CS 2750 Machine Learning
Lecture 6

Density estimation

Milos Hauskrecht

milos@cs.pitt.edu

5329 Sennott Square

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)** $\boxed{\max_{\Theta} p(D \mid \Theta, \xi)}$
  - Find Θ that maximizes $p(D \mid \Theta, \xi)$

$$p(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi)$$

$$= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi)$$

$$= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)$$

Independent examples

Log likelihood – has the same maximum as likelihood

$$\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$$

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)** $\boxed{\max_{\Theta} p(D \mid \Theta, \xi)}$
  - Find Θ that maximizes $p(D \mid \Theta, \xi)$

Log likelihood – has the same maximum as likelihood

$$\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$$

The optimum satisfies:

$$\frac{d}{d\Theta} \log p(D \mid \Theta, \xi) = 0$$

It can be often solved analytically

MAP Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum a posterior probability** $\boxed{\max_{\Theta} p(\Theta \mid D, \xi)}$
  - Find Θ that maximizes $p(\Theta \mid D, \xi)$

Likelihood of data

Prior distribution

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes theorem)**

Normalizing factor

Conjugate choices:

- Prior distribution on the parameters **matches** the data distribution
- Posterior is the same type of the distribution as the prior

MAP Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum a posterior probability** $\boxed{\max_{\Theta} p(\Theta \mid D, \xi)}$
  - Find Θ that maximizes $p(\Theta \mid D, \xi)$

- **The optimum satisfies:**

$$\frac{d}{d\Theta} \log p(\Theta \mid D, \xi) = 0 \quad \text{or} \quad \frac{d}{d\Theta} p(\Theta \mid D, \xi) = 0$$

- **It can be often solved analytically**

Bernoulli distribution

Random variable: x

Two outcomes: 0 or 1

Distribution:

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

where θ is the probability of x=1

Example: Coin toss

Outcomes:

- **Head → x=1**
- **Tail → x=0**
- θ **→ probability of a Head**

Bernoulli distribution

Data D: iid sample of n outcomes (coin flips)

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

N_1 - number of 1s

N_2 - number of 0s

Loglikelihood of data:

$$l(D, \theta) = N_1 \log \theta + N_2 \log(1-\theta)$$

ML estimate:

$$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

Bernoulli distribution

Data D: iid sample of n outcomes (coin flips)

Posterior of data:

$$\boxed{p(\theta \mid D,\xi)} = \frac{\boxed{P(D \mid \theta,\xi)}\boxed{p(\theta \mid \xi)}}{P(D \mid \xi)}$$

Likelihood

$$\boxed{P(D \mid \theta,\xi)} = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{N_1}(1-\theta)^{N_2}$$

Conjugate prior:

$$\boxed{p(\theta \mid \xi)} = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Posterior:

$$\boxed{p(\theta \mid D,\xi)} = Beta(\theta \mid \alpha_1+N_1, \alpha_2+N_2)$$

Binomial distribution

Example problem: N coin flips, where each coin flip can have two results: head or tail

Outcome: N_1 - number of heads seen N_2 - number of tails seen in N trials

Model: probability of a head θ
probability of a tail $(1-\theta)$

Probability of an outcome:

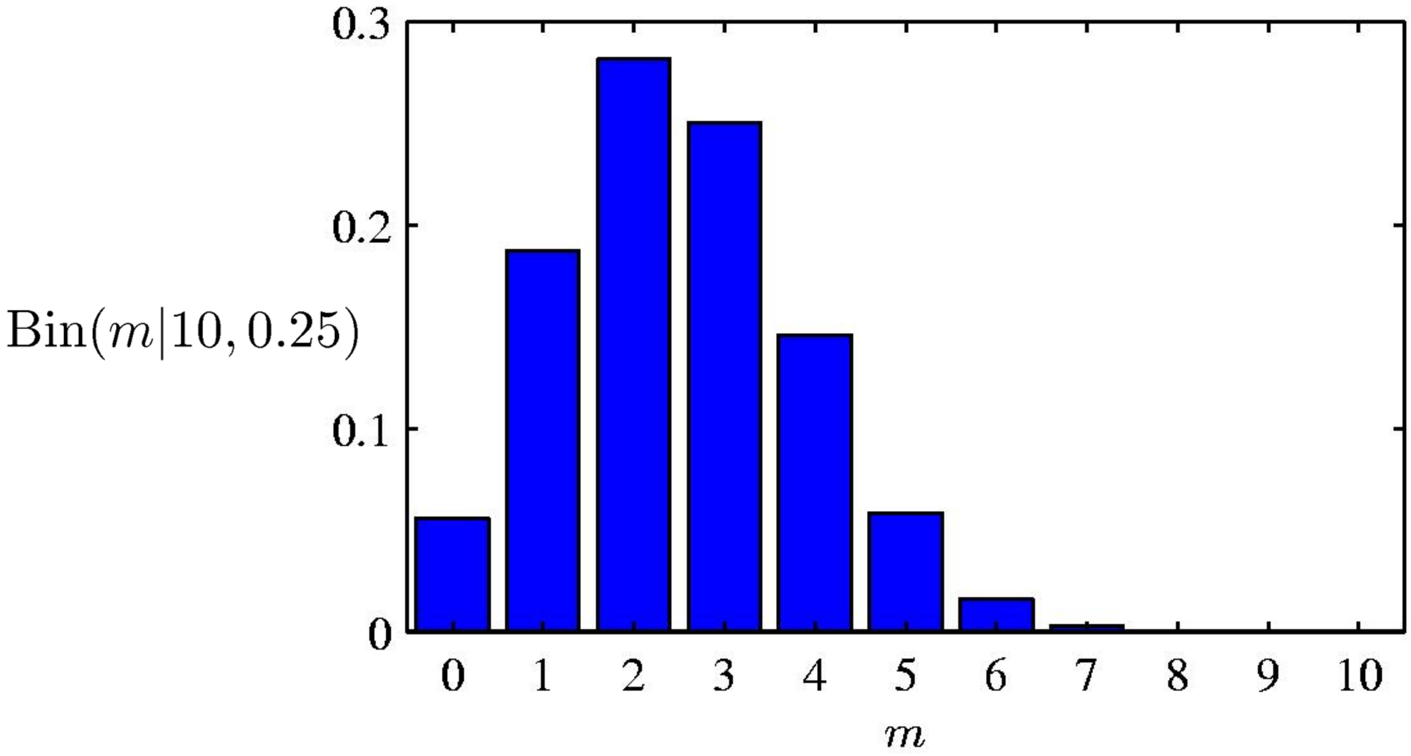
$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1}$$ **Binomial distribution**

Binomial distribution:

- **models order independent sequence of independent Bernoulli trials**

Binomial distribution

Binomial distribution:



Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D\,|\,\theta) = \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2} = \frac{N!}{N_1!N_2!}\theta^{N_1}(1-\theta)^{N_2}$$

Log-likelihood

$$l(D,\theta) = \log\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_2} = \underline{\log\frac{N!}{N_1!N_2!}} + N_1\log\theta + N_2\log(1-\theta)$$

Constant from the point of optimization !!!

ML Solution: $$\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1+N_2}$$

The same as for Bernoulli and D with iid sequence of examples

Posterior density

Posterior density

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Prior choice

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Likelihood

$$P(D \mid \theta) = \frac{\Gamma(N_1+N_2)}{\Gamma(N_1)\Gamma(N_2)}\theta^{N_1}(1-\theta)^{N_2}$$

Posterior

$$p(\theta \mid D,\xi) = Beta(\alpha_1+N_1,\alpha_2+N_2)$$

MAP estimate

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D,\xi)$$

$$\theta_{MAP} = \frac{\alpha_1+N_1-1}{\alpha_1+\alpha_2+N_1+N_2-2}$$

Multinomial distribution

Example: multiple rolls of a dice with 6 results

Outcome: counts of occurrences of k possible outcomes of N trials: N_i - a number of times an outcome i has been seen

$$\sum_{i=1}^{k} N_i = N$$

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ **s.t.** $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability distribution:

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

Posterior and MAP estimate

Choice of the prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1,..,\alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for the multinomial sampling

$$P(D \mid \boldsymbol{\theta}, \xi) = P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Posterior density

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, ..\alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

MAP estimate:

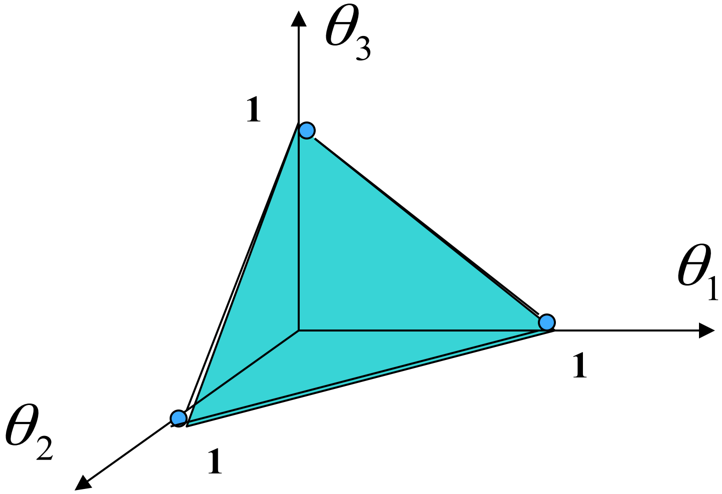
$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

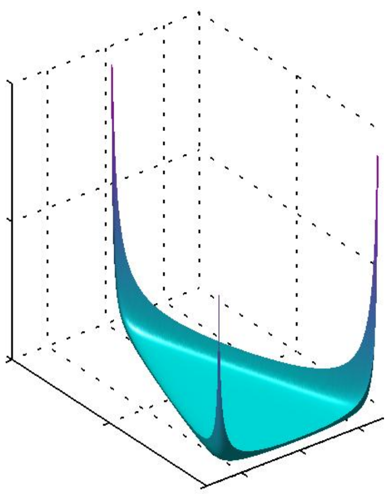
Dirichlet distribution

Dirichlet distribution:

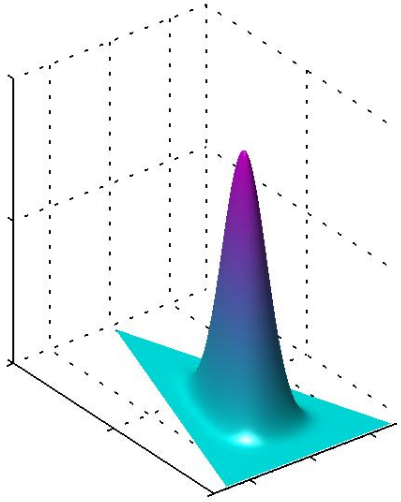
$$Dir(\boldsymbol{\theta} \mid \alpha_1, .., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}$$

Assume: k=3





$\alpha_k = 10^{-1}$



$\alpha_k = 10^{1}$

Other distributions

The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to **“nice” solutions**

- **Exponential family of distributions**

Conjugate choices for some of the distributions from the exponential family:

- **Binomial – Beta**
- **Multinomial - Dirichlet**
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

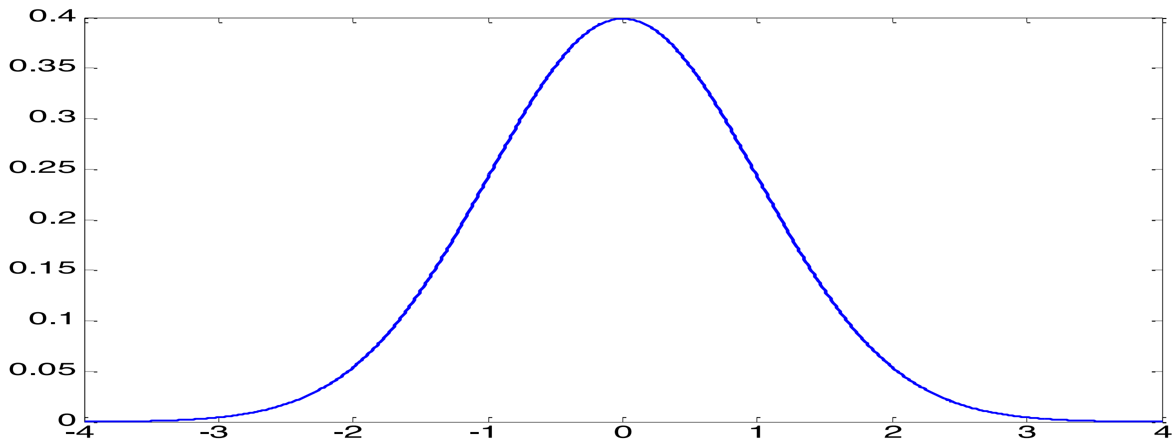
Gaussian (normal) distribution

- **Model of a real-valued outcome**
- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean

  σ - standard deviation
- **Density function:**

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

- **Example:**



$N(0,1)$

Parameter estimates

- **Loglikelihood**

$$l(D,\mu,\sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu,\sigma)$$

- **ML estimates of the mean and variance:**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

  – ML variance estimate is biased

$$E_n(\sigma^2) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

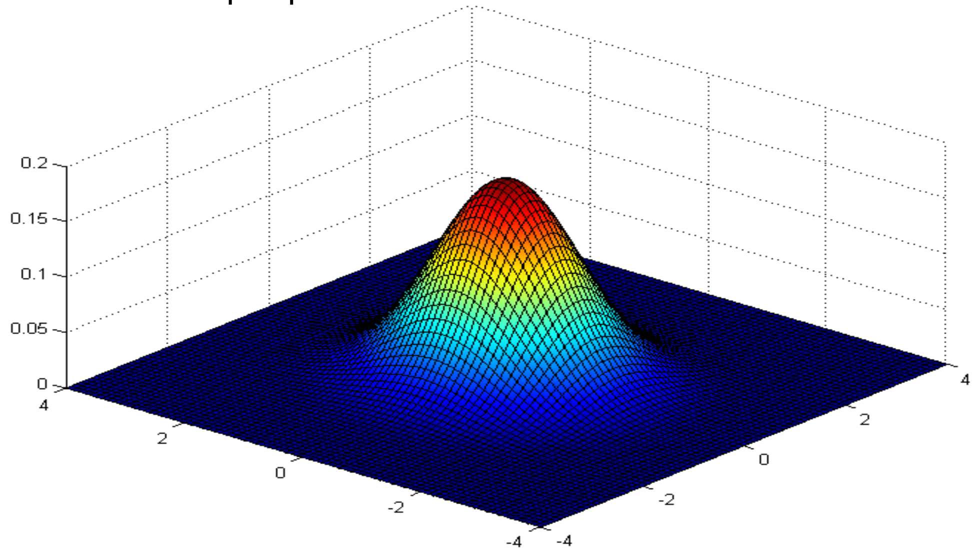
Multivariate normal distribution

- **Multivariate normal:** $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:** $\boldsymbol{\mu}$ - mean

  $\boldsymbol{\Sigma}$ - covariance matrix
- **Density function:**

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right]$$

- **Example:**



Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

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

- **Example:**

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

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

Precision matrix

- **What are the distributions for marginals and conditionals?**

$$p(x_a) \qquad p(x_a \mid x_b)$$

Partitioned Conditionals and Marginals

- **Conditional density:**

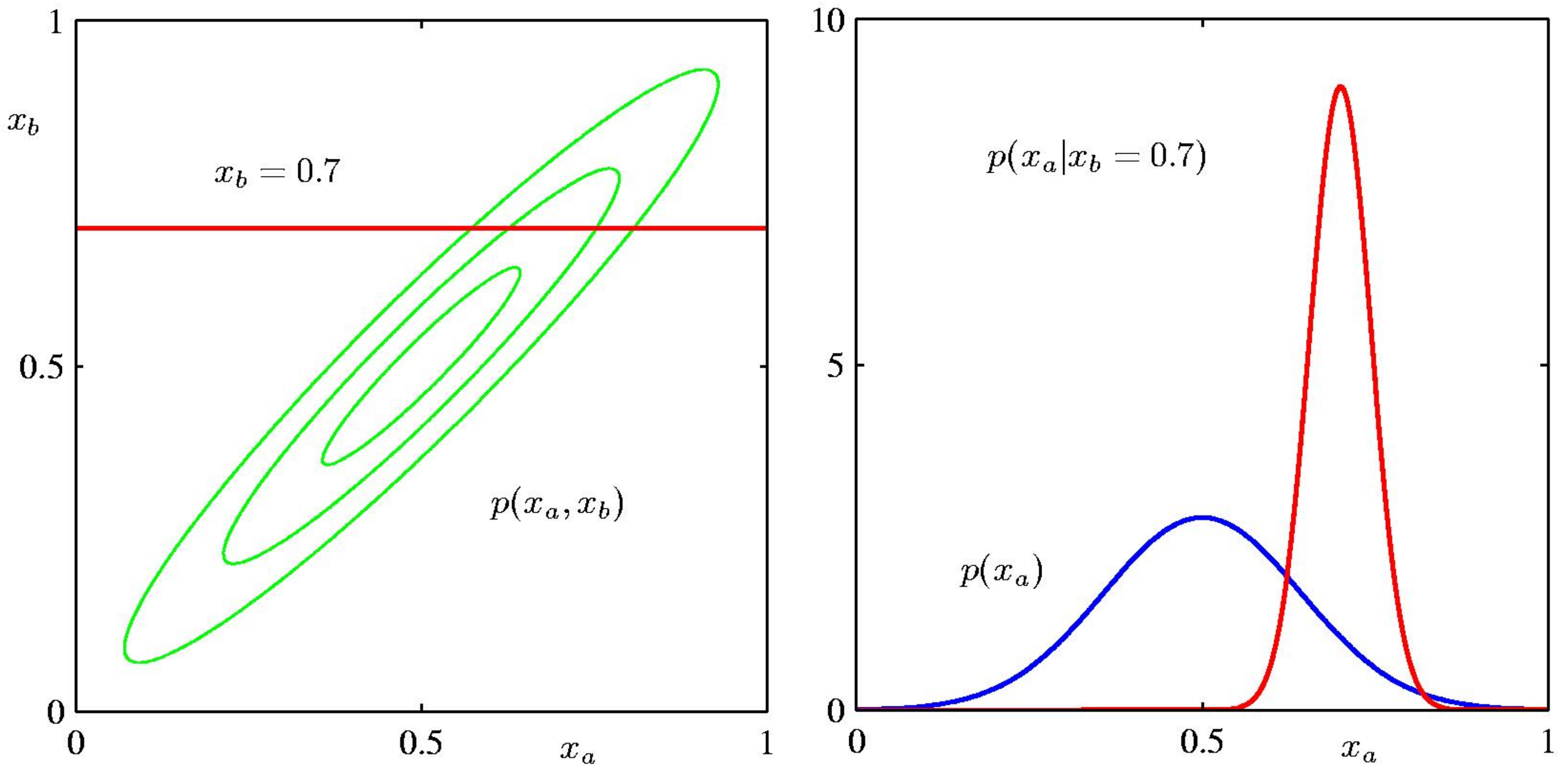
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

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

- **Marginal Density:**

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Partitioned Conditionals and Marginals



Parameter estimates

- **Loglikelihood**

$$l(D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **ML estimates of the mean and covariances:**

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

  – Covariance estimate is biased

$$E_n(\hat{\boldsymbol{\Sigma}}) = E_n\left(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_\mathbf{i} - \hat{\boldsymbol{\mu}})(\mathbf{x}_\mathbf{i} - \hat{\boldsymbol{\mu}})^T \right) = \frac{n-1}{n} \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- **Unbiased estimate:**

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

Posterior of a multivariate normal

- Assume a prior on the mean $\boldsymbol{\mu}$ that is normally distributed:

$$p(\boldsymbol{\mu}) \approx N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

- Then the posterior of $\boldsymbol{\mu}$ is normally distributed

$$p(\boldsymbol{\mu} \mid D) \approx \left(\prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \right)$$

$$* \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_p|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p) \right]$$

$$= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

Posterior of a multivariate normal

- **Then the posterior of $\boldsymbol{\mu}$ is normally distributed**

$$p(\boldsymbol{\mu} \mid D) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_p^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + \frac{1}{n} \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_p$$

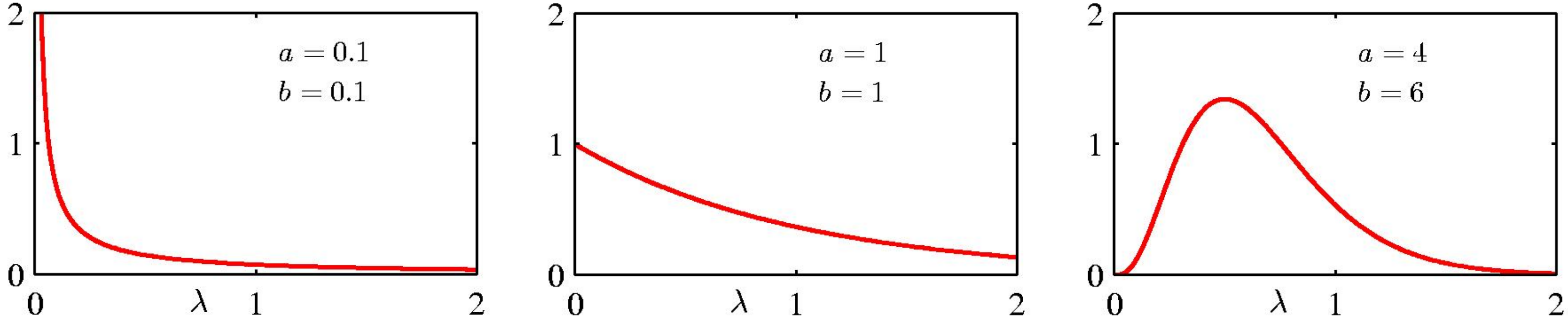
$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \boldsymbol{\Sigma}$$

Other distributions

Gamma distribution:

$$p(\lambda \mid a, b) = \frac{1}{\Gamma(a) b^a} \lambda^{a-1} e^{-\frac{\lambda}{b}} \qquad \text{for} \quad \lambda \in [0, \infty]$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



Other distributions

Exponential distribution:

- A special case of Gamma for a=1

$$p(\lambda \mid b) = \left(\frac{1}{b}\right) e^{-\frac{\lambda}{b}}$$

Uniform distribution:

$$p(x \mid a, b) = \frac{1}{b-a} \qquad \text{for} \quad x \in [a, b]$$

Poisson distribution: models a number of events occurring in a specific time interval

$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Sequential Bayesian parameter estimation

- **Sequential Bayesian approach**
  - Under the iid the estimates of the posterior can be computed incrementally for a sequence of data points

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{\int_{\Theta} p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)\, d\Theta}$$

- If we use a conjugate prior we get back the same posterior
- Assume we split the data D in the last element **x** and the rest

$$p(D \mid \mathbf{\Theta}) = P(x \mid \mathbf{\Theta}) P(D_{n-1} \mid \mathbf{\Theta})$$

- **Then:**

$$p(\Theta \mid D, \xi) = \frac{P(x \mid \mathbf{\Theta}) \overbrace{P(D_{n-1} \mid \mathbf{\Theta})\, p(\Theta \mid \xi)}^{\textbf{A ``new'' prior}}}{\int_{\Theta} P(x \mid \mathbf{\Theta}) P(D_{n-1} \mid \mathbf{\Theta})\, p(\Theta \mid \xi)\, d\Theta}$$

Exponential family

Exponential family:

- all probability mass / density functions that can be written in the exponential normal form

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- $\boldsymbol{\eta}$ a vector of **natural (or canonical) parameters**
- $t(\mathbf{x})$ a function referred to as a **sufficient statistic**
- $h(\mathbf{x})$ a function of x (it is less important)
- $Z(\boldsymbol{\eta})$ a normalization constant (a **partition function**)

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}$$

- Other common form:

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right] \qquad \log Z(\boldsymbol{\eta}) = A(\boldsymbol{\eta})$$