Statistics for Data Science

MSc Data Science WiSe 2019/20

Prof. Dr. Dirk Ostwald

Bayesian inference

(12) Foundations and conjugate inference

Bibliographic remarks

Introductions to Bayesian inference are can found in most statistical textbooks (e.g., Wasserman, 2004; Held and Sabanés Bové, 2014).

Foundations and conjugate inference

- Foundations
  - The Bayesian paradigm
  - Inference summaries
- Conjugate inference
  - The Beta-Binomial model
  - The univariate Gaussian-Gaussian model

Foundations and conjugate inference

- **Foundations**
  - **The Bayesian paradigm**
  - Inference summaries
- Conjugate inference
  - The Beta-Binomial model
  - The univariate Gaussian-Gaussian model

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.
- Parameters are fixed, unknown constants, referred to as *true, but unknown* values. No probability statements are made about parameters.
- Statistical procedures are designed to have good long run frequency properties and are typically assessed by studying their sampling distributions.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like "the probability that it will rain this afternoon is 0.5" are meaningful.

- Parameters are fixed, unknown constants, about which probabilistic statements quantifying our uncertainty about their true, but unknown, value can be made.

- Probabilistic statements about parameters are made with the help of probability distributions, from which further inferences, such as point or interval estimates, can be derived.

Definition (Probabilistic model)

A *probabilistic model* is a joint distribution over a family of observable random variables $X_{1:n} = (X_1, ..., X_n)$, commonly modeling data, and a not directly observable random vector θ, commonly modeling parameters, that is specified in terms of a joint PDF or PMF

$$p(x_{1:n}, \theta). \tag{1}$$

Probabilistic models are also referred to as *generative models*. Typically, probabilistic models are defined in terms of the product of a marginal PDF or PMF of θ and a conditional PDF or PMF of $x_{1:n}$ in the form

$$p(x_{1:n}, \theta) = p(x_{1:n}|\theta)p(\theta). \tag{2}$$

Here, $p(\theta)$ is referred to as *prior distribution* and $p(x_{1:n}|\theta)$ is referred to as the *likelihood*. Often, the observable random variables $X_{1:n}$ are assumed to be *conditionally independent and identically distributed*, i.e.,

$$p(x_{1:n}|\theta) := \prod_{i=1}^{n} p(x_i|\theta) \text{ and } p(x_i|\theta) = p(x_1|\theta) \text{ for } i = 2, ..., n. \tag{3}$$

Definition (Posterior distribution)

Given a probabilistic model $p(x_{1:n}, \theta)$, the conditional distribution of the not directly observable random vector θ given the observable random vector X is referred to as *posterior distribution*. By means of Bayes theorem for PMF or PDFs, the posterior distribution can be evaluated in terms of the PDF or PMF

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta)\, d\theta}. \tag{4}$$

The denominator of the right-hand side of the above is sometimes referred to as *evidence*, such that a mnemonic for the posterior distribution is given by

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{5}$$

Because the evidence is independent of θ and constant for a given realization of X, it corresponds to a normalization constant of the product of prior and likelihood that renders this product a proper PDF or PMF. If only the functional form of the PDF or PMF of the posterior is of interest, the unnormalized version suffices,

$$p(\theta|x_{1:n}) \propto p(x_{1:n}|\theta)p(\theta). \tag{6}$$

Example (Batch Bayesian estimation)

Assume that n data points $x_i, i = 1, ..., n$ are available. Then Bayesian estimation typically proceeds as follows:

(1) Specification of a prior distribution $p(\theta)$.

(2) Specification of the data likelihood given the parameter, where often the observed random variables points $X_i, i = 1, ..., n$ are assumed to be conditionally independent given θ, such that

$$p(x_{1:n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta). \tag{7}$$

(3) *Batch Bayesian estimation* then amounts to evaluating

$$p(\theta|x_{1:n}) = \frac{p(\theta)\prod_{i=1}^{n} p(x_i|\theta)}{\int p(\theta)\prod_{i=1}^{n} p(x_i|\theta)\, d\theta}. \tag{8}$$

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order $x_1, x_2, \ldots$. Then Bayesian estimation can proceed as follows:

(1) Specification of the prior distribution $p(\theta)$.

(2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \ldots, n$ are assumed to be conditionally independent given θ, such that

$$p(x_1|\theta) = p(x_2|\theta) = \ldots \tag{9}$$

(3) *Recursive Bayesian estimation* then amounts to recursively updating the distribution of θ, such that for each $i = 1, 2, \ldots$ the posterior distribution at $i-1$ serves as prior distribution at i, i.e.,

$$\begin{aligned} p(\theta|x_1) &= \frac{p(x_1|\theta)p(\theta)}{\int p(x_1|\theta)p(\theta)\, d\theta} \\ p(\theta|x_{1:2}) &= \frac{p(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1)\, d\theta} \\ &\ldots \end{aligned} \tag{10}$$

Note that for n data points, batch and recursive Bayesian estimation are equivalent:

$$p(\theta|x_{1:n}) = \frac{p(\theta)\prod_{i=1}^n p(x_i|\theta)}{\int p(\theta)\prod_{i=1}^n p(x_i|\theta)\, d\theta} = \frac{p(x_n|\theta)p(\theta|x_{1:n-1})}{\int p(x_n|\theta)p(\theta|x_{1:n-1})\, d\theta}. \tag{11}$$

Definition (Marginal likelihood, model evidence, Bayes factor)

Let $p(x_{1:n}, \theta)$ denote the PMF or PDF a probabilistic model. Then the marginal PDF or PMF of the observable random variables

$$p(x_{1:n}) = \int p(x_{1:n}, \theta)\, d\theta = \int p(x_{1:n}|\theta)p(\theta)\, d\theta \tag{12}$$

is referred to as *marginal (data) likelihood* or *model evidence*. For a fixed set of data observations $x^*_{1:n}$ and two probabilistic models $p_1(x_{1:n}, \theta_1)$ and $p_2(x_{1:n}, \theta_2)$ with identical data outcome and possibly different parameter spaces, the ratio of the marginal likelihoods

$$\text{BF} = \frac{p_1(x^*_{1:n})}{p_2(x^*_{1:n})} \tag{13}$$

is referred to as *Bayes factor* and serves as a basic model comparison criterion in Bayesian statistics.

Remarks

- The evidence is "the probability of a data set under a probabilistic model".
- $p(x_{1:n})$ is sometimes referred to as prior predictive distribution.

Definition (Posterior predictive distribution)

Let $p(x_{1:n}, \theta)$ denote the PDF or PMF of a probabilistic model. Then the conditional distribution of an observable random variable X_{n+1} given the observed random variables $X_{1:n}$ is referred to as *posterior predictive distribution*. A PMF or PDF of the posterior predictive distribution is given by

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n})\, d\theta, \tag{14}$$

where typically

$$p(x_{n+1}|\theta) = p(x_i|\theta) \text{ for } i = 1, ..., n \tag{15}$$

Remark

- In contrast to $p_{\hat{\theta}_n}(x_{n+1})$, $p(x_{n+1}|\theta)$ accounts for estimation uncertainty.

Foundations and conjugate inference

- **Foundations**
  - The Bayesian paradigm
  - **Inference summaries**
- Conjugate inference
  - The Beta-Binomial model
  - The univariate Gaussian-Gaussian model

Bayesian point estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, however, point estimates are often useful.
- Bayesian point estimation is a decision-theoretic problem.
- Decision theory rests on the notions of utility or loss functions.

Definition (Loss function, expected posterior loss, Bayes estimator)

Let $p(x_{1:n}, \theta)$ a probabilistic model and let $\hat{\theta}$ denote a point estimate for the true, but unknown, value of the parameter of the probabilistic model. Then a real-valued function

$$l : \Theta \times \Theta \to \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) \tag{16}$$

that measures the undesirability of choosing the point estimate $\hat{\theta}$, if the true, but unknown, parameter value is θ, is called a *loss function*.

$$l_p : \Theta \to \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int l(\hat{\theta}, \theta) p(\theta | x_{1:n}) \, d\theta. \tag{17}$$

The point estimator

$$\hat{\theta}_B := \underset{\hat{\theta} \in \Theta}{\arg\min} \, l_p(\hat{\theta}) \tag{18}$$

that minimizes the expected posterior loss is known as *Bayes estimator*.

Example (Quadratic loss function, posterior expectation)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$ and let the *quadratic loss function* be defined as

$$l : \Theta \times \Theta \to \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := (\hat{\theta} - \theta)^2. \tag{19}$$

Then the Bayes estimator is the posterior expected value of the parameter, i.e.,

$$\hat{\theta}_B = \arg\min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \int \theta \, p(\theta | x_{1:n}) \, d\theta. \tag{20}$$

This estimator is also referred to as *minimum mean squared error (MMSE)* estimator.

Example (Quadratic loss function, posterior expectation)

Proof

The expected posterior loss function is given by

$$l_p : \Theta \to \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (\hat{\theta} - \theta)^2 p(\theta|x_{1:n})\, d\theta \tag{21}$$

Its derivative with respect to $\hat{\theta}$ evaluates to

$$\frac{d}{d\hat{\theta}} l_p(\hat{\theta}) = \int \frac{d}{d\hat{\theta}} (\hat{\theta} - \theta)^2 p(\theta|x_{1:n})\, d\theta = 2 \int (\hat{\theta} - \theta) p(\theta|x_{1:n})\, d\theta. \tag{22}$$

Setting to zero yields

$$\int (\hat{\theta} - \theta) p(\theta|x_{1:n})\, d\theta = 0 \Leftrightarrow \hat{\theta} - \int \theta p(\theta|x_{1:n})\, d\theta = 0 \Leftrightarrow \hat{\theta} = \int \theta p(\theta|x_{1:n})\, d\theta. \tag{23}$$

□

Example (Zero-one loss function, posterior mode)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model and let the *zero-one loss function* be defined as

$$l : \Theta \times \Theta \to \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := 1 - 1_{\hat{\theta}}(\theta) = \begin{cases} 0, & \theta = \hat{\theta} \\ 1, & \theta \neq \hat{\theta} \end{cases} \tag{24}$$

Then the Bayes estimator is the posterior mode of the parameter, i.e.,

$$\hat{\theta}_B = \arg\min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \arg\max_{\hat{\theta} \in \Theta} p(\hat{\theta}|x_{1:n}). \tag{25}$$

This estimator is also referred to as *maximum-a-postiori (MAP)* estimator.

Example (Zero-one loss, posterior mode)

Proof

The expected posterior loss function is given by

$$\begin{aligned} l_p : \Theta \to \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) &:= \int (1 - 1_{\hat{\theta}}(\theta)) p(\theta|x_{1:n})\, d\theta \\ &= \int 1\, p(\theta|x_{1:n})\, d\theta - \int 1_{\hat{\theta}}(\theta)\, p(\theta|x_{1:n})\, d\theta \\ &= 1 - p(\hat{\theta}|x_{1:n}) \end{aligned} \tag{26}$$

Minimizing $l_p(\hat{\theta})$ with respect to $\hat{\theta}$ is thus equivalent to maximizing $p(\hat{\theta}|x_{1:n})$ with respect to $\hat{\theta}$.

□

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, interval estimates are often useful.
- Bayesian interval estimates are referred to as *credible intervals*.
- Credible intervals are "more intuitive" than confidence intervals.
- *Credible regions* generalize credible intervals for non-scalar parameters.

Definition (δ-credible region, δ-credible interval)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with parameter space Θ and let $p(\theta|x_{1:n})$ denote the posterior distribution. Then any region $R_q \subset \Theta$ such that

$$\int_{R_\delta} p(\theta|x_{1:n})\, d\theta = \delta \text{ for } \delta \in]0, 1[\tag{27}$$

is referred to as a posterior δ-credible region. If $R_q \subset \mathbb{R}$ is an interval, then R_q is also referred to as a *δ-credible interval*.

Remarks

- δ-credible intervals are typically not uniquely defined
- ... but neither are confidence intervals
- Approaches for uniquely specifying δ-credible intervals include selecting
  - highest posterior density (minimum size) δ-credible intervals,
  - symmetric δ-credible intervals around the posterior expectation,
  - equal upper and lower tail probability δ-credible intervals.

Foundations and conjugate inference

- **Foundations**
  - The Bayesian paradigm
  - Inference summaries
- **Conjugate inference**
  - The Beta-Binomial model
  - The univariate Gaussian-Gaussian model

Definition (Conjugate family of distributions and hyperparameters)

Let

$$p(\theta, x_{1:n}) = \prod_{i=1}^{n} p(x_i|\theta)p(\theta) \tag{28}$$

denote a probabilistic model with conditionally independent and identically distributed observed random variables $X_1, ..., X_n$ and unobserved random variable $\theta \in \Theta$.

Let ϕ denote a family of distributions over Θ. ϕ is called a *conjugate family of distributions with respect to* $p(x_i|\theta)$, if the posterior distribution $p(\theta|x_{1:n})$ is an element of ϕ, irrespective of the prior distribution $p(\theta) \in \phi$ and the observations $x_1, ..., x_n$. If both the prior and the posterior distributions are elements of ϕ with respect to $p(x_i|\theta), i = 1, ..., n$, ϕ is also said to be *closed under sampling*. Because θ is often referred to as parameter, the parameters of the distributions in ϕ are often referred to as *hyperparameters*.

Theorem (The Beta-Binomial model)

Consider the probabilistic model

$$p(x, \theta) = p(x|\theta)p(\theta) := \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta). \tag{29}$$

Then the posterior distribution is given by

$$p(\theta|x) = \text{Beta}(\theta; \alpha + x, \beta + n - x) \tag{30}$$

and the MMSE and MAP Bayes estimators are

$$\hat{\theta}_{MMSE} = \frac{\alpha + x}{\alpha + \beta + n} \text{ and } \hat{\theta}_{MAP} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}. \tag{31}$$

Example (Binomial random variable)

Let X be a random variable with outcome set $\mathcal{X} := \mathbb{N}_n^0$ and probability mass function

$$p : \mathcal{X} \to [0,1], x \mapsto p(x) := \binom{n}{x} \mu^x (1-\mu)^{n-x} \text{ for } \mu \in [0,1]. \tag{32}$$

Then X is said to be distributed according to a *Binomial distribution* with parameters $\mu \in [0,1]$ and $n \in \mathbb{N}$, for which we write $X \sim \text{Bin}(\mu, n)$. We denote the probability mass function of a Binomial random variable by

$$\text{Bin}(x; \mu, n) := \binom{n}{x} \mu^x (1-\mu)^{n-x} \tag{33}$$

Remark

- $\text{Bin}(x; \mu, 1) = \text{Bern}(x; \mu)$.

Example (Beta random variable)

Let X be a random variable with outcome set $\mathcal{X} := [0, 1]$ and probability density function

$$p : \mathcal{X} \to [0, 1], x \mapsto p(x) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \text{ for } \alpha, \beta \in \mathbb{R}_{>0}, \tag{34}$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a *Beta distribution* with parameters α, β, for which we write $X \sim \text{Beta}(\alpha, \beta)$. We denote the probability density function of a Beta random variable by

$$\text{Beta}(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}. \tag{35}$$

Remarks

- A Beta random variable can be used to model the distribution of a probability.
- $\text{Beta}(x; 1, 1) = U(x; 0, 1)$.
- For $\alpha < 1, \beta < 1$ the outcome set is $\mathcal{X} :=]0, 1[$.

Proof

Posterior distribution

We first note that up to proportionality constants, the posterior distribution is given by

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &= \text{Bin}(x;\theta,n)\text{Beta}(\theta;\alpha,\beta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned} \tag{36}$$

The right-hand side of the above corresponds to the kernel of the PDF of a Beta distribution with parameters $\alpha + x$ and $\beta + n - x$. The posterior distribution is thus given by

$$p(\theta|x) = \text{Beta}(\theta;\alpha+x,\beta+n-x) \tag{37}$$

MMSE and MAP Bayes estimators

The expected value and the mode of a random variable $X \sim \text{Beta}(\alpha,\beta)$ are given by

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \text{ and } \mathbb{M}(X) = \frac{\alpha-1}{\alpha+\beta-2}(\alpha,\beta>1) \tag{38}$$

Substitution of the posterior distribution parameters thus directly yields the MMSE and MAP Bayes estimators.

Remarks

- The MMSE estimator of the Beta-Binomial model can be written as

$$\begin{aligned}\hat{\theta}_{MMSE} &= \frac{\alpha + x}{\alpha + \beta + n} \\ &= \frac{\alpha}{\alpha + \beta + n} + \frac{x}{\alpha + \beta + n} \\ &= \frac{\alpha(\alpha + \beta)}{(\alpha + \beta + n)(\alpha + \beta)} + \frac{xn}{(\alpha + \beta + n)n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}\end{aligned} \tag{39}$$

- It is thus a weighted average of prior expectation and ML estimator.
- The weighting constant are proportional to
  - the number of virtual prior observation $\alpha + \beta$,
  - the number of actual observations n.

Theorem (The univariate Gaussian-Gaussian model)

Consider the probabilistic model

$$p(x_{1:n}, \theta) = \prod_{i=1}^{n} p(x_i|\theta)p(\theta) := \prod_{i=1}^{n} N(x_i; \theta, \sigma_x^2)N(\theta; \mu_\theta, \sigma_\theta^2). \tag{40}$$

Then the posterior distribution is given by

$$p(\theta|x_{1:n}) = N\left(\theta; \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}\left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right), \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}\right) \tag{41}$$

and the MMSE and MAP Bayes estimators are

$$\hat{\theta}_{MMSE} = \hat{\theta}_{MAP} = \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}\left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right). \tag{42}$$

Proof

Posterior distribution

By focusing on the fact that the posterior distribution is a function of θ and hence subsuming multiplicative constants independent of θ in proportionality statements, we have

$$\begin{aligned}
p(\theta|x_{1:n}) &\propto \prod_{i=1}^{n} N(x_i;\theta,\sigma_x^2)N(\theta;\mu_\theta,\sigma_\theta^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}(x_i-\theta)^2\right) \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta-\mu_\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\sigma_x^2} + \frac{(\theta-\mu_\theta)^2}{\sigma_\theta^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(x_i^2-2x_i\theta+\theta^2)}{\sigma_x^2} + \frac{\theta^2-2\theta\mu_\theta+\mu_\theta^2}{\sigma_\theta^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}x_i^2-2n\bar{x}\theta+n\theta^2}{\sigma_x^2} + \frac{\theta^2-2\theta\mu_\theta+\mu_\theta^2}{\sigma_\theta^2}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{-2n\bar{x}\theta+n\theta^2}{\sigma_x^2} + \frac{\theta^2-2\theta\mu_\theta}{\sigma_\theta^2}\right)\right).
\end{aligned} \tag{43}$$

Proof (cont.)

Hence

$$\begin{aligned} p(\theta|x_{1:n}) &\propto \exp\left(-\frac{1}{2}\left(\frac{\theta^2 n - 2n\bar{x}\theta}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\theta^2\frac{n}{\sigma_x^2} - 2\theta\frac{n\bar{x}}{\sigma_x^2} + \theta^2\frac{1}{\sigma_\theta^2} - 2\theta\frac{\mu_\theta}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\theta^2\frac{n}{\sigma_x^2} - \frac{1}{2}\theta^2\frac{1}{\sigma_\theta^2} + \theta\frac{n\bar{x}}{\sigma_x^2} + \theta\frac{\mu_\theta}{\sigma_\theta^2}\right) \\ &= \exp\left(-\frac{1}{2}\theta^2\left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right) + \theta\left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right)\right). \end{aligned} \tag{44}$$

By defining

$$\phi_1 := \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \text{ and } \phi_2 := \phi_1\left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right) \tag{45}$$

we then have

$$\begin{aligned} p(\theta|x_{1:n}) &\propto \exp\left(-\frac{1}{2\phi_1}\theta^2 + \frac{1}{\phi_1}\theta\phi_2\right) \\ &\propto \exp\left(-\frac{1}{2\phi_1}\theta^2 + \frac{1}{\phi_1}\theta\phi_2 - \frac{1}{2\phi_1}\phi_2^2\right). \end{aligned} \tag{46}$$

Proof (cont.)

Hence

$$p(\theta|x_{1:n}) \propto \exp\left(-\frac{1}{2\phi_1}(\theta-\phi_2)^2\right) \tag{47}$$

Based on the normalization constant of the Gaussian probability density function, we thus have

$$\begin{aligned} p(\theta|x_{1:n}) &= \frac{1}{\sqrt{2\pi\phi_1}} \exp\left(-\frac{1}{2\phi_1}(\theta-\phi_2)^2\right) \\ &= N(\theta;\phi_2,\phi_1), \end{aligned} \tag{48}$$

such that the posterior distribution is given by a univariate Gaussian probability density function with expectation parameter

$$\phi_2 = \left(\frac{n}{\sigma_x^2}+\frac{1}{\sigma_\theta^2}\right)^{-1}\left(\frac{\mu_\theta}{\sigma_\theta^2}+\frac{n\bar{x}}{\sigma_x^2}\right) \tag{49}$$

and variance parameter

$$\phi_1 = \left(\frac{n}{\sigma_x^2}+\frac{1}{\sigma_\theta^2}\right)^{-1}. \tag{50}$$

□

Remarks

- The MMSE estimator of the univariate Gaussian-Gaussian model has the form

$$\hat{\theta}_{MMSE} \propto \frac{1}{\sigma_\theta^2}\mu_\theta + \frac{n}{\sigma_x^2}\bar{x}. \tag{51}$$

- It is thus a weighted average of the prior expectation and the ML estimator.
- The weighting constants are given by
  - the prior precision (reciprocal variance) $1/\sigma_\theta^2$,
  - the data precision $1/\sigma_x^2$ and the number of observations.
- The posterior variance parameter

$$\left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \tag{52}$$

  is reciprocally related to the number observations.

Further often encountered conjugate models include

- the Dirichlet-Multinoulli model (Übung),
- the Gaussian-Gamma-Gaussian model (Übung),
- the multivariate Gaussian-Gaussian model,
- the Gaussian-Wishart-Gaussian model.

A comprehensive theory for conjugate analysis is afforded by

- Conjugate inference in the exponential family.

References

Held, L. and Sabanés Bové, D. (2014). *Applied statistical inference*, volume 10. Springer.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer.