Probability Distributions

Evgeny Burnaev

Skoltech, Moscow, Russia

Maximum likelihood approach

- We have the likelihood of the data $\log p(\mathcal{D}|\mathbf{w})$ which depends on the vector of parameters $\mathbf{w}$ we want to estimate.
- A natural way to estimate parameters is to maximize the likelihood:
$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathcal{D}|\mathbf{w})$$

Gaussian MLE

Likelihood of an i.i.d. data sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ having gaussian distribution

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$$

Log-likelihood is equal to

$$\log p(\mathbf{X}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi)$$

MLE is equal to

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i,\ \sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{ML})^2$$

Properties:

$$\mathbb{E}[\mu_{ML}] = \mu,\ \mathbb{E}[\sigma^2_{ML}] = \left(\frac{n-1}{n}\right) \sigma^2$$

Bayesian approach

- We have the likelihood of the data $\log p(\mathcal{D}|\mathbf{w})$ which depends on the vector of parameters $\mathbf{w}$ we want to estimate.
- We also have a prior distribution for the parameters $p(\mathbf{w})$.
- The goal is to look into a conditional probability (posterior distribution):

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- We can use MAP (Maximum posterior) estimate $\equiv$ regularized MLE:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}}[\log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w})]$$

Bayesian estimation for Gaussian model

- We define a prior distribution for μ:

$$p(\mu) = \mathcal{N}(\mu; \mu_0, \beta^2).$$

- Then the logarithm of the conditional probability is proportional to:

$$\log p(\mu|\mathcal{D}) \sim -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{1}{2\beta^2}\mu^2$$

- In this case we can evaluate μ_{MAP} in a direct way:

$$\mu_{\text{MAP}} = \frac{n\beta^2}{n\beta^2 + \sigma^2}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{\sigma^2}{n\beta^2 + \sigma^2}\mu_0 =$$

$$= \frac{n\beta^2}{\sigma^2 + n\beta^2}\mu_{\text{MLE}} + \frac{\sigma^2}{\sigma^2 + n\beta^2}\mu_0.$$

Bernoulli distribution

- $x \in \{0, 1\}$, $p(x = 1|\mu) = \mu$
- $Bern(x|\mu) = \mu^x (1 - \mu)^{1-x}$

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

- Data set $\mathcal{D} = \{x_1, \dots, x_n\}$, then likelihood

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

- Log-likelihood

$$\begin{aligned} \log p(\mathcal{D}|\mu) &= \sum_{i=1}^{n} \log p(x_i|\mu) \\ &= \sum_{i=1}^{n} \{x_i \log \mu + (1 - x_i) \log(1 - \mu)\} \end{aligned}$$

- MLE $\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{m}{n}$, where $m = \sum_{i=1}^{n} x_i$

Binomial distribution

- n Bernoulli trials with probability of success equal to μ
- m is a number of trials with $x = 1$, then

$$Bin(m|n,\mu) = \binom{n}{m} \mu^m (1-\mu)^{n-m}$$

- Mean value and variance

$$\mathbb{E}[m] = \sum_{i=1}^{n} \mathbb{E}[x_i] = n\mu, \; \mathrm{var}[m] = n\mu(1-\mu)$$

Beta distribution

- Prior $p(\mu)$ for μ
- Density

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

where gamma-function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, $x > 0$

- Mean and variance

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

Beta distribution

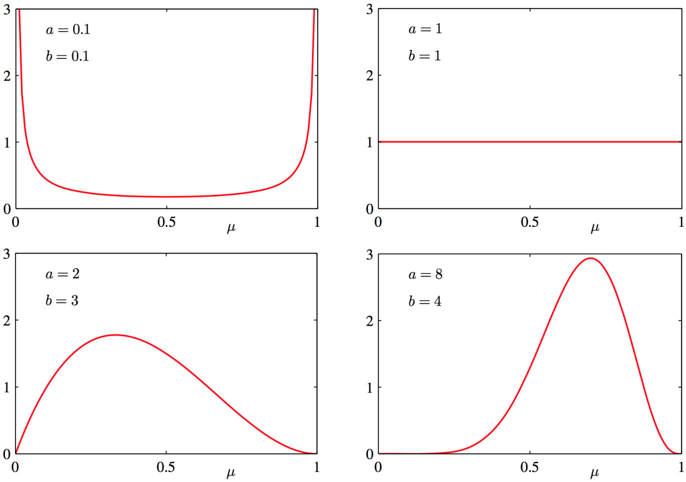


Figure: Gamma-distribution $\Gamma(a, b)$

Beta distribution

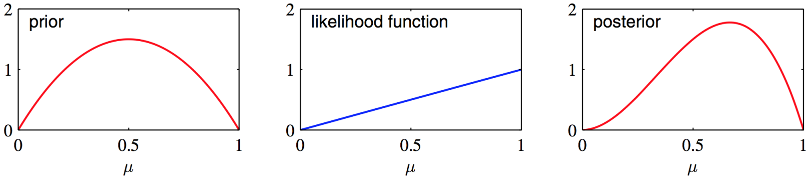
- Posterior $p(\mu|m, l, a, b)$ with $l = n - m$ is equal to

$$p(\mu|m, l, a, b) \sim Bin(m|n, \mu) \times p(\mu|a, b) \sim \mu^{m+a-1}(1-\mu)^{l+b-1}$$

- Comparing with $Beta(\mu|a, b)$ we get that normalization constant is equal to

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

Beta distribution



- Assume we obtain observations sequentially
- Additional observation $x = 1 \Rightarrow$ incrementing value of a by 1
- Additional observation $x = 0 \Rightarrow$ incrementing value of b by 1

Beta distribution

- Predict the outcome of the next trial
- We have to evaluate the predictive distribution of x given observed data set $\mathcal{D}$

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu$$
$$= \int_0^1 \mu p(\mu|\mathcal{D})d\mu = \mathbb{E}[\mu|\mathcal{D}]$$
$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$$

- As $m, l \to \infty$ the result reduces to MLE
- For a finite data set, the posterior mean for μ always lies between the prior mean and the MLE for μ

Multinomial distribution

- Discrete variables that can take on one of K possible mutually exclusive states
- 1-of-K scheme, in which the variable is represented by a K-dimensional vector $\mathbf{x} = (x_1, \dots, x_K)$, e.g.

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}, \ \sum_{i=1}^{K} x_i = 1$$

- Distribution of $\mathbf{x}$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}, \ \boldsymbol{\mu} = (\mu_1, \dots, \mu_K)^{\mathrm{T}}, \ \mu_k \geq 0, \ \sum_{k=1}^{K} \mu_k = 1$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1, \ \mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = \boldsymbol{\mu}$$

Multinomial distribution

- For a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ the likelihood has the form

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

- Using the Lagrange multiplier method to optimize the likelihood we get that

$$\mu_k^{ML} = \frac{m_k}{n}$$

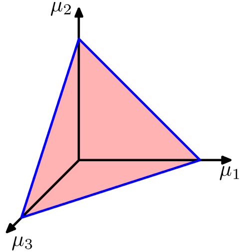
- Multinomial distribution

$$Mult(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, n) = \binom{n}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k},$$

where $\sum_{k=1}^{K} m_k = n$

Dirichlet distribution

- Prior distributions for the parameters $\{\mu_k\}$ of the multinomial distribution



- Conjugate prior is given by

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \sim \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \; 0 \leq \mu_k \leq 1, \; \sum_{k=1}^{K} \mu_k = 1$$

since $p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \sim p(\mathcal{D}|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \sim \prod_{k=1}^{K} \mu_k^{\alpha_k + \mu_k - 1}$

Dirichlet distribution

- The normalized form of the Dirichlet distribution

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

Dirichlet distribution

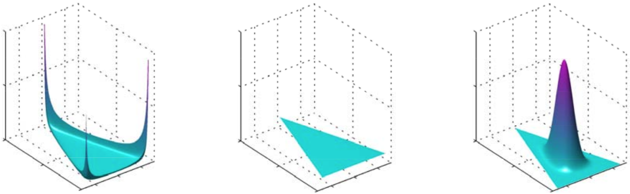


Figure: $\{\alpha_k\} = 0.1$ (left), $\{\alpha_k\} = 1$ (centre), $\{\alpha_k\} = 10$ (right)

Then the normalized posterior

$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= Dir(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m}) \\
&= \frac{\Gamma(\alpha_0 + n)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1},
\end{aligned}
$$

where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$, $\mathbf{m} = (m_1, \ldots, m_K)^{\mathrm{T}}$

Gaussian Distribution

- In case of a single variable x

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

- For $\mathbf{x} \in \mathbb{R}^d$ with $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$ and $\operatorname{cov}[\mathbf{x}] = \boldsymbol{\Sigma}$

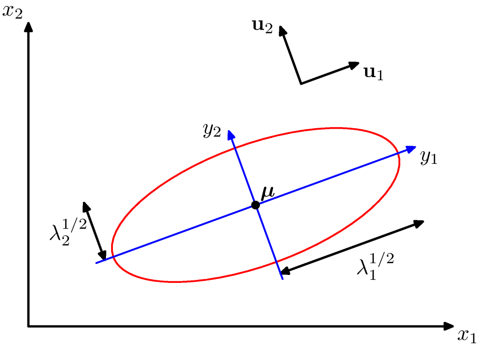
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{d/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- It holds that

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

- The total number of parameters is equal to $\dim(\boldsymbol{\mu}) + \dim(\boldsymbol{\Sigma}) = d + d(d+1)/2 = d(d+3)/2$

Gaussian distribution



- The red curve shows the elliptical surface of constant probability density for $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $d = 2$
- Curve corresponds to the density $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$
- The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix $\boldsymbol{\Sigma}$, with eigenvalues λ_i

Conditional Gaussian distribution

- $\mathbf{x}$ is distributed as $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$
- Let us also partition the mean and the covariance

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

and define the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$,

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

Conditional Gaussian distribution

- In order to get $p(\mathbf{x}_a|\mathbf{x}_b)$ we need to fix $\mathbf{x}_b$ in $p(\mathbf{x}) = p(\mathbf{x}_a, \mathbf{x}_b)$ and normalize it w.r.t. $\mathbf{x}_a$
- Let us consider a quadratic form

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

- This is a quadratic form as a function of $\mathbf{x}_a \Rightarrow p(\mathbf{x}_a|\mathbf{x}_b)$ will be Gaussian $\mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$
- Let us "complete the square", i.e. represent the previous sum as a quadratic form w.r.t. $\mathbf{x}_a$

Conditional Gaussian distribution

- It is obvious that
$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathrm{const},$$
- If we pick out all terms that are second order in $\mathbf{x}_a$, then we get
$$-\frac{1}{2}\mathbf{x}_a^{\mathrm{T}}\boldsymbol{\Lambda}_{aa}\mathbf{x}_a,$$
thus
$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}$$
- Analogously (Exercise!!!) we get that
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

Conditional Gaussian distribution

- Identity for the inverse of a partitioned matrix

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

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

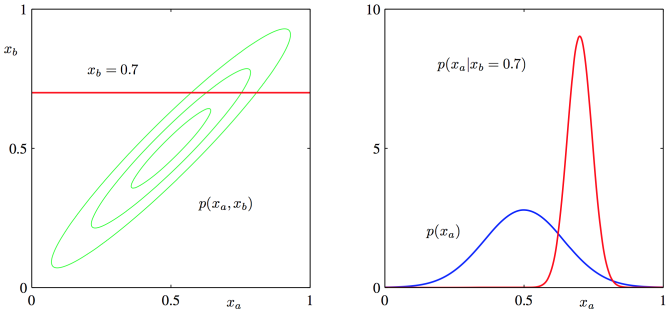
- Since $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda}$ we get that

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

- Thus we get that

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

Conditional Gaussian distribution



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

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

Conditional Gaussian distribution

- Let us calculate $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$
- Along the same lines it can be shown that

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

Conditional Gaussian distribution

We assume that

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}),$$

then using the same considerations

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right\}, \boldsymbol{\Sigma}),$$

where

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

MLE for Gaussian distribution

Gaussian log-likelihood

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- We get that

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

- Properties

$$\mathbb{E}[\mu_{ML}] = \boldsymbol{\mu},\ \mathbb{E}[\boldsymbol{\Sigma}_{ML}] = \frac{n-1}{n}\boldsymbol{\Sigma}$$

- Corrected estimate

$$\tilde{\boldsymbol{\Sigma}} = \frac{n}{n-1}\boldsymbol{\Sigma}_{ML}$$

Bayesian inference for the Gaussian

Data point $x \sim \mathcal{N}(x|\mu, \sigma^2)$, $\mathbf{X} = \{x_1, \dots, x_n\}$. We assume that σ^2 is known, then the likelihood

$$p(\mathbf{X}|\mu) = \prod_{i=1}^{n} p(x_i|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

- We use prior $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- The posterior

$$p(\mu|\mathbf{X}) \sim p(\mathbf{X}|\mu)p(\mu) \sim \mathcal{N}(\mu|\mu_n, \sigma_n^2),$$

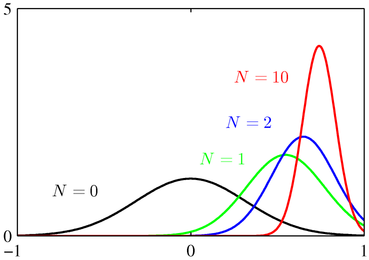
where

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML}$$

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

- Thus for $n \to \infty$ we tend to use μ_{ML}

Bayesian inference for the Gaussian



- Dependence of posterior on n
- The data points are generated from $\mathcal{N}(x|0.8, 0.1)$, prior has mean 0, the variance is set to the true value

Sequential view of the inference problem

$$p(\boldsymbol{\mu}|\mathcal{D}) \sim \left[p(\boldsymbol{\mu}) \prod_{i=1}^{n-1} p(\mathbf{x}_i|\boldsymbol{\mu}) \right] p(\mathbf{x}_n|\boldsymbol{\mu})$$

Bayesian inference for the Gaussian

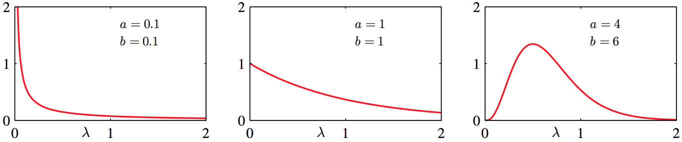
- Let us suppose that the mean is known and we wish to infer the variance
- We re-parameterize it by precision $\lambda = \frac{1}{\sigma^2}$
- The likelihood has the form

$$p(\mathbf{X}|\lambda) = \prod_{i=1}^{n} \mathcal{N}(x_i|\boldsymbol{\mu}, \lambda^{-1}) \sim \lambda^{n/2} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

- Gamma prior on λ

$$Gam(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda),\ \mathbb{E}[\lambda] = \frac{a}{b},\ \mathrm{var}[\lambda] = \frac{a}{b^2}$$

Bayesian inference for the Gaussian



- Posterior has the form

$$p(\lambda|\mathbf{X}) \sim \lambda^{a_0-1}\lambda^{n/2}\exp\left\{-b_0\lambda - \frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$
$$\sim Gam(\lambda|a_n, b_n),$$

where

$$a_n = a_0 + \frac{n}{2}$$
$$b_n = b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 = b_0 + \frac{n}{2}\sigma_{ML}^2$$

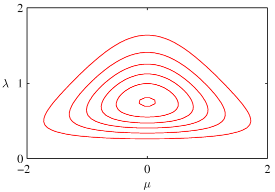
Bayesian inference for the Gaussian

- In general case (mean and variance are not known) the likelihood has the form

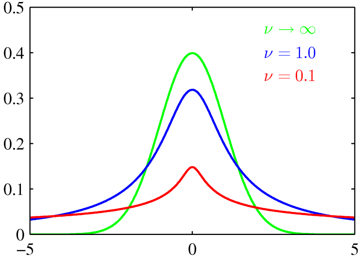
$$p(\mathbf{X}|\mu,\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(x_i-\mu)^2\right\}$$

- It can be easily proved that the conjugate prior has the form (normal-gamma distribution)

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})Gam(\lambda|a,b)$$



Student's t-distribution



Density is a mixture

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

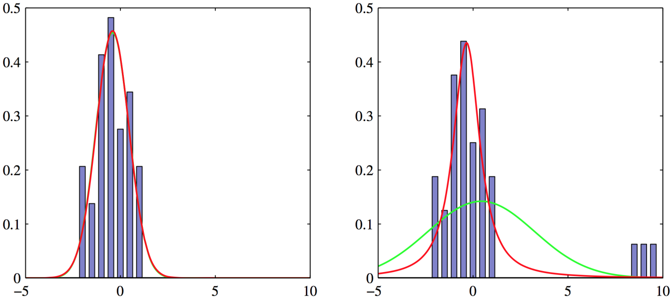
Student's t-distribution

Re-parameterizing we get that

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

For $\nu \to \infty$ it holds that $St(x|\mu, \lambda, \nu) \to \mathcal{N}(x|\mu, \lambda^{-1})$

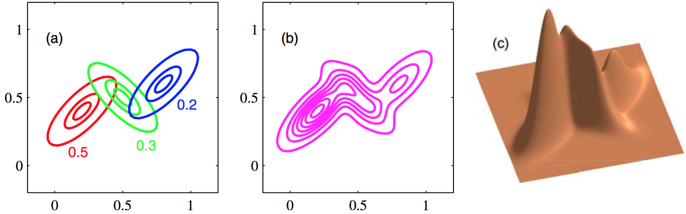
Bayesian inference for the Gaussian



Student's t-distribution has heavy tails:

- Left: Histogram (30 points from a Gaussian distr.), together with MLE fit of a t-distribution (red curve) and a Gaussian (green curve)
- Right: the same data set but with three additional outliers. The Gaussian (green curve) is strongly distorted

Mixture of Gaussians



- Superposition of K Gaussian densities

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$

Mixture of Gaussians

- Parameters: $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$
- The likelihood

$$\log p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\},$$

  where $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$
- Optimization: EM algorithm (further in this course)