Probability Distributions

Fall 2019

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

• Maximum likelihood estimation (MLE), Maximum A posterior estimation (MAP)
• Probability distributions
  – Binomial, multinominal
  – 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 \( p(x|\theta) \) parameterized by \( \theta \)

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

\[
\mathcal{D} = \{x_1, \ldots, x_n\} \quad \text{observations}
\]

How do we estimate \( \theta \) 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 $\theta$ that maximizes the likelihood of all the observations.

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i | \theta)$$

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i | \theta)$$

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 $\theta$

$$\theta_{MAP} = \arg\max_{\theta} \log p(\theta) + \sum_{i=1}^{n} \log p(x_i | \theta)$$

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:  
\[ p(\theta|\mathcal{D}) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta) \]

\( \theta_{MAP} \) 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: $p(x = 1) = \mu$

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

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

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

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

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

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

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

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

Repeat Bernoulli experiments $N$ times

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

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

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

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

\[
E[x] = N \mu
\]

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

Trick: represent \( x \) as a summation of Bernoulli variables!
Categorical variables

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

\[ x = [x_1, \ldots, x_K]^\top \]

  only one entry can be 1, others=0

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

\[ x = [0, 1, 0, 0]^\top \]

Also called one-hot encoding
Categorical variables

• The distribution of a categorical variable is

\[ p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k} \quad \mu = (\mu_1, \ldots, \mu_K)^T \]

Note each \( x_k \) is either 0 or 1

Only one \( x_k \) is 1

Note: we have constraints on the parameter \( \mu \)

\[ \mu_k \geq 0 \quad \sum_{k=1}^{K} \mu_k = 1 \]
Categorical variables - MLE

- Consider we have N IID observations $\mathcal{D} = \{x_1, \ldots, x_n\}$

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

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

Log likelihood

Lagrange multiplier: why?

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

Ratio of each category
Categorical variables

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

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

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

$$\sum_{k=1}^{K} m_k = N \quad \binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$$
Link categorical variables to ML models (we will discuss them later)

- Key: how to model the parameters $\mu$ or $\mu$
in terms of features $\alpha$
  - Logistic regression
    $$\mu = 1/(1 + \exp(-w^T \alpha))$$
  - Probit regression
    $$\mu = \text{GaussianCDF}(w^T \alpha)$$
  - Multi-class classification
  - Ordinal regression
    $$\mu_k = \frac{\exp(w_k^T \alpha)}{\sum_j \exp(w_j^T \alpha)}$$
    $$\mu_k = \int_{b_{k-1}}^{b_k} \mathcal{N}(t|w^T \alpha, 1)\,dt$$
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 \( \mu \)? 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

Figure 2.2 Plots of the beta distribution $\text{Beta}(\mu|a, b)$ given by (2.13) as a function of $\mu$ for various values of the hyperparameters $a$ and $b$.

where $l = N - m$, and therefore corresponds to the number of 'tails' in the coin example. We see that (2.17) has the same functional dependence on $\mu$ as the prior distribution, reflecting the conjugacy properties of the prior with respect to the likelihood function. Indeed, it is simply another beta distribution, and its normalization coefficient can therefore be obtained by comparison with (2.13) to give

$$p(\mu|m, l, a, b) = \frac{\Gamma(m + a + l + b)}{\Gamma(m + a) \Gamma(l + b)} \mu^{m + a - 1} (1 - \mu)^{l + b - 1}.$$  

(2.18)

We see that the effect of observing a data set of $m$ observations of $x = 1$ and $l$ observations of $x = 0$ has been to increase the value of $a$ by $m$, and the value of $b$ by $l$, in going from the prior distribution to the posterior distribution. This allows us to provide a simple interpretation of the hyperparameters $a$ and $b$ in the prior as an effective number of observations of $x = 1$ and $x = 0$, respectively. Note that $a$ and $b$ need not be integers. Furthermore, the posterior distribution can act as the prior if we subsequently observe additional data. To see this, we can imagine taking observations one at a time and after each observation updating the current posterior.
2.1. Binary Variables

Given by (2.3) and (2.4), respectively, we have

\[
E[m] \equiv N \sum_{m=0}^N \binom{m}{N, \mu} = N \mu
\]

(2.11)

\[
\text{var}[m] \equiv N \sum_{m=0}^N (m - E[m])^2 \binom{m}{N, \mu} = N \mu (1 - \mu).
\]

(2.12)

These results can also be proved directly using calculus.

Exercise 2.4

2.1.1 The beta distribution

We have seen in (2.8) that the maximum likelihood setting for the parameter \(\mu\) in the Bernoulli distribution, and hence in the binomial distribution, is given by the fraction of the observations in the data set having \(x = 1\). As we have already noted, this can give severely over-fitted results for small data sets. In order to develop a Bayesian treatment for this problem, we need to introduce a prior distribution \(p(\mu)\) over the parameter \(\mu\). Here we consider a form of prior distribution that has a simple interpretation as well as some useful analytical properties. To motivate this prior, we note that the likelihood function takes the form of the product of factors of the form \(\mu x (1 - \mu)^{1-x}\). If we choose a prior to be proportional to powers of \(\mu\) and \((1 - \mu)^{-1-x}\), then the posterior distribution, which is proportional to the product of the prior and the likelihood function, will have the same functional form as the prior. This property is called conjugacy and we will see several examples of it later in this chapter. We therefore choose a prior, called the beta distribution, given by

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

(2.13)

where \(\Gamma(x)\) is the gamma function defined by (1.141), and the coefficient in (2.13) ensures that the beta distribution is normalized, so that

\[
\int_0^1 \text{Beta}(\mu | a, b) d\mu = 1.
\]

(2.14)

The mean and variance of the beta distribution are given by

Exercise 2.5

\[
E[\mu] = \frac{a}{a + b}
\]

(2.15)

\[
\text{var}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)}
\]

(2.16)
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

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

- Can we have a distribution over \( \mu \) ? Dirichlet distribution

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

\( \alpha = [\alpha_1, \ldots, \alpha_K]^T \) are called concentration parameters

Each \( \alpha_k > 0 \)
Dirichlet distribution: distribution over simplexes

The Dirichlet distribution over three variables $\mu_1, \mu_2, \mu_3$ 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 $K=2$
Dirichlet distribution

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

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

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

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

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

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

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

\[ \text{tr}\left((x - \mu)(x - \mu)^\top \Sigma^{-1}\right) \]

\[ \mu : \text{mean} \quad \Sigma \succ 0 : \text{covariance matrix} \]

Sometimes we use \( \Lambda = \Sigma^{-1} \), which is called precision matrix
Contours of 2-D Gaussian

covariance
general
diagonal
identity
• The key fact \( \mathbb{E}[x] = \mu \) \( \mathbb{E}[xx^T] = \mu \mu^T + \Sigma \)

• Given IID observations \( \mathcal{D} = \{x_1, \ldots, x_N\} \)

The variable is \( d \) dimensional

\[
\log(p(\mathcal{D}|\mu, \Sigma)) = -\frac{N}{2}d \log(2\pi) - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)
\]

Sufficient statistics

\[
\sum_{n=1}^{N} x_n, \quad \sum_{n=1}^{N} x_n x_n^T.
\]
Multivariate Gaussian distribution - MLE

\[
\log \left( p(D|\mu, \Sigma) \right) = -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^\top \Sigma^{-1} (x_n - \mu)
\]

set
\[
\frac{\partial \log \left( p(D|\mu, \Sigma) \right)}{\partial \mu} = \sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu) = 0
\]

\[
\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n
\]
Multivariate Gaussian distribution - MLE

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

$$\frac{\partial \log(p(D|\mu_{ML}, \Sigma))}{\partial \Sigma} = -\frac{N}{2} \Sigma^{-1} + \frac{1}{2} \sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu_{ML})(x_n - \mu_{ML})^\top \Sigma^{-1}$$

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^\top \quad \text{It is semi-positive definite}$$
Multivariate Gaussian distribution - MLE

\[ \mathbb{E}[ \mu_{ML} ] = \mu \]
\[ \mathbb{E}[ \Sigma_{ML} ] = \frac{N - 1}{N} \Sigma \]  

Why?
Multivariate Gaussian distribution - MLE

$$
\mathbb{E}[\mu_{ML}] = \mu
$$

$$
\mathbb{E}[\Sigma_{ML}] = \frac{N-1}{N} \Sigma
$$

Biased estimate

$$
\tilde{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T
$$

Unbiased estimate
Partitioned Gaussian

\[ \mathbf{x} \sim \mathcal{N}(\mathbf{x}| \mu, \Sigma) \]

\[ \mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \]

\[ \Lambda \equiv \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \]

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

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

Quadratic term  Linear term
Conditional Gaussian distribution

- Let us expand the partitioned variables

\[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = \]

\[-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa} (x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab} (x_b - \mu_b) \]

\[-\frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba} (x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb} (x_b - \mu_b). \]
Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(x_a|x_b)$

\[-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) = \]

\[-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b) \]

\[-\frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b).\]

Quadratic term

\[-\frac{1}{2} x_a^T \Lambda_{aa} x_a \]
Conditional Gaussian distribution

• Let us expand the exponent of the conditional $p(x_a|x_b)$

$$-rac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) =$$

$$-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b)$$

$$-\frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b).$$

Quadratic term

$$-\frac{1}{2}x_a^T \Lambda_{aa} x_a \quad \rightarrow \quad \Sigma_{a|b} = \Lambda_{aa}^{-1}$$
Conditional Gaussian distribution

Let us expand the exponent of the conditional $p(x_a|x_b)$

$$-rac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) =$$

$$-\frac{1}{2}(x_a - \mu_a)^T \Lambda_{aa}(x_a - \mu_a) - \frac{1}{2}(x_a - \mu_a)^T \Lambda_{ab}(x_b - \mu_b)$$

$$-\frac{1}{2}(x_b - \mu_b)^T \Lambda_{ba}(x_a - \mu_a) - \frac{1}{2}(x_b - \mu_b)^T \Lambda_{bb}(x_b - \mu_b).$$

Linear term: $x_a^T \{ \Lambda_{aa} \mu_a - \Lambda_{ab}(x_b - \mu_b) \}$
Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(x_a|x_b)$

  Linear term: $x_a^T \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \}$

  $$-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) = -\frac{1}{2} x^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu + \text{const}$$
Conditional Gaussian distribution

- Let us expand the exponent of the conditional $p(x_a|x_b)$

The linear term is:

$$x_a^T \left\{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \right\}$$

This leads to:

$$\mu_{a|b} = \Sigma_{a|b} \left\{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \right\} = \mu_a - \mathbf{\Lambda}_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)$$
Conditional Gaussian distribution

\[ p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

\[
\Sigma_{a|b} = \Lambda^{-1}_{aa} \\
\mu_{a|b} = \mu_a - \Lambda^{-1}_{aa} \Lambda_{ab}(x_b - \mu_b)
\]
Conditional Gaussian distribution

\[
p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b})
\]

\[
\Sigma_{a|b} = \Lambda_{aa}^{-1}
\]
\[
\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)
\]

\[
\begin{pmatrix}
\Sigma_{aa} & \Sigma_{ab} \\
\Sigma_{ba} & \Sigma_{bb}
\end{pmatrix}^{-1} = \begin{pmatrix}
\Lambda_{aa} & \Lambda_{ab} \\
\Lambda_{ba} & \Lambda_{bb}
\end{pmatrix}
\]
Conditional Gaussian distribution

- **Block matrix inverse**

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

\[
M = (A - BD^{-1}C)^{-1}
\]
Conditional Gaussian distribution

- Block matrix inverse

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

\[M = (A - BD^{-1}C)^{-1}\]

\[
\begin{pmatrix}
\Sigma_{aa} & \Sigma_{ab} \\
\Sigma_{ba} & \Sigma_{bb}
\end{pmatrix}^{-1} = \begin{pmatrix}
\Lambda_{aa} & \Lambda_{ab} \\
\Lambda_{ba} & \Lambda_{bb}
\end{pmatrix}
\]

\[
\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}
\]
\[
\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}
\]
Conditional Gaussian distribution

\[ p(x_a|x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

\[ \Sigma_{a|b} = \Lambda_{aa}^{-1} \]

\[ \mu_{a|b} = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab}(x_b - \mu_b) \]

\[ \Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba})^{-1} \]

\[ \Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba})^{-1} \Sigma_{ab} \Sigma_{bb}^{-1} \]

\[ \mu_{a|b} = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b) \]

\[ \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}. \]
Marginal Gaussian distribution

\[ x \sim \mathcal{N}(x|\mu, \Sigma) \]

\[ \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \]

Question 2: What is \( p(x_a) = \int p(x_a, x_b) \, dx_b \)?
Marginal Gaussian distribution

\[ \mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mathbf{\mu}, \Sigma) \]

\[ \mathbf{x} = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \quad \mathbf{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \]

Use the same trick, we can derive that

\[ p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa}) \]

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? \( \lambda = 1/\sigma^2 \quad \lambda > 0 \)

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

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

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

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

$$\text{var}[\lambda] = \frac{a}{b^2}$$
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? \( \lambda = 1/\sigma^2 \quad \lambda > 0 \)

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

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

\[ \mathbb{E}[\log(\lambda)] = \psi(a) - \log(b) \quad \text{digamma function} \]
Gamma distribution

Figure 2.13 Plot of the gamma distribution \( \text{Gam}(\lambda|a, b) \) defined by (2.146) for various values of the parameters \( a \) and \( b \).

The corresponding conjugate prior should therefore be proportional to the product of a power of \( \lambda \) and the exponential of a linear function of \( \lambda \). This corresponds to the gamma distribution which is defined by

\[
\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a) b^a} \lambda^{a-1} \exp\left(-\frac{\lambda}{b}\right).
\]

Here \( \Gamma(a) \) is the gamma function that is defined by (1.141) and that ensures that (2.146) is correctly normalized. The gamma distribution has a finite integral if \( a > 0 \), and the distribution itself is finite if \( a \geq 1 \). It is plotted, for various values of \( a \) and \( b \), in Figure 2.13. The mean and variance of the gamma distribution are given by

\[
E[\lambda] = \frac{a}{b}, \quad \text{var}[\lambda] = \frac{a}{b^2}.
\]

Exercise 2.41 Consider a prior distribution \( \text{Gam}(\lambda|a_0, b_0) \). If we multiply by the likelihood function (2.145), then we obtain a posterior distribution

\[
p(\lambda|X) \propto \lambda^{a_0-1} \lambda^{N/2} \exp\left\{-b_0 \lambda - \lambda^2 \sum_{n=1}^{N} (x_n - \mu)^2\right\}
\]

which we recognize as a gamma distribution of the form \( \text{Gam}(\lambda|a_N, b_N) \) where

\[
a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma^2_{\text{ML}}
\]

where \( \sigma^2_{\text{ML}} \) is the maximum likelihood estimator of the variance. Note that in (2.149) there is no need to keep track of the normalization constants in the prior and the likelihood function because, if required, the correct coefficient can be found at the end using the normalized form (2.146) for the gamma distribution.
Inverse Gamma distribution

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

\[ \lambda^{-1} \sim \text{InvGamma}(\lambda|a, b) \]
Inverse Gamma distribution

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

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

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

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

• Now let us switch to multivariate Gaussian distribution

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

Do we have a distribution over the precision matrix \( \Lambda \equiv \Sigma^{-1} \)?
Wishart Distribution

• Now let us switch to multivariate Gaussian distribution

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

Do we have a distribution over the precision matrix \( \Lambda \equiv \Sigma^{-1} \)?

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

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

degree of freedom
Wishart Distribution

• Now let us switch to multivariate Gaussian distribution

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

Do we have a distribution over the precision matrix \( \Lambda \equiv \Sigma^{-1} \)?

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

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

degree of freedom

multivariate gamma function
Wishart Distribution

• Now let us switch to multivariate Gaussian distribution

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

Do we have a distribution over the precision matrix \( \Lambda \equiv \Sigma^{-1} \)?

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

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

degree of freedom

Multi-dimensional version of Gamma distribution!
Inverse Wishart Distribution

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

\[ \Lambda^{-1} \sim \mathcal{W}^{-1}(\Lambda | \mathbf{W}^{-1}, \nu) \]
Inverse Wishart Distribution

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

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

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

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

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

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

\[ p(x|\mu, a, b) = \int_0^{\infty} \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) \, d\tau \]

\[ = \int_0^{\infty} b^a e^{-b\tau} \tau^{a-1} \left( \frac{\tau}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\tau}{2} (x - \mu)^2 \right\} \, d\tau \]

\[ = \frac{b^a}{\Gamma(a)} \left( \frac{1}{2\pi} \right)^{1/2} \left[ b + \frac{(x - \mu)^2}{2} \right]^{-a-1/2} \Gamma(a + 1/2) \]
The Student t’s distribution

\[ p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) \, d\tau \]

\[ = \int_0^\infty b^a e^{-b\tau} \tau^{a-1} \left( \frac{\tau}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\tau}{2} (x - \mu)^2 \right\} \, d\tau \]

\[ = \frac{b^a}{\Gamma(a)} \left( \frac{1}{2\pi} \right)^{1/2} \left[ b + \frac{(x - \mu)^2}{2} \right]^{-a-1/2} \Gamma(a + 1/2) \]

\[ \nu = 2a \quad \lambda = a/b. \]
Student t’s distribution

\[ p(x|\mu, a, b) = \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) \, d\tau \]

\[ = \int_0^\infty \frac{b^a e^{-b\tau} \tau^{a-1}}{\Gamma(a)} \left( \frac{\tau}{2\pi} \right)^{1/2} \exp \left\{ -\frac{\tau}{2} (x - \mu)^2 \right\} d\tau \]

\[ = \frac{b^a}{\Gamma(a)} \left( \frac{1}{2\pi} \right)^{1/2} \left[ b + \frac{(x - \mu)^2}{2} \right]^{-a-1/2} \Gamma(a + 1/2) \]

\[ \nu = 2a \quad \lambda = a/b. \]

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

mean \hspace{1cm} precision \hspace{1cm} degree of freedom \hspace{1cm} \nu > 0

Infinite weighted sum of Gaussians!
Student t’s distribution – heavy tail

\[ \nu \to \infty \]

\[ \nu = 1.0 \]

\[ \nu = 0.1 \]

\[ \text{St}(x|\mu, \lambda, \nu) \to \mathcal{N}(x|\mu, \lambda^{-1}) \]
**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

\[
p(x|\mu, a, b) = \int_0^\infty p(x|\mu, \tau)p(\tau|a, b)\,d\tau
\]
Student t’s distribution

\[ p(x|\mu, a, b) = \int_0^\infty p(x|\mu, \tau)p(\tau|a, b)d\tau \]

\[ \nu = 2a, \lambda = a/b, \eta = \tau b/a \]
Student t’s distribution

\[ p(x|\mu, a, b) = \int_0^\infty p(x|\mu, \tau)p(\tau|a, b)d\tau \]

\[ \nu = 2a, \lambda = a/b, \eta = \tau b/a \]

\[ \text{St}(x|\mu, \lambda, \nu) = \int_0^\infty \mathcal{N}(x|\mu, (\eta\lambda)^{-1}) \operatorname{Gam}(\eta|\nu/2, \nu/2) d\eta \]
Multivariate student-t distribution

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

\[ \text{St}(x|\mu, \Lambda, \nu) = \int_0^{\infty} N \left( x|\mu, (\eta \Lambda)^{-1} \right) \text{Gam}(\eta|\nu/2, \nu/2) \, d\eta \]
Multivariate student-t distribution

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

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

\[
\text{St}(x|\mu, \Lambda, \nu) = \frac{\Gamma(d/2 + \nu/2)}{\Gamma(\nu/2)} \frac{|\Lambda|^{1/2}}{(\pi \nu)^{d/2}} \left[ 1 + \frac{1}{\nu} (x - \mu)^\top \Lambda (x - \mu) \right]^{-d/2-\nu/2}
\]
Multivariate student-t distribution

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

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


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.