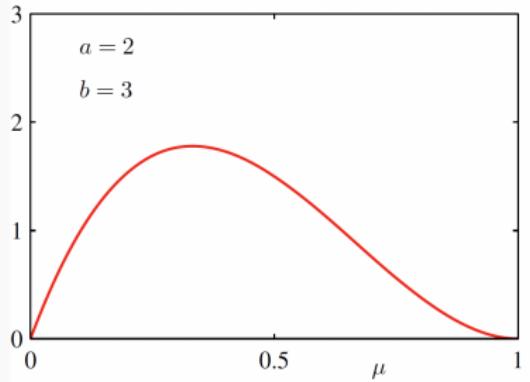
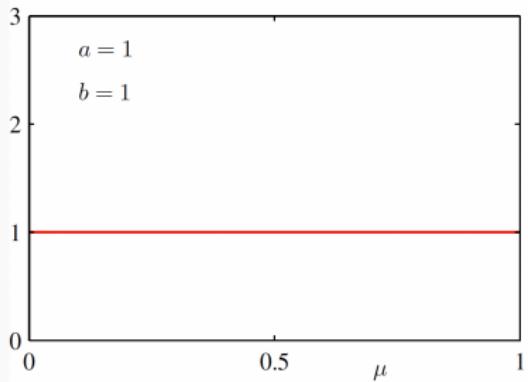
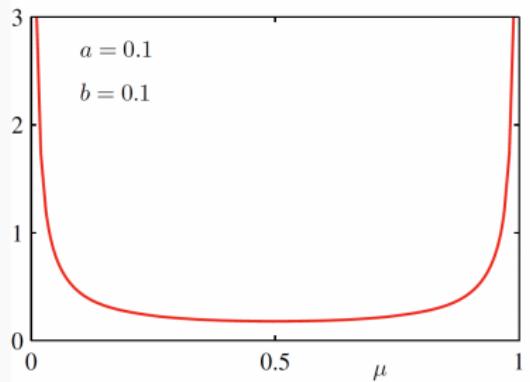
- Can we have a distribution over μ ? Beta distribution

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

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

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

Beta Distribution with Different a, b



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

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

$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^\top$$

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

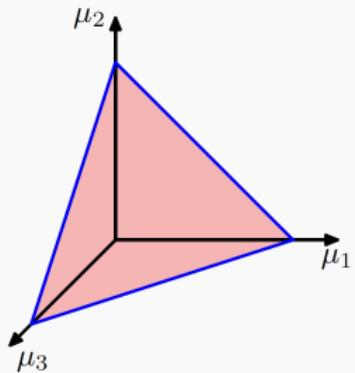
- Can we have a distribution over $\boldsymbol{\mu}$? **Dirichlet distribution**

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

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

Each $\alpha_k > 0$

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 $0 \leq \mu_k \leq 1$ and $\sum_k \mu_k = 1$.



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

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

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

digamma function

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$$

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.

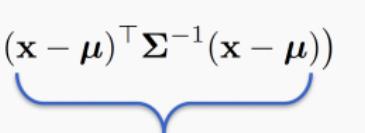
- Gaussian distribution

Everybody knows the single-variable case

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

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

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

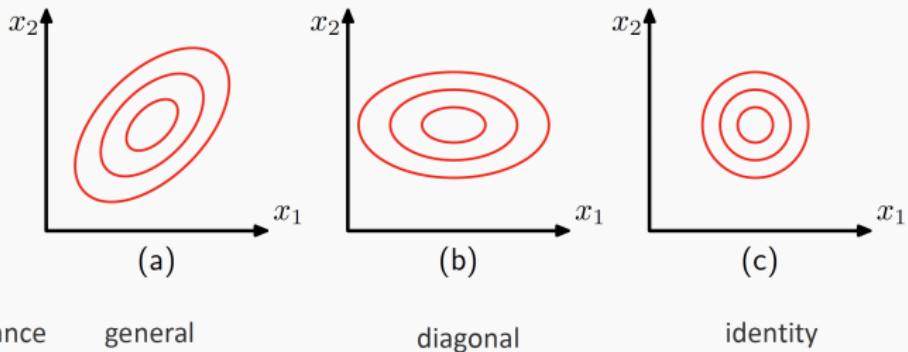


$$\text{tr}((\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1})$$

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

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

Contours of 2-D Gaussian



covariance

general

diagonal

identity

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

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

Sufficient statistics

$$\sum_{n=1}^N \mathbf{x}_n,$$

$$\sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T.$$

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

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



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

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



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



$$\boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}}) (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^\top \quad \text{It is semi-positive definite}$$

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

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

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

Biased estimate

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\text{ML}})^T \quad \text{Unbiased estimate}$$

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

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

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

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

- We need to use the “completing the square” trick

The exponent of a general Gaussian distribution is

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

Quadratic term Linear term

- Let us expand the partitioned variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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



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

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

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

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

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

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

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

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

- Block matrix inverse

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

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

- Block matrix inverse

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

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

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

Conditional Gaussian Distribution

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

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

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



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

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

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

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ba}.$$

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

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

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

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

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

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

A scalar Gaussian distribution

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

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? $\lambda = 1/\sigma^2$ $\lambda > 0$

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

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? $\lambda = 1/\sigma^2$ $\lambda > 0$

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

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

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? $\lambda = 1/\sigma^2$ $\lambda > 0$

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

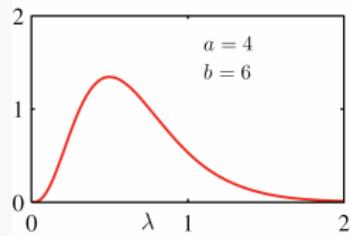
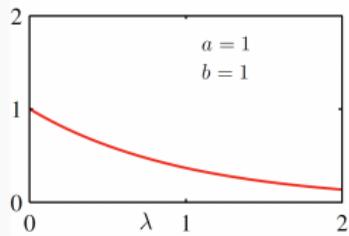
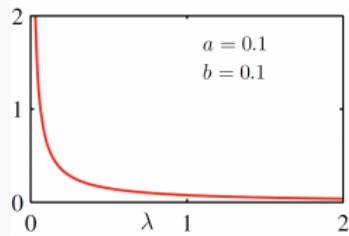
$$\mathbb{E}[\lambda] = \frac{a}{b}$$

$$\text{var}[\lambda] = \frac{a}{b^2}$$

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

digamma function

Gamma Distribution



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



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

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



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

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

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

- Now let us switch to multivariate Gaussian distribution

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

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

- 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** $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$?

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

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

degree of freedom

- Now let us switch to multivariate Gaussian distribution

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

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

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

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

degree of freedom

multivariate gamma function

- Now let us switch to multivariate Gaussian distribution

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

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

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

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

degree of freedom

multivariate gamma function

Multi-dimensional version of Gamma distribution!

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



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

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



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

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

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

- 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

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

What is the marginal distribution of x ?

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

Student t's distribution

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

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

Student t's distribution

Infinite weighted sum of Gaussians!

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

$$\nu = 2a \quad \lambda = a/b$$



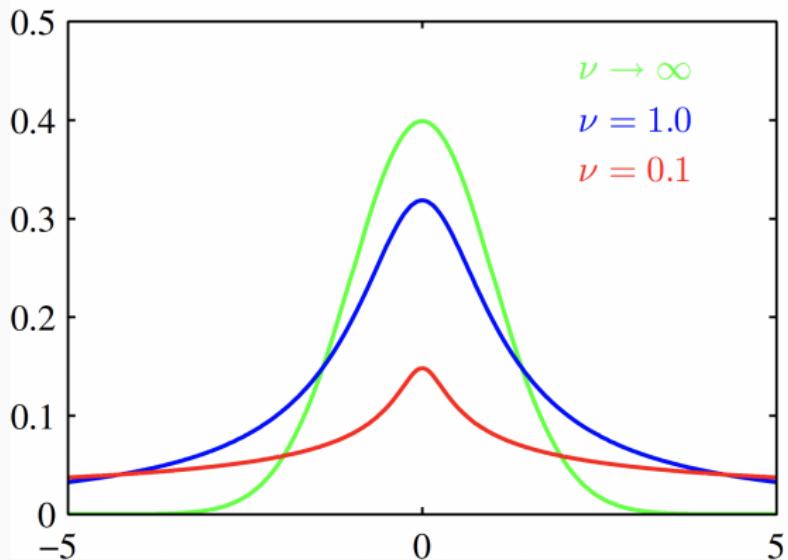
$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

mean

precision

degree of freedom $\nu > 0$

Student t's distribution - heavy tail



$$\nu \rightarrow \infty \quad \xrightarrow{\hspace{1cm}} \quad \text{St}(x|\boldsymbol{\mu}, \lambda, \nu) \rightarrow \mathcal{N}(x|\boldsymbol{\mu}, \lambda^{-1})$$

Student t's distribution - robustness

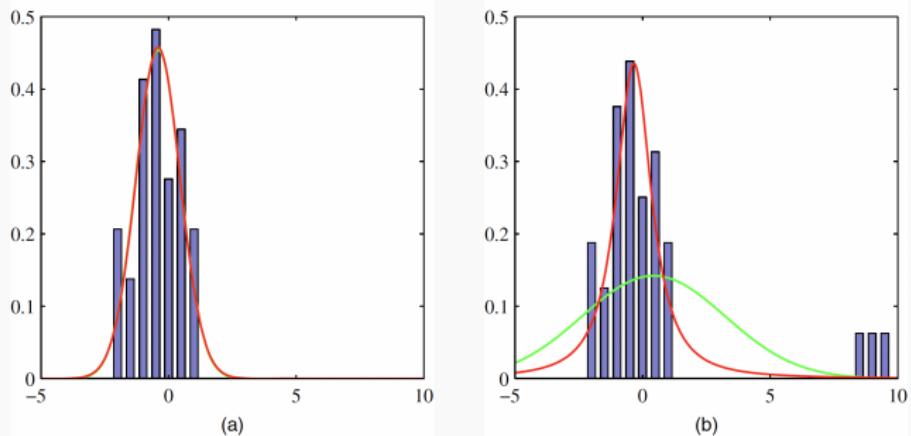


Figure 2.16 Illustration of the robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden 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.

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

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

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

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

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



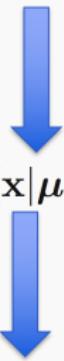
$$\text{St}(x|\mu, \lambda, \nu) = \int_0^{\infty} \mathcal{N}\left(x|\mu, (\eta\lambda)^{-1}\right) \text{Gam}(\eta|\nu/2, \nu/2) d\eta$$

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



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

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



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

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

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

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

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.

- 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.