

CS 2750 Machine Learning

Lecture 4

Density estimation II.

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

CS 2750 Machine Learning

Outline

Outline:

- **Density estimation:**
 - Maximum likelihood (ML)
 - Bayesian parameter estimates
 - MAP
- **Bernoulli distribution.** ✓
- **Binomial distribution** ✓
- **Multinomial distribution**
- **Normal distribution**
- **Exponential family**

CS 2750 Machine Learning

Bernoulli trials

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

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

Probability of an outcome of a coin flip

$$P(x_i | \theta) = \theta^{x_i} (1 - \theta)^{(1-x_i)} \leftarrow \text{Bernoulli distribution}$$

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

N_1, N_2 - Number of heads and tails respectively

Posterior distribution

Posterior density

$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) p(\theta | \xi)}{P(D | \xi)} \quad \text{(via Bayes rule)}$$

Annotations:
- Likelihood of data points to $P(D | \theta, \xi)$
- prior points to $p(\theta | \xi)$
- Normalizing factor points to $P(D | \xi)$

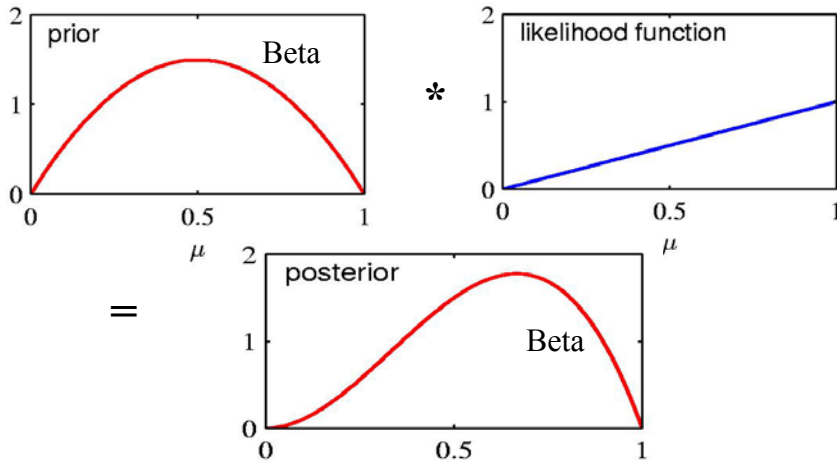
$$P(D | \theta, \xi) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} = \theta^{N_1} (1 - \theta)^{N_2}$$

$p(\theta | \xi)$ - is the prior probability on θ

Conjugate choice of prior: Beta

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

Posterior distribution



$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

CS 2750 Machine Learning

Maximum a posterior probability

Maximum a posteriori estimate

– Selects the mode of the **posterior distribution**

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

$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1) \Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1} (1 - \theta)^{N_2 + \alpha_2 - 1}$$

Notice that parameters of the prior
act like counts of heads and tails

(sometimes they are also referred to as **prior counts**)

MAP Solution:

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

CS 2750 Machine Learning

Binomial distribution

Example: a biased coin

Outcomes: two possible values -- head or tail

Data: D a set of order-independent outcomes

We treat D as a multi-set !!!

N_1 - number of heads seen N_2 - number of tails seen

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

Probability of an outcome

$$P(D | \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} \quad \text{Binomial distribution}$$

CS 2750 Machine Learning

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

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

Constant from the point of optimization !!!

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

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

CS 2750 Machine Learning

Posterior density

Posterior density

$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) p(\theta | \xi)}{P(D | \xi)} \quad (\text{via Bayes rule})$$

Prior choice

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

Likelihood

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

Posterior

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

$$\begin{aligned} \text{MAP estimate} \quad \theta_{MAP} &= \arg \max_{\theta} p(\theta | D, \xi) \\ \theta_{MAP} &= \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2} \end{aligned}$$

CS 2750 Machine Learning

Multinomial distribution

Example: Multi-way coin toss, roll of dice

- **Data:** a set of N outcomes (multi-set)

N_i - a number of times an outcome i has been seen

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ s.t. $\sum_{i=1}^k \theta_i = 1$
 θ_i - probability of an outcome i

Probability of data (likelihood)

$$P(N_1, N_2, \dots, N_k | \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \dots N_k!} \theta_1^{N_1} \theta_2^{N_2} \dots \theta_k^{N_k} \quad \text{Multinomial distribution}$$

ML estimate:

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

CS 2750 Machine Learning

Posterior density and MAP estimate

Choice of the prior: Dirichlet distribution

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

Dirichlet is the conjugate choice for multinomial

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

Posterior density

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

MAP estimate:

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

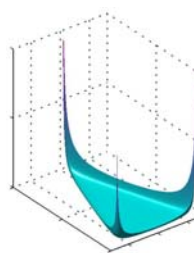
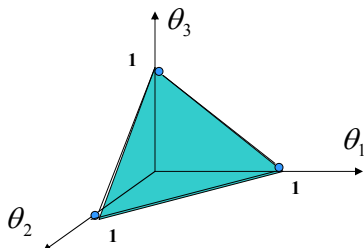
CS 2750 Machine Learning

Dirichlet distribution

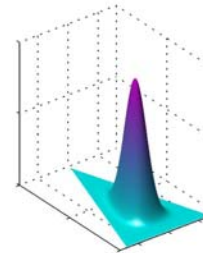
Dirichlet distribution:

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

Assume: k=3



$\alpha_k = 10^{-1}$



$\alpha_k = 10^1$

CS 2750 Machine Learning

Expected value

The result is analogous to the result for binomial

$$E(\boldsymbol{\theta}) = \int_{0 \leq \theta_i \leq 1, \sum \theta_i = 1} \boldsymbol{\theta} \text{Dir}(\boldsymbol{\theta} | \boldsymbol{\eta}) d\boldsymbol{\theta} = \left(\frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \dots, \frac{\eta_i}{\eta_1 + \eta_2 + \eta_k}, \dots, \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right)$$

Expectation based parameter estimate

$$E(\boldsymbol{\theta}) = \left(\frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}, \dots, \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}, \dots, \frac{\alpha_k + N_k}{\alpha_1 + N_1 + \dots + \alpha_k + N_k} \right)$$

Represents the predictive probability of an event $x=i$

$$P(x=i | \boldsymbol{\theta}, \xi) = \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \dots + \alpha_k + N_k}$$

Other distributions

The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to a nice solutions

- Exponential family of distributions

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

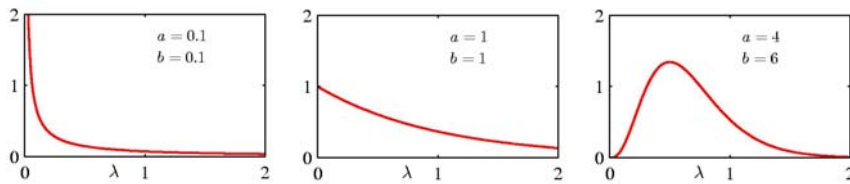
- Binomial – Beta
- Multinomial - Dirichlet
- Exponential – Gamma
- Poisson – Inverse Gamma
- Gaussian - Gaussian (mean) and Wishart (covariance)

Gamma distribution

- Gamma distribution

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

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



CS 2750 Machine Learning

Other distributions

Exponential distribution:

- A special case of Gamma for $a=1$

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

Poisson distribution:

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

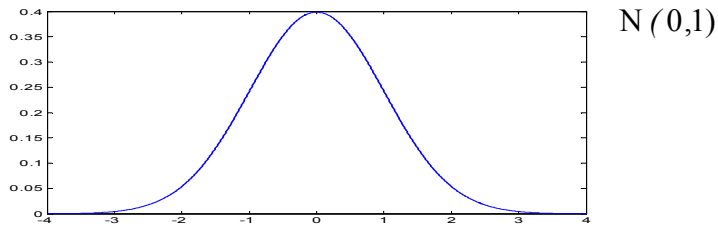
CS 2750 Machine Learning

Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean
 σ - standard deviation
- **Density function:**

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

- **Example:**



CS 2750 Machine Learning

Parameter estimates

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

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

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

– ML variance estimate is biased

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

- **Unbiased estimate:**

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

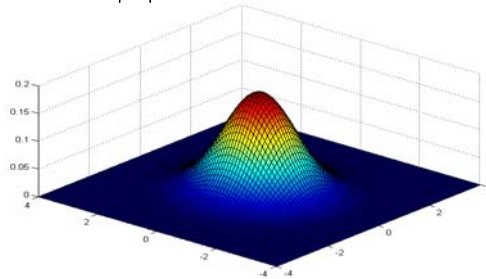
CS 2750 Machine Learning

Multivariate normal distribution

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

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

- **Example:**



CS 2750 Machine Learning

Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

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

– what are marginals and conditionals?

- **Example:**

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

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

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

Precision matrix

CS 2750 Machine Learning

Partitioned Conditionals and Marginals

- Conditional density:

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

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

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

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

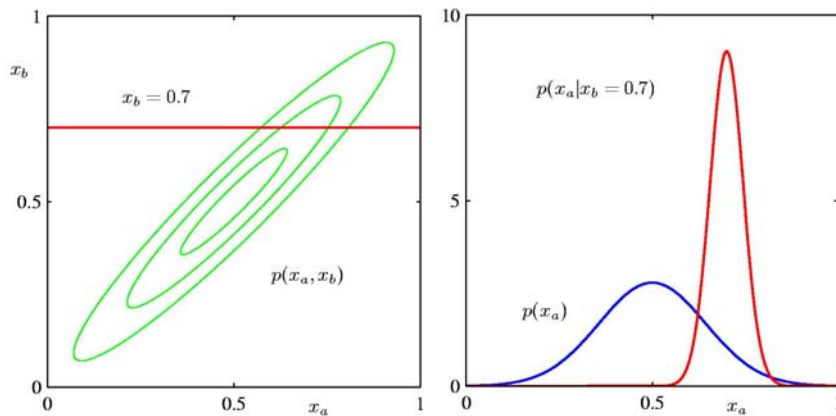
$$= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} (\mathbf{x}_b - \boldsymbol{\mu}_b)$$

- Marginal Density:

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

CS 2750 Machine Learning

Partitioned Conditionals and Marginals



CS 2750 Machine Learning

Parameter estimates

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

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

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

– Covariance estimate is biased

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

- **Unbiased estimate:**

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

CS 2750 Machine Learning

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**

$$\begin{aligned} p(\boldsymbol{\mu} | D) &\approx \left(\prod_{i=1}^n \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\right] \right) \\ &\quad * \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_p|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_p)\right] \\ &= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right] \end{aligned}$$

CS 2750 Machine Learning

Posterior of a multivariate normal

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

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

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

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

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

Sequential Bayesian parameter estimation

- Sequential Bayesian approach

- Under the iid for some densities, the estimates of the posterior can be computed incrementally for a sequence of data points

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

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

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

- Then:

$$p(\Theta | D, \xi) = \frac{P(\mathbf{x} | \Theta) \overbrace{P(D_{n-1} | \Theta) p(\Theta | \xi)}^{\text{A "new" prior}}}{\int_{\Theta} P(\mathbf{x} | \Theta) P(D_{n-1} | \Theta) p(\Theta | \xi) d\Theta}$$