

# Maximum-likelihood and Bayesian parameter estimation

Andrea Passerini  
passerini@disi.unitn.it

Machine Learning

# Maximum-likelihood (ML) estimation

## Setting

- A training data  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  of i.i.d. examples for the target class  $y$  is available
- We assume the parameter vector  $\theta$  has a fixed but unknown value
- We estimate such value maximizing its **likelihood** with respect to the training data:

$$\theta^* = \operatorname{argmax}_{\theta} p(\mathcal{D}|\theta) = \operatorname{argmax}_{\theta} \prod_{j=1}^n p(\mathbf{x}_j|\theta)$$

- The joint probability over  $\mathcal{D}$  decomposes into a product as examples are i.i.d (thus independent of each other given the distribution)

# Maximum-likelihood estimation

## Maximizing log-likelihood

- It is usually simpler to maximize the logarithm of the likelihood (monotonic):

$$\theta^* = \operatorname{argmax}_{\theta} \ln p(\mathcal{D}|\theta) = \operatorname{argmax}_{\theta} \sum_{j=1}^n \ln p(\mathbf{x}_j|\theta)$$

- Necessary conditions for the maximum can be obtained zeroing the gradient wrt to  $\theta$ :

$$\nabla_{\theta} \sum_{j=1}^n \ln p(\mathbf{x}_j|\theta) = \mathbf{0}$$

- Points zeroing the gradient can be local or global maxima depending on the form of the distribution

## setting

- Assumes parameters  $\theta_i$  are *random variables* with some known *prior* distribution
- Predictions for new examples are obtained *integrating* over all possible values for the parameters:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) = \int_{\theta_i} p(\mathbf{x}, \theta_i|y_i, \mathcal{D}_i) d\theta_i$$

- probability of  $\mathbf{x}$  given each class  $y_i$  is independent of the other classes  $y_j$ , for simplicity we can again write:

$$p(\mathbf{x}|y_i, \mathcal{D}_i) \rightarrow p(\mathbf{x}|\mathcal{D}) = \int_{\theta} p(\mathbf{x}, \theta|\mathcal{D}) d\theta$$

- where  $\mathcal{D}$  is a dataset for a certain class  $y$  and  $\theta$  the parameters of the distribution

## setting

$$p(\mathbf{x}|\mathcal{D}) = \int_{\theta} p(\mathbf{x}, \theta|\mathcal{D})d\theta = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})d\theta$$

- $p(\mathbf{x}|\theta)$  can be easily computed (we have both form and parameters of distribution, e.g. Gaussian)
- need to estimate the parameter posterior density given the training set:

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

## denominator

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

- $p(\mathcal{D})$  is a constant independent of  $\theta$  (i.e. it will no influence final Bayesian decision)
- if final *probability* (not only decision) is needed we can compute:

$$p(\mathcal{D}) = \int_{\theta} p(\mathcal{D}|\theta)p(\theta)d\theta$$

# Sufficient statistics

## Definition

- Any function on a set of samples  $\mathcal{D}$  is a *statistic*
- A statistic  $\mathbf{s} = \phi(\mathcal{D})$  is *sufficient* for some parameters  $\theta$  if:

$$P(\mathcal{D}|\mathbf{s}, \theta) = P(\mathcal{D}|\mathbf{s})$$

- If  $\theta$  is a random variable, a sufficient statistic contains all relevant information  $\mathcal{D}$  has for estimating it:

$$p(\theta|\mathcal{D}, \mathbf{s}) = \frac{p(\mathcal{D}|\theta, \mathbf{s})p(\theta|\mathbf{s})}{p(\mathcal{D}|\mathbf{s})} = p(\theta|\mathbf{s})$$

## Use

- A sufficient statistic allows to compress a sample  $\mathcal{D}$  into (possibly few) values
- Sample mean and covariance are sufficient statistics for true mean and covariance of the Gaussian distribution

# Conjugate priors

## Definition

- Given a likelihood function  $p(x|\theta)$
- Given a prior distribution  $p(\theta)$
- $p(\theta)$  is a *conjugate prior* for  $p(x|\theta)$  if the posterior distribution  $p(\theta|x)$  is in the same family as the prior  $p(\theta)$

## Examples

Likelihood	Parameters	Conjugate prior
Binomial	$p$ (probability)	Beta
Multinomial	$\mathbf{p}$ (probability vector)	Dirichlet
Normal	$\mu$ (mean)	Normal
Multivariate normal	$\mu_j$ (mean vector)	Normal

# Probability distributions

## Bernoulli distribution

- Two possible values (outcomes): 1 (success), 0 (failure).
- Parameters:  $p$  probability of success.
- Probability mass function:

$$P(x; p) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- $E[x] = p$
- $\text{Var}[x] = p(1 - p)$

## Example: tossing a coin

- Head (success) and tail (failure) possible outcomes
- $p$  is probability of head

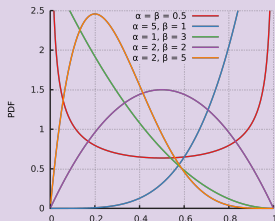
# Probability distributions

## Beta distribution

- Defined in the interval  $[0, 1]$
- Parameters:  $\alpha, \beta$
- Probability density function:

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

- $E[x] = \frac{\alpha}{\alpha + \beta}$        $\Gamma(x + 1) = x\Gamma(x), \Gamma(1) = 1$
- $\text{Var}[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$



## Note

It models the posterior distribution of parameter  $p$  of a binomial distribution after observing  $\alpha - 1$  independent events with probability  $p$  and  $\beta - 1$  with probability  $1 - p$ .

## Setting

- Boolean event:  $x = 1$  for success,  $x = 0$  for failure (e.g. tossing a coin)
- Parameters:  $\theta$  = probability of success (e.g. head)
- Probability mass function

$$P(x|\theta) = \theta^x(1 - \theta)^{1-x}$$

- Beta conjugate prior:

$$P(\theta) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_h)\Gamma(\alpha_t)} \theta^{\alpha_h-1} (1 - \theta)^{\alpha_t-1}$$

# Bernoulli distribution

## Maximum likelihood estimation: example

- Dataset  $\mathcal{D} = \{H, H, T, T, T, H, H\}$  of  $N$  realizations (e.g. head/tail coin toss results)
- Likelihood function:

$$p(\mathcal{D}|\theta) = \theta \cdot \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot \theta = \theta^h (1 - \theta)^t$$

- Maximum likelihood parameter:

$$\frac{\partial}{\partial \theta} \ln p(\mathcal{D}|\theta) = 0 \quad \Rightarrow \quad \frac{\partial}{\partial \theta} h \ln \theta + t \ln (1 - \theta) = 0$$

$$h \frac{1}{\theta} - t \frac{1}{1 - \theta} = 0$$

$$h(1 - \theta) = t\theta$$

$$\theta = \frac{h}{h + t}$$

- $h, t$  are the sufficient statistics

# Bernoulli distribution

## Bayesian estimation: example

- Parameter posterior is proportional to:

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta) \propto \theta^h(1-\theta)^t\theta^{\alpha_h-1}(1-\theta)^{\alpha_t-1}$$

- i.e. the posterior has a beta distribution with parameters  $h + \alpha_h, t + \alpha_t$ :

$$P(\theta|\mathcal{D}) \propto \theta^{h+\alpha_h-1}(1-\theta)^{t+\alpha_t-1}$$

- The prediction for a new event is the expected value of the posterior beta:

$$\begin{aligned}P(x|\mathcal{D}) &= \int P(x|\theta)P(\theta|\mathcal{D})d\theta = \int \theta P(\theta|\mathcal{D})d\theta \\ &= \mathbb{E}_{P(\theta|\mathcal{D})}[\theta] = \frac{h + \alpha_h}{h + t + \alpha_h + \alpha_t}\end{aligned}$$

## Interpreting priors

- Our prior knowledge is encoded as a number  $\alpha = \alpha_h + \alpha_t$  of imaginary experiments
- we assume  $\alpha_h$  times we observed heads
- $\alpha$  is called *equivalent sample size*
- $\alpha \rightarrow 0$  reduces estimation to the classical ML approach (frequentist)

## Multinomial distribution (one sample)

- Models the probability of a certain outcome for an event with  $m$  possible outcomes.
- Parameters:  $p_1, \dots, p_m$  probability of each outcome
- Probability mass function:

$$P(x_1, \dots, x_m; p_1, \dots, p_m) = \prod_{i=1}^m p_i^{x_i}$$

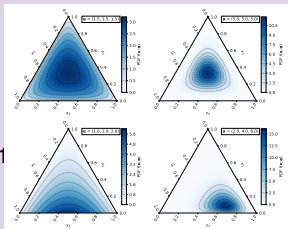
- where  $x_1, \dots, x_m$  is a vector with  $x_i = 1$  for outcome  $i$  and  $x_j = 0$  for all  $j \neq i$ .
- $E[x_i] = p_i$
- $\text{Var}[x_i] = p_i(1 - p_i)$
- $\text{Cov}[x_i, x_j] = -p_i p_j$

# Probability distributions

## Dirichlet distribution

- Defined:  $\mathbf{x} \in [0, 1]^m, \sum_{i=1}^m x_i = 1$
- Parameters:  $\alpha = \alpha_1, \dots, \alpha_m$
- Probability density function:

$$p(x_1, \dots, x_m; \alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m x_i^{\alpha_i - 1}$$



- $E[x_j] = \frac{\alpha_j}{\alpha_0}$

where  $\alpha_0 = \sum_{j=1}^m \alpha_j$

- $\text{Var}[x_j] = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$

$$\text{Cov}[x_i, x_j] = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

## Note

It models the posterior distribution of parameters  $\mathbf{p}$  of a multinomial distribution after observing  $\alpha_j - 1$  times each mutually exclusive event

# Multinomial distribution

## Setting

- Categorical event with  $r$  states  $x \in \{x^1, \dots, x^r\}$  (e.g. tossing a six-faced dice)
- One-hot encoding  $\mathbf{z}(x) = [z_1(x), \dots, z_r(x)]$  with  $z_k(x) = 1$  if  $x = x^k$ , 0 otherwise.
- Parameters:  $\theta = [\theta_1, \dots, \theta_r]$  probability of each state
- Probability mass function

$$P(x|\theta) = \prod_{k=1}^r \theta_k^{z_k(x)}$$

- Dirichlet conjugate prior:

$$P(\theta|\psi) = P(\theta|\alpha_1, \dots, \alpha_r) = \frac{\Gamma(\alpha)}{\prod_{k=1}^r \Gamma(\alpha_k)} \prod_{k=1}^r \theta_k^{\alpha_k - 1}$$

## Maximum likelihood estimation: example

- Dataset  $\mathcal{D}$  of  $N$  realizations (e.g. results of tossing a dice)
- Likelihood function:

$$p(\mathcal{D}|\theta) = \prod_{j=1}^N \prod_{k=1}^r \theta_k^{z_k(x_j)} = \prod_{k=1}^r \theta_k^{N_k}$$

- Maximum likelihood parameter:

$$\theta_k = \frac{N_k}{N}$$

- $N_1, \dots, N_r$  are the sufficient statistics

# Multinomial distribution

## Bayesian estimation: example

- Parameter posterior is proportional to:

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta) \propto \prod_{k=1}^r \theta_k^{N_k} \theta_k^{\alpha_k-1}$$

- i.e. the posterior has a Dirichlet distribution with parameters  $N_k + \alpha_k, k = 1, \dots, r$ :

$$P(\theta|\mathcal{D}) \propto \prod_{k=1}^r \theta_k^{N_k+\alpha_k-1}$$

- The prediction for a new event is the expected value of the posterior Dirichlet:

$$P(x_k|\mathcal{D}) = \int \theta_k P(\theta|\mathcal{D}) d\theta = E_{P(\theta|\mathcal{D})}[\theta_k] = \frac{N_k + \alpha_k}{N + \alpha}$$

Appendix

Additional reference material

# Maximum-likelihood estimation

## Multivariate Gaussian case: proof (mean)

- The gradient wrt to the mean is:

$$\nabla_{\boldsymbol{\mu}} \sum_{j=1}^n -\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) - \frac{1}{2} \ln(2\pi)^d |\boldsymbol{\Sigma}| =$$
$$\sum_{j=1}^n \boldsymbol{\Sigma}^{-1}(\mathbf{x}_j - \boldsymbol{\mu})$$

## Note

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \mathbf{A}^T \mathbf{x} + \mathbf{A} \mathbf{x} \\ &= 2\mathbf{A} \mathbf{x} \quad \text{for symmetric } \mathbf{A} \end{aligned}$$

## Multivariate Gaussian case: proof (mean)

- Setting the gradient to zero gives:

$$\sum_{j=1}^n \Sigma^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) = \mathbf{0}$$

$$\sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu}) = \Sigma \mathbf{0} = \mathbf{0}$$

$$\sum_{j=1}^n \mathbf{x}_j = \sum_{j=1}^n \boldsymbol{\mu} = n \boldsymbol{\mu}$$

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j$$

## Multivariate Gaussian case: proof (covariance)

- The gradient wrt to the covariance is:

$$\frac{\partial}{\partial \Sigma} \sum_{j=1}^n -\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) - \frac{1}{2} \ln(2\pi)^d |\Sigma| =$$
$$-\frac{1}{2} \left( \sum_{j=1}^n \frac{\partial}{\partial \Sigma} (\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_j - \boldsymbol{\mu}) + \sum_{j=1}^n \frac{\partial}{\partial \Sigma} \ln(2\pi)^d |\Sigma| \right)$$

# Maximum-likelihood estimation

## Multivariate Gaussian case: proof (covariance)

$$\begin{aligned}\frac{\partial}{\partial \Sigma} (\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) &= \\ (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \frac{\partial}{\partial \Sigma} \Sigma^{-1} &= \\ -(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2}\end{aligned}$$

## Note

Use matrix derivative rule:

$$\frac{\partial}{\partial B} \text{tr}(ABC) = CA$$

Where  $A = (\mathbf{x}_j - \boldsymbol{\mu})^t$ ,  $B = \Sigma^{-1}$ ,  $C = (\mathbf{x}_j - \boldsymbol{\mu})$  and  $\text{tr}(ABC) = ABC$  as  $ABC$  is a scalar.

## Multivariate Gaussian case: proof (covariance)

$$\begin{aligned}\frac{\partial}{\partial \Sigma} \ln (2\pi)^d |\Sigma| &= \frac{1}{(2\pi)^d} |\Sigma|^{-1} \frac{\partial}{\partial \Sigma} (2\pi)^d |\Sigma| = \\ \frac{1}{(2\pi)^d} |\Sigma|^{-1} (2\pi)^d \frac{\partial}{\partial \Sigma} |\Sigma| &= |\Sigma|^{-1} |\Sigma| \Sigma^{-1} = \Sigma^{-1}\end{aligned}$$

## Note

Use matrix derivative rule:

$$\frac{\partial}{\partial A} |A| = |A| A^{-1}$$

## Multivariate Gaussian case: proof (covariance)

- Combining and putting equal to zero:

$$-\frac{1}{2} \left( \sum_{j=1}^n \overbrace{-(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2}}^{\frac{\partial}{\partial \Sigma} (\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu})} + \sum_{j=1}^n \underbrace{\frac{\partial}{\partial \Sigma} \ln(2\pi)^d |\Sigma|}_{\Sigma^{-1}} \right) = 0$$

# Maximum-likelihood estimation

## Multivariate Gaussian case: proof (covariance)

$$\begin{aligned}\sum_{j=1}^n \Sigma^{-1} &= \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2} \\ \Sigma^2 \sum_{j=1}^n \Sigma^{-1} &= \Sigma^2 \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \Sigma^{-2} \\ \sum_{j=1}^n \Sigma &= \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \\ n\Sigma &= \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t \\ \Sigma &= \frac{1}{n} \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^t\end{aligned}$$

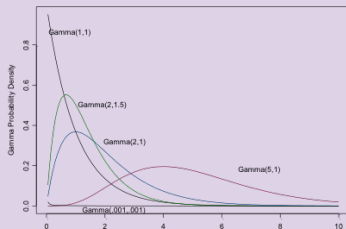
# Bayesian estimation

## Gamma distribution

- Defined in the interval  $[0, \infty)$
- Parameters:  $\alpha > 0$  (shape)  
 $\beta > 0$  (rate)
- Probability density function:

$$p(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- $E[X] = \frac{\alpha}{\beta}$
- $\text{Var}[X] = \frac{\alpha}{\beta^2}$



## Note

Used to model the prior distribution of the *precision* (inverse variance, i.e.  $\lambda = 1/\sigma^2$ ).

## Univariate normal case: unknown $\mu$ and $\lambda = 1/\sigma^2$

- Examples are drawn from:

$$p(x|\mu, \lambda) \sim N(\mu, 1/\lambda)$$

- The Prior of mean and precision is the NormalGamma distribution:

$$\begin{aligned} p(\mu, \lambda) &= p(\mu|\lambda)p(\lambda) = N(\mu|\mu_0, \frac{1}{\kappa_0\lambda})\text{Ga}(\lambda|\alpha_0, \beta_0) \\ &= \text{NG}(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) \end{aligned}$$

# Univariate normal case: unknown $\mu$ and $\lambda = 1/\sigma^2$

*a posteriori* parameter density

$$p(\mu, \lambda | \mathcal{D}) = \frac{1}{\mathcal{D}} \prod_{j=1}^n \underbrace{\frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda}{2}(x_j - \mu)^2\right]}_{p(x_j | \mu, \lambda)} \underbrace{\frac{(\kappa_0 \lambda)^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\kappa_0 \lambda}{2}(\mu - \mu_0)^2\right]}_{p(\mu | \lambda)} \underbrace{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} \exp(-\beta_0 \lambda)}_{p(\lambda)}$$
$$\propto \lambda^{\alpha_0 + n/2 - 1} \exp(-\beta_0 \lambda) \lambda^{1/2} \exp\left[-\frac{\lambda}{2} \left[ \sum_{j=1}^n (x_j - \mu)^2 - \kappa_0 (\mu - \mu_0)^2 \right]\right]$$

*a posteriori* parameter density is still NormalGamma

$$p(\mu, \lambda | \mathcal{D}) = \text{NG}(\mu, \lambda | \mu_n, \kappa_n, \alpha_n, \beta_n)$$

# Univariate normal case: unknown $\mu$ and $\lambda = 1/\sigma^2$

*a posteriori* parameter density is still NormalGamma

$$p(\mu, \lambda | \mathcal{D}) = \text{NG}(\mu, \lambda | \mu_n, \kappa_n, \alpha_n, \beta_n)$$

where

$$\mu_n = \frac{\kappa_0 \mu_0 + n \hat{\mu}_n}{\kappa_0 + n}$$

$$\kappa_n = \kappa_0 + n$$

$$\alpha_n = \alpha_0 + n/2$$

$$\beta_n = \beta_0 + \frac{1}{2} \sum_{j=1}^n (x_j - \hat{\mu}_n)^2 + \frac{\kappa_0 n (\hat{\mu}_n - \mu_0)^2}{2(\kappa_0 + n)}$$

## Interpreting the posterior

- Posterior mean is weighted average of prior ( $\mu_0$ ) and sample ( $\mu_n$ ) means, weighted by  $\kappa_0$  and  $n$  respectively

$$\mu_n = \frac{\kappa_0 \mu_0 + n \hat{\mu}_n}{\kappa_0 + n}$$

- Posterior  $\kappa_n$  is increased by the number of samples  $n$

$$\kappa_n = \kappa_0 + n$$

- Posterior  $\alpha_n$  is increased by half the number of samples  $n$

$$\alpha_n = \alpha_0 + n/2$$

## Interpreting the posterior

- Posterior sum of squares ( $\beta_n$ ) is sum of prior sum of squares ( $\beta_0$ ) and sample sum of squares  $\frac{1}{2} \sum_{j=1}^n (x_j - \hat{\mu}_n)^2$  and a term due to the discrepancy between the sample mean and the prior mean.

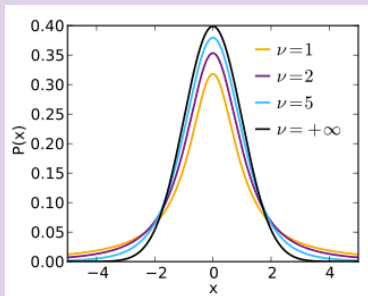
$$\beta_n = \beta_0 + \frac{1}{2} \sum_{j=1}^n (x_j - \hat{\mu}_n)^2 + \frac{\kappa_0 n (\hat{\mu}_n - \mu_0)^2}{2(\kappa_0 + n)}$$

# Univariate normal case: unknown $\mu$ and $\lambda = 1/\sigma^2$

## Computing the posterior predictive

$$\begin{aligned} p(x|\mathcal{D}) &= \int_{\mu} \int_{\lambda} p(x|\mu, \lambda) p(\mu, \lambda|\mathcal{D}) d\mu d\lambda \\ &= \frac{P(x, \mathcal{D})}{P(\mathcal{D})} = t_{2\alpha_n} \left( x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n \kappa_n} \right) \end{aligned}$$

- It is a T-distribution with mean  $\mu_n$  and precision  $\frac{\beta_n(\kappa_n+1)}{\alpha_n \kappa_n}$  (proof omitted)



## Wishart distribution

- Defined over  $d \times d$  positive semi-definite matrix
- Parameters:  $\nu > d - 1$  (degree of freedom)  $T > 0$  ( $d \times d$  scale matrix)
- Probability density function:

$$p(X; \nu, T) = \frac{1}{2^{\nu d/2} |T|^{\nu/2} \Gamma_d(\nu/2)} |X|^{\frac{\nu-d-1}{2}} \exp -\frac{1}{2} \text{tr}(T^{-1}X)$$

- $E[X] = \nu T$
- $\text{Var}[X_{ij}] = \nu(T_i i^2 + T_{ij} T_{jj})$

## Note

Used to model the prior distribution of the *precision* matrix (inverse covariance matrix, i.e.  $\Lambda = \Sigma^{-1}$ ).  $T$  is the prior covariance

## Multivariate normal case: unknown $\mu$ and $\Sigma$

- Examples are drawn from:

$$p(x|\mu, \Lambda) \sim N(\mu, \Lambda^{-1})$$

- The Prior of mean and precision is the NormalWishart distribution:

$$p(\mu, \Lambda) = p(\mu|\Lambda)p(\Lambda) = N(\mu|\mu_0, (\kappa_0\Lambda)^{-1})Wi(\Lambda|\nu, T)$$

# Multivariate normal case: unknown $\mu$ and $\Sigma$

*a posteriori* parameter density

$$p(\mu, \Lambda | \mathcal{D}) = N(\mu | \mu_n (\kappa_n \Lambda)^{-1}) Wi(\Lambda | \nu_n, T_n)$$

where

$$\mu_n = \frac{\kappa_0 \mu_0 + n \hat{\mu}_n}{\kappa_0 + n}$$

$$T_n = T + \sum_{i=1}^n (x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^T + \frac{\kappa n}{\kappa + n} (\mu_0 - \hat{\mu}_n)(\mu_0 - \hat{\mu}_n)^T$$

$$\nu_n = \nu + n \quad \kappa_n = \kappa + n$$

Computing the posterior predictive

$$p(x | \mathcal{D}) = t_{\nu_n - d + 1} \left( x | \mu_n, \frac{T_n (\kappa_n + 1)}{\kappa_n (\nu_n - d + 1)} \right)$$