BAYESIAN INFERENCE

x — observable random variable;

θ — "true state of nature";

$p(x \,|\, \theta)$ or $p_{x|\theta}(x \,|\, \theta)$ – data model, likelihood [the same as the data model in the non-Bayesian approach, which we denoted as $p(x; \theta)$];

$\pi(\theta)$ or $\pi_\theta(\theta)$ – prior distribution on θ (epistemic probability), our knowledge about the true state of nature.

The Bayesian approach implies assigning a prior distribution on parameter θ and then applying the Bayes' rule. Note that θ is often *not* really random.

Inference is based on the posterior distribution of θ:

$$p_{\theta|x}(\theta \,|\, x) = \frac{p_{x,\theta}(x, \theta)}{\underbrace{\int p_{x,\theta}(x, \vartheta)\, d\vartheta}_{\text{does not depend on } \theta}} \propto p_{x,\theta}(x, \theta)$$

and, therefore,

$$p(\theta \,|\, x) \propto \underbrace{p(x \,|\, \theta)\, \pi(\theta)}_{p(x,\theta)}.$$

Note:

- $p(\theta \,|\, x)$ is also an epistemic probability.
- It is important to master the use of $\propto$.

Conjugate Priors

If F is a class of measurement models and P a class of prior distributions, then P is *conjugate* for F if $\pi(\theta) \in P$ and $p(x \,|\, \theta) \in F$ implies $p(\theta \,|\, x) \in P$. It is useful to choose conjugate priors: they are computationally convenient as they allow finding analytically tractable posteriors.

Important special case: If $F \equiv$ exponential family $\Longrightarrow$ distributions in F have natural conjugate priors. Consider

$$p(x_i \,|\, \theta) = h(x_i)\, g(\theta)\, \exp[\boldsymbol{\phi}(\theta)^T \boldsymbol{t}(x_i)], \quad i = 1, 2, \ldots, N.$$

For *(conditionally)* independent, identically distributed (i.i.d.) x_i *(given θ)*, the likelihood function is

$$p(\boldsymbol{x} \,|\, \theta) = \Big[\prod_{i=1}^{N} h(x_i)\Big] \cdot [g(\theta)]^N \exp[\boldsymbol{\phi}(\theta)^T \boldsymbol{t}(\boldsymbol{x})]$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$ and the sufficient statistic $\boldsymbol{t}(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{u}(x_i)$. Consider the following prior pdf/pmf:

$$\pi(\theta) \propto g(\theta)^{\eta} \exp[\boldsymbol{\phi}(\theta)^T \boldsymbol{\nu}].$$

Then, the posterior pdf/pmf is

$$p(\theta \,|\, \boldsymbol{x}) \propto [g(\theta)]^{n+\eta} \exp\{\boldsymbol{\phi}(\theta)^T [\boldsymbol{t}(\boldsymbol{x}) + \boldsymbol{\nu}]\}$$

and hence $\pi(\theta)$ is indeed the conjugate prior for $p(\boldsymbol{x} \,|\, \theta)$.

Sequential Bayesian Idea

Suppose that we have observed x_1 and x_2 where x_1 comes first, e.g. the subscript is a time index. We wish to do inference about θ. Then, if we treat x_1 as a fixed (known) quantity, we have:

$$p(x_2, \theta \,|\, x_1) = p(x_2 \,|\, \theta, x_1) \cdot p(\theta \,|\, x_1). \tag{1}$$

where

$$p(x_2 \,|\, x_1, \theta) \equiv \text{new, updated likelihood for } \theta \text{ based on } x_2$$

and

$$p(\theta \,|\, x_1) \equiv \text{new, updated prior for } \theta.$$

Clearly, (1) implies

$$p(\theta \,|\, x_1, x_2) \propto p(x_2 \,|\, \theta, x_1) \cdot p(\theta \,|\, x_1). \tag{2}$$

Conditionally independent observations x_1 and x_2. In the special case where x_1 and x_2 are conditionally independent given θ, we have

$$p(x_1, x_2 \,|\, \theta) = p(x_1 \,|\, \theta) \cdot p(x_2 \,|\, \theta) \tag{3}$$

and, consequently (and clearly),

$$p(x_2 \,|\, x_1, \theta) = p(x_2 \,|\, \theta). \tag{4}$$

[A rather long way to get from (3) to (4) would be as follows:

$$p(x_2 \,|\, x_1, \theta) \propto p(x_2, x_1 \,|\, \theta) = p(x_2 \,|\, \theta) \cdot p(x_1 \,|\, \theta) \propto p(x_2 \,|\, \theta)$$

— it is a good practice for familiarizing with the $\propto$ notation.]

Substituting (4) into (1) and (2) yields

$$p(x_2, \theta \,|\, x_1) = \underbrace{p(x_2 \,|\, \theta)}_{\text{ordinary likelihood based on } x_2} \cdot \underbrace{p(\theta \,|\, x_1)}_{\text{new prior}} \tag{5}$$

and

$$p(\theta \,|\, x_1, x_2) \propto p(x_2 \,|\, \theta) \cdot p(\theta \,|\, x_1). \tag{6}$$

A Side Comment (Exercise). Make sure that you understand the following:

$$\begin{aligned} p(\theta \,|\, x_1, x_2) &\propto p(\theta, x_1, x_2) \\ &\propto p(\theta, x_1 \,|\, x_2) \\ &\propto p(\theta, x_2 \,|\, x_1). \end{aligned}$$

A Bit About Prediction

We continue with the scenario described on the last two pages. Suppose that we have observed x_1 and wish to predict x_2. For this purpose, we use the *posterior predictive distribution*:

$$p(x_2 \,|\, x_1) \tag{7}$$

A good way to think of this predictive distribution is as follows. Given x_1, we have [see (1)]:

$$p(x_2, \theta \,|\, x_1) = p(x_2 \,|\, \theta, x_1) \cdot p(\theta \,|\, x_1).$$

Let us marginalize this pdf with respect to the unknown parameter θ (i.e. integrate θ out):

$$p_{x_2 \,|\, x_1}(x_2 \,|\, x_1) = \int p(x_2, \theta \,|\, x_1)\, d\theta = \int p(x_2 \,|\, \theta, x_1) \cdot p(\theta \,|\, x_1)\, d\theta.$$

Conditionally independent observations x_1 and x_2. In the special case where x_1 and x_2 are conditionally independent given θ, i.e.

$$p(x_1, x_2 \,|\, \theta) = p(x_1 \,|\, \theta) \cdot p(x_2 \,|\, \theta) \quad \Leftrightarrow \quad p(x_2 \,|\, x_1, \theta) = p(x_2 \,|\, \theta)$$

we have:

$$p(x_2 \,|\, x_1) = \int p(x_2 \,|\, \theta) \cdot p(\theta \,|\, x_1)\, d\theta. \tag{8}$$

The First (Ever) Bayesian Model: Binomial Measurements

Suppose that, given θ, x_1 and x_2 are independent, coming from $x_i \,|\, \theta \sim \mathrm{Bin}(n_i, \theta)$, $i = 1, 2$, i.e. the likelihood is

$$p(x_1, x_2 \,|\, \theta) = \binom{n_1}{x_1} \theta^{x_1} (1-\theta)^{n_1 - x_1} \cdot \binom{n_2}{x_2} \theta^{x_2} (1-\theta)^{n_2 - x_2}.$$

We pick a conjugate prior pdf for θ:

$$\pi(\theta) = \mathrm{Beta}(\alpha, \beta) \propto \theta^{\alpha - 1} \, (1-\theta)^{\beta - 1}$$

see the table of distributions. Therefore, the posterior pdf of θ is

$$\begin{aligned} p(\theta \,|\, x_1, x_2) &\propto p(x_1, x_2 \,|\, \theta) \, \pi(\theta) \\ &\propto \theta^{x_1 + x_2 + \alpha - 1} \, (1-\theta)^{n_1 + n_2 - x_1 - x_2 + \beta - 1} \cdot i_{(0,1)}(\theta) \end{aligned}$$

which we recognize as the kernel of the $\mathrm{Beta}(x_1 + x_2 + \alpha, \beta + n_1 - x_1 + n_2 - x_2)$ pdf, see the table of distributions. Hence,

$$p(\theta \,|\, x_1, x_2) = \mathrm{Beta}(x_1 + x_2 + \alpha, \beta + n_1 - x_1 + n_2 - x_2).$$

How about the posterior pdf $p(x_1 \,|\, \theta)$ based only on x_1? Now,

$$p(\theta \,|\, x_1) \propto p(x_1 \,|\, \theta)\, \pi(\theta) \propto \theta^{x_1+\alpha-1}\,(1-\theta)^{n_1-x_1+\beta-1} \cdot i_{(0,1)}(\theta)$$

which is the kernel of the $\mathrm{Beta}(x_1+\alpha, \beta+n_1-x_1)$ pdf, i.e.

$$p(\theta \,|\, x_1) = \mathrm{Beta}(x_1+\alpha, \beta+n_1-x_1).$$

Since, given θ, x_1 and x_2 are independent, we use (6) and verify its validity:

$$p(x_2 \,|\, \theta) \cdot p(\theta \,|\, x_1) = \binom{n_2}{x_2} \theta^{x_2}(1-\theta)^{n_2-x_2}$$

$$\cdot \frac{\Gamma(\overbrace{x_1+\alpha+\beta+n_1-x_1}^{\alpha+\beta+n_1})}{\Gamma(x_1+\alpha)\Gamma(\beta+n_1-x_1)}\, \theta^{x_1+\alpha-1}\,(1-\theta)^{\beta+n_1-x_1-1}$$

$$\underbrace{\propto}_{\text{keep track of } \theta} \underbrace{\theta^{x_1+x_2+\alpha-1}\,(1-\theta)^{n_1+n_2-x_1-x_2+\beta-1}}_{\text{kernel of } \mathrm{Beta}(x_1+x_2+\alpha, \beta+n_1-x_1+n_2-x_2)}$$

for $\theta \in (0,1)$, which is exactly the posterior pdf $p(\theta \,|\, x_1, x_2)$. **Therefore, we can either do our inference "in sequential steps" or "in batch" — both yield the same answer!**

How about predicting x_2 based on x_1? Since, given θ, x_1 and x_2 are independent, we apply (8):

$$\begin{aligned}
&\int_0^1 p(x_2 \,|\, \theta) \cdot p(\theta \,|\, x_1)\, d\theta \\
&= \binom{n_2}{x_2} \cdot \frac{\Gamma(\alpha+\beta+n_1)}{\Gamma(x_1+\alpha)\Gamma(\beta+n_1-x_1)} \\
&\cdot \underbrace{\int_0^1 \theta^{x_1+x_2+\alpha-1}\, (1-\theta)^{\beta+n_1-x_1+n_2-x_2-1}\, d\theta}_{\frac{\Gamma(x_1+x_2+\alpha)\Gamma(\beta+n_1-x_1+n_2-x_2)}{\Gamma(\alpha+\beta+n_1+n_2)}} \\
&= \binom{n_2}{x_2} \cdot \frac{\Gamma(\alpha+\beta+n_1)}{\Gamma(x_1+\alpha)\Gamma(\beta+n_1-x_1)} \\
&\cdot \frac{\Gamma(x_1+x_2+\alpha)\Gamma(\beta+n_1-x_1+n_2-x_2)}{\Gamma(\alpha+\beta+n_1+n_2)} = p(x_2 \,|\, x_1)
\end{aligned}$$

which is the desired predictive pmf of x_2 given x_1.

Comments:

- Here, we have used the fact that the $\text{Beta}(\alpha, \beta)$ pdf of a random variable θ (say) has the following form (see the distribution table):

$$p_\theta(\theta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\text{normalizing constant}} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

implying that $\int_0^1 \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}\, d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

- Bayes used a special case of the above model.

- Laplace computed posterior probabilities under a special case of this model. In particular, he considered a single observation x_1 (the number of girls born in Paris over a certain time interval in the 18th century) coming from

$$x_1 \,|\, \theta \sim \text{Bin}\Big(n_1, \underbrace{\theta}_{\text{prob. that a newborn child is a girl}}\Big)$$

and set the following prior pdf:

$$\pi(\theta) = \text{uniform}(0,1) = \text{Beta}(1,1).$$

Here is the measurement:

$$x_1 = 241,945$$

and $n_1 = 241,945 + 251,527$. Laplace computed

$$P[\theta \geq 0.5 \,|\, x_1] \approx 10^{-42}.$$

An Example of Bayesian Inference: DC-level Estimation in AWGN with Known Variance

See Ch. 2.6 in

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Second ed. New York: Chapman & Hall, 2004.

Single Observation. Let us choose the data model: $p(x \,|\, \theta) = \mathcal{N}(\theta, \sigma^2)$ where σ^2 is assumed known. Hence, the likelihood for one measurement is

$$p(x \,|\, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\Big[-\frac{1}{2\,\sigma^2}(x-\theta)^2\Big]. \tag{9}$$

Viewed as a function of θ (as usual), the likelihood has the form:

$$p(x \,|\, \theta) \quad \underbrace{\propto}_{\text{keep track of } \theta} \quad \exp\Big[-\underbrace{\frac{1}{2\,\sigma^2}\,(\theta^2 - 2x\theta + x^2)}_{\text{quadratic in } \theta}\Big]$$

$$\propto \quad \underbrace{\exp\Big[-\frac{1}{2\,\sigma^2}\,(\theta^2 - 2x\theta)\Big]}_{\text{kernel of the Gaussian pdf with mean } x \text{ and variance } \sigma^2}$$

$$\propto \quad \underbrace{\exp\Big(-\frac{1}{2\,\sigma^2}\,\theta^2 + \frac{x\theta}{\sigma^2}\Big)}_{\text{kernel of the Gaussian pdf with mean } x \text{ and variance } \sigma^2}$$

Then, according to the results from p. 3,

$$\begin{aligned} g(\theta) &= \exp\Big(-\frac{1}{2\,\sigma^2}\,\theta^2\Big) \\ t(x) &= x \\ \phi(\theta) &= \frac{\theta}{\sigma^2} \end{aligned}$$

and the conjugate prior pdf for θ has the following form:

$$\pi(\theta) \propto \underbrace{\exp(\overbrace{A}^{\text{const}}\;\theta^2)}_{g(\theta)^\eta}\;\underbrace{\exp(\overbrace{B}^{\text{const}}\;\theta)}_{\exp[\nu\phi(\theta)]}$$

which can be reparametrized as (where we assume that A is negative):

$$\pi(\theta) \propto \exp\Big[-\frac{1}{2\tau_0^2}\cdot(\theta-\mu_0)^2\Big]$$

and we conclude that $\pi(\theta) = \mathcal{N}(\mu_0, \tau_0^2)$ is a conjugate prior for the model (9). Here, μ_0 and τ_0^2 are *hyperparameters*, considered *known*. (Of course, we could also go on and put a prior on the hyperparameters, which would lead to a hierarchical Bayesian model.)

We now compute the posterior pdf by collecting the terms containing θ and θ^2:

$$
\begin{aligned}
p(\theta \,|\, x) &\propto p(x \,|\, \theta)\, \pi(\theta) \\
&\propto \exp\Big[-\frac{1}{2} \cdot \big(\frac{x^2 - 2x\theta + \theta^2}{\sigma^2} + \frac{\theta^2 - 2\mu_0\theta + \mu_0^2}{\tau_0^2}\big)\Big] \\
&\propto \exp\Big[-\frac{1}{2} \cdot \frac{(\tau_0^2 + \sigma^2)\theta^2 - 2(x\tau_0^2 + \mu_0\sigma^2)\theta}{\sigma^2\tau_0^2}\Big] \\
&\propto \exp\Big[-\frac{1}{2} \cdot \underbrace{\frac{\sigma^2 + \tau_0^2}{\sigma^2\tau_0^2}}_{1/\tau_1^2} \cdot \big(\theta^2 - 2 \cdot \underbrace{\frac{x\tau_0^2 + \mu_0\sigma^2}{\sigma^2 + \tau_0^2}}_{\mu_1}\, \theta\big)\Big]
\end{aligned}
$$

implying that $p(\theta \,|\, x)$ is a Gaussian pdf with mean and variance

$$
\begin{aligned}
\mu_1 &= \frac{x\tau_0^2 + \mu_0\sigma^2}{\sigma^2 + \tau_0^2} \\
\tau_1^2 &= \frac{\sigma^2\tau_0^2}{\sigma^2 + \tau_0^2}.
\end{aligned}
$$

We will generalize the above expressions to multiple measurements.

Comments on the Single Observation Case:

- The posterior mean is a weighted average of the observation and the prior mean:

$$
\begin{aligned}
\mu_1 &= \frac{x\tau_0^2 + \mu_0\sigma^2}{\sigma^2 + \tau_0^2} = \frac{\frac{1}{\sigma^2}\, x + \frac{1}{\tau_0^2}\, \mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}} \\
&= \frac{\text{likelihood precision} \cdot x + \text{prior precision} \cdot \mu_0}{\text{likelihood precision} + \text{prior precision}}.
\end{aligned}
$$

- We will show that the posterior mean is the (Bayesian) MMSE estimate of θ.

- Here, the weights are given by *precisions* $\frac{1}{\sigma^2}$ and $\frac{1}{\tau_0^2}$. (The inverse of the variance of a Gaussian distribution is called *precision*.)

- As the likelihood precision $\frac{1}{\sigma^2}$ increases (i.e. $\sigma^2 \to 0$),

$$\mu_1 \to x.$$

- As the prior precision $\frac{1}{\tau_0^2}$ increases (i.e. $\tau_0^2 \to 0$),

$$\mu_1 \to \mu_0.$$

- The posterior mean is the data *"shrunk"* towards the prior mean:

$$\mu_1 = x - \frac{\sigma^2}{\sigma^2 + \tau_0^2} \cdot (x - \mu_0)$$

or the prior mean adjusted towards the observed data:

$$\mu_1 = \mu_0 + \frac{\tau_0^2}{\sigma^2 + \tau_0^2} \cdot (x - \mu_0).$$

- Posterior precision is the sum of the prior and data precisions:

$$\frac{1}{\tau_1^2} = \frac{\sigma^2 + \tau_0^2}{\sigma^2 \tau_0^2} = \frac{1}{\sigma^2} + \frac{1}{\tau_0^2}.$$

Multiple I.I.D. Observations. Consider now N *(conditionally)* i.i.d. observations $x_1, x_2, \ldots, x_N$ *(given θ)*:

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}) &\propto \pi(\theta)\, p(\boldsymbol{x} \,|\, \theta) \\ &\propto \exp\Big[-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2 \Big] \cdot \prod_{i=1}^{N} \exp\Big[-\frac{1}{2\,\sigma^2}(x_i - \theta)^2 \Big] \end{aligned}$$

where $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^T$. This posterior pdf depends on $\boldsymbol{x}$ only through the sample mean

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

i.e. $\overline{x}$ is the *sufficient statistic* for this model. Note that

$$\overline{x} \,|\, \theta \quad \sim \quad \underbrace{\mathcal{N}(\theta, \sigma^2/N)}_{\substack{\text{our new likelihood} \\ \text{(using sufficiency)}}} \quad .$$

Hence

$$p(\theta \,|\, \boldsymbol{x}) \overset{\text{sufficiency}}{=} p(\theta \,|\, \overline{x}) \propto \pi(\theta)\, p(\overline{x} \,|\, \theta) = \mathcal{N}(\mu_N, \tau_N^2) \quad (10)$$

with

$$\mu_N = \frac{\frac{N}{\sigma^2}\overline{x} + \frac{1}{\tau_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}} \tag{11}$$

$$\frac{1}{\tau_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau_0^2} \tag{12}$$

see also Example 10.2 in Kay-I.

Comments:

- If N is large, the influence of the prior disappears and the posterior distribution effectively depends only on $\overline{x}$ and σ^2.

- If $\tau_0^2 = \sigma^2$, the prior has the same weight as adding one more observation with value μ_0.

- When $\tau_0^2 \to \infty$ with N fixed or as $N \to \infty$ with τ_0 fixed, we have
$$p(\theta \,|\, \overline{x}) \to \mathcal{N}\big(\overline{x}, \frac{\sigma^2}{N}\big) \tag{13}$$
which is a good general approximation whenever our prior is vague about θ or the number of observations N is large. In this scenario, the influence of the prior disappears. Furthermore,

  $*$ $\tau_0^2 \to \infty$ corresponds to

$$\pi(\theta) \propto 1 \tag{14}$$

and leads to the posterior pdf proportional to the likelihood:

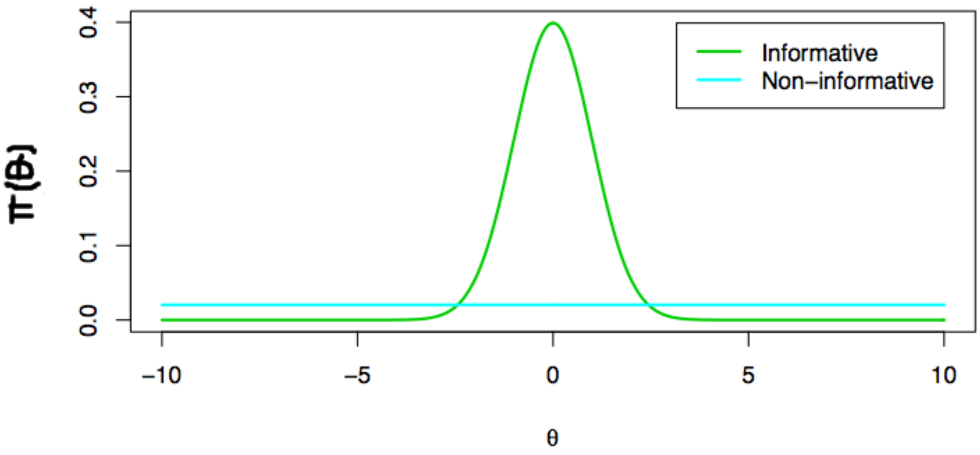
$$p(\theta \,|\, \boldsymbol{x}) = p(\theta \,|\, \overline{x}) \propto \underbrace{p(\overline{x} \,|\, \theta)}_{\text{likelihood}} .$$

The prior choice (14) does not describe a valid probability density since

$$\int_{-\infty}^{\infty} 1 = \infty.$$

Hence, (14) is an *improper prior*. However, we can still use it because the posterior pdf in (13) is proper.

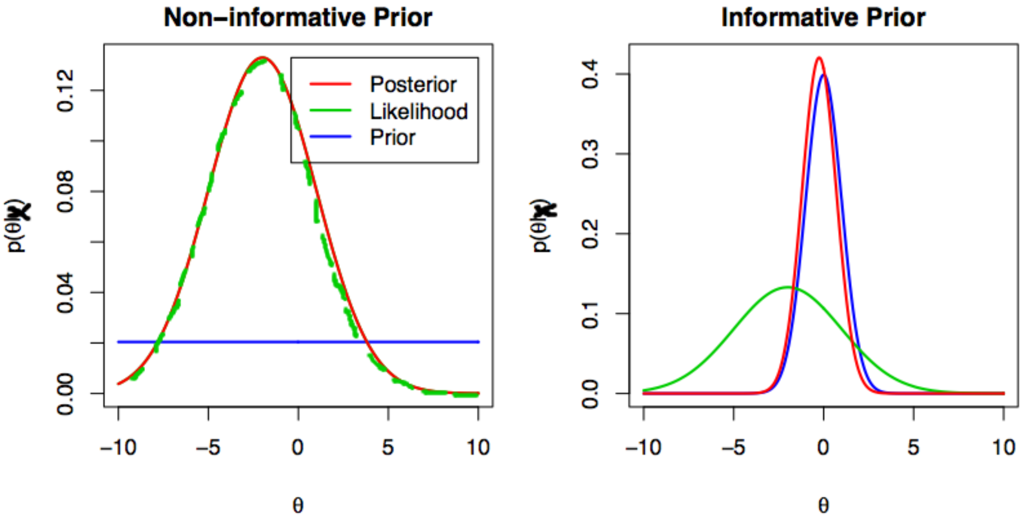
If τ_0^2 is large, we obtain a noninformative prior:



Recall: The posterior mean and precision are

$$\mu_N = \frac{\frac{N}{\sigma^2}\overline{x} + \frac{1}{\tau_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\frac{1}{\tau_N^2} = \frac{N}{\sigma^2} + \frac{1}{\tau_0^2}$$



Sufficiency and Bayesian Models

Since we have just applied sufficiency to simplify our Bayesian calculations, perhaps it is a good idea to formally state and prove the following (Kolmogorov's) result:

Theorem 1. *If a statistic $T(\mathbf{X})$ is sufficient for a parameter θ, then*

$$p(\theta \,|\, T(\boldsymbol{x})) = p(\theta \,|\, \boldsymbol{x}).$$

Proof. (A rough proof) Utilize the factorization theorem:

$$\begin{aligned} p(\theta \,|\, \boldsymbol{x}) &\propto \pi(\theta)\, p(\boldsymbol{x} \,|\, \theta) = \pi(\theta)\, g(T(\boldsymbol{x}), \theta)\, h(\boldsymbol{x}) \\ &\propto \pi(\theta)\, g(T(\boldsymbol{x}), \theta) \\ &\propto p(\theta \,|\, T(\boldsymbol{x})). \end{aligned}$$

□

For *true Bayesians*, the statement

$$p(\theta \,|\, \boldsymbol{x}) = p(\theta \,|\, T(\boldsymbol{x}))$$

is the definition of sufficient statistics $T(\boldsymbol{x})$ for θ. Note that the factorization theorem applies to the posterior $p(\theta \,|\, \boldsymbol{x})$ the same way as it does to the likelihood $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$.

(Back to) DC-level Estimation in AWGN with Known Variance: Predictive Distribution

Suppose that we have collected N (conditionally) i.i.d. observations $x_1, x_2, \ldots, x_N$ (given θ) according to the DC model described earlier. We wish to predict the next observation, denoted by $x_\star$. Recall that

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

is a sufficient statistic for estimating θ based on $x_1, x_2, \ldots, x_N$ and that

$$p(\theta \,|\, x_1, x_2, \ldots, x_N) = p(\theta \,|\, \overline{x}).$$

Assume that $x_\star$ is conditionally independent of $x_1, x_2, \ldots, x_N$ (given θ) and, therefore, along the lines of (8), we get:

$$p_{x_\star \,|\, \overline{x}}(x_\star \,|\, \overline{x}) = \int \underbrace{p_{x_\star \,|\, \theta}(x_\star \,|\, \vartheta) \cdot p_{\theta \,|\, \overline{x}}(\vartheta \,|\, \overline{x})}_{p_{x_\star, \theta \,|\, \overline{x}}(x_\star, \vartheta \,|\, \overline{x})} \, d\vartheta. \tag{15}$$

[The fact that $x_\star$ is conditionally independent of $x_1, x_2, \ldots, x_N$ also implies that

$$p(x_\star \,|\, \theta, \overline{x}) = p(x_\star \,|\, \theta) \tag{16}$$

which is analogous to (4).]

Let us focus on the integrand (and drop the subscripts, for simplicity):

$$p(x_\star, \theta \,|\, \overline{x}) = \underbrace{p(x_\star \,|\, \theta)}_{\mathcal{N}(\theta, \sigma^2)} \cdot \underbrace{p(\theta \,|\, \overline{x})}_{\mathcal{N}(\mu_N, \tau_N^2)}$$

Now

$$p(x_\star, \theta \,|\, \overline{x}) \quad \propto \quad \exp\Big[-\frac{1}{2\,\sigma^2}(x_\star - \theta)^2\Big] \cdot \exp\Big[-\frac{1}{2\,\tau_N^2}(\mu_N - \overline{x})^2\Big]$$

which we recognize as a kernel of a bivariate Gaussian pdf, see p. 27 in handout # 1. Hence, $p(x_\star, \theta \,|\, \overline{x})$ is a bivariate Gaussian pdf. Now, integrating θ out (i.e. marginalizing with respect to θ) in (15) is easy [see p. 26 in handout # 1] — we know that the posterior predictive pdf must be Gaussian and we just need to find its mean

$$\begin{aligned} \mathrm{E}_{x_\star, \theta \,|\, \overline{x}}[X_\star \,|\, \overline{x}] &= \mathrm{E}_{\theta \,|\, \overline{x}}\Big[\qquad \underbrace{\mathrm{E}_{x_\star \,|\, \theta, \overline{x}}[X_\star \,|\, \theta, \overline{x}]}_{\mathrm{E}_{x_\star \,|\, \theta}[X_\star \,|\, \theta] = \theta \quad \text{see also (16)}} \qquad \Big] \\ &= \mathrm{E}_{\theta \,|\, \overline{x}}[\theta] = \mu_N. \end{aligned}$$

and its variance [where we use (16) again]:

$$\begin{aligned}
\mathrm{var}_{x_\star,\theta \,|\, \overline{x}}[X_\star \,|\, \overline{x}] &= \mathrm{E}_{\,\theta \,|\, \overline{x}}[\underbrace{\mathrm{var}_{x_\star \,|\, \theta}(X_\star \,|\, \theta)}_{\sigma^2}] \\
&\quad + \mathrm{var}_{\theta \,|\, \overline{x}}[\underbrace{\mathrm{E}_{\,x_\star \,|\, \theta}(X_\star \,|\, \theta)}_{\theta}] \\
&= \sigma^2 + \tau_N^2
\end{aligned}$$

see the probability review in handout # 1. Therefore

$$p(x_\star \,|\, \overline{x}) = \mathcal{N}(\mu_N, \sigma^2 + \tau_N^2).$$

Proper vs. Improper Priors

A prior $\pi(\theta)$ is called *proper* if it is a valid probability distribution:

$$\pi(\theta) \geq 0 \quad \forall\theta, \qquad \int \pi(\theta)\, d\theta = 1.$$

A prior $\pi(\theta)$ is called *improper* if

$$\pi(\theta) \geq 0 \quad \forall\theta, \qquad \int \pi(\theta)\, d\theta = \infty.$$

If a prior is proper, so is the posterior

$$p(\theta \,|\, \boldsymbol{x}) \propto \pi(\theta)\, p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$$

(and everything is fine).

If a prior is improper, the posterior may or may not be proper. For many common problems, popular improper noninformative priors (e.g. Jeffreys' priors, to be discussed later in this handout) lead to proper posteriors, assuming that there is enough data. But, this has to be checked!

Regarding "propriety," all that we really care about is that the posterior is proper, making it a valid pdf/pmf (which is clearly key to Bayesian inference)!

Example: Estimating the Variance of a Gaussian Distribution with Known Mean

See also Ch. 2.7 in

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Second ed. New York: Chapman & Hall, 2004.

Data model: $p(x \,|\, \sigma^2) = \mathcal{N}(\mu, \sigma^2)$ where $\sigma^2 \geq 0$ is now the parameter of interest and μ is assumed known.

For (conditionally) i.i.d. $x_1, x_2, \ldots, x_N$ (given σ^2), the likelihood function is

$$\begin{aligned} p(\boldsymbol{x} \,|\, \sigma^2) &= (2\pi\sigma^2)^{-N/2} \exp\Big[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\Big] \\ &= (2\pi\sigma^2)^{-N/2} \exp\Big(-\frac{N v}{2\,\sigma^2}\Big) \end{aligned}$$

where

$$v \triangleq \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

is a sufficient statistic. The above likelihood function is in the exponential-family form:

$$p(\boldsymbol{x} \,|\, \sigma^2) = \Big[\prod_{i=1}^{N} h(x_i)\Big] \cdot [g(\sigma^2)]^N \exp[\phi(\sigma^2)\, t(\boldsymbol{x})], \quad \sigma^2 \geq 0$$

where

$$\begin{aligned} g(\sigma^2) &= (\sigma^2)^{-1/2} \\ t(\boldsymbol{x}) &= v \\ \phi(\sigma^2) &= -\frac{N}{2\,\sigma^2} \end{aligned}$$

and hence the conjugate prior is

$$\begin{aligned} \pi(\sigma^2) &\propto g(\sigma^2)^\eta \exp[\phi(\sigma^2)^T \nu] \\ &\propto (\sigma^2)^{-\eta/2} \cdot \exp\Big(-\frac{N}{2\,\sigma^2}\,\nu\Big), \quad \sigma^2 \geq 0. \end{aligned}$$

What does this distribution *look like*? By looking up the table of distributions in Appendix A of

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Second ed. New York: Chapman & Hall, 2004

we see that it "looks like" (and therefore is) an inverse gamma pdf:

$$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2), \quad \sigma^2 \geq 0$$

where α and β are assumed known (and are sometimes referred to as the *hyperparameters*). (Note that this distribution is used as a prior distribution for the variance parameter in Example 10.3 of Kay-I.) For ease of interpretation, we reparametrize this prior pdf as a *scaled inverted χ^2 distribution* with scale

σ_0^2 and ν_0 degrees of freedom; here σ_0^2 and ν_0 are known hyperparameters. In other words, we take the prior distribution of σ^2 to be the distribution of

$$\frac{\sigma_0^2\,\nu_0}{X}$$

where X is a $\chi^2_{\nu_0}$ random variable (see the underlined part of the distribution table handout). We use the following notation for this distribution

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \;\propto\; \left(\sigma^2\right)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0\,\sigma_0^2}{2\,\sigma^2}\right), \;\sigma^2 \geq 0.$$

Note: From the table of distributions, we also obtain the following facts:

- the prior mean is $\sigma_0^2\nu_0/(\nu_0 - 2)$ and

- when ν_0 is large, the prior variance behaves like σ_0^4/ν_0 (implying that large ν_0 yields high prior precision).

Finally,

$$\begin{aligned} p(\sigma^2 \,|\, \boldsymbol{x}) &\propto \pi(\sigma^2)\, p(\boldsymbol{x}|\sigma^2) \\ &\propto (\sigma^2)^{-(\frac{\nu_0}{2}+1)} \exp\Big(- \frac{\nu_0\, \sigma_0^2}{2\, \sigma^2}\Big) \\ &\quad \cdot (\sigma^2)^{-N/2} \cdot \exp\Big(- \frac{N v}{2\, \sigma^2}\Big) \\ &\propto (\sigma^2)^{-(\frac{\nu_N}{2}+1)} \cdot \exp\Big(- \frac{\nu_N\, \sigma_N^2}{2\, \sigma^2}\Big) \end{aligned}$$

with

$$\nu_N = \nu_0 + N$$

and

$$\sigma_N^2 = \frac{N v + \nu_0\, \sigma_0^2}{N + \nu_0}.$$

Therefore, $p(\sigma^2 \,|\, \boldsymbol{x})$ is also a scaled inverted χ^2 distribution. Now, the posterior mean (MMSE estimate, to be shown later) is

$$\mathrm{E}\,[\sigma^2|\boldsymbol{x}] = \frac{\sigma_N^2 \nu_N}{\nu_N - 2} = \frac{N v + \nu_0\, \sigma_0^2}{N + \nu_0 - 2}.$$

Comments:

- The MMSE estimate of σ^2 is a weighted average of the prior “guess” and a data estimate:

$$\mathrm{E}\,[\sigma^2 \,|\, \boldsymbol{x}] = \frac{N v + \nu_0\, \sigma_0^2}{N + \nu_0 - 2}$$

where the weights are obtained using the prior and sample degrees of freedom.

- **Interpretation of the prior information:** the chosen prior provides information equivalent to ν_0 observations with average variance equal to σ_0^2.

- As $N \to \infty$, $\sigma_N^2 \to v$ and $\mathrm{E}\,[\sigma^2 \,|\, \boldsymbol{x}] \to v$.

- Similarly, as $\nu_0 \to \infty$, $\sigma_N^2 \to \sigma_0^2$ and $\mathrm{E}\,[\sigma^2 \,|\, \boldsymbol{x}] \to \sigma_0^2$.

Noninformative Priors

Although it may seem that picking a noninformative prior distribution is easy, (e.g. just use a uniform), it is not quite that straightforward.

Example. Estimating the Variance of a Gaussian Distribution with Known Mean:

$$\begin{aligned} x_1, x_2, \ldots, x_N \,|\, \sigma \quad & \text{i.i.d.} \quad \mathcal{N}(0, \sigma^2) \\ \pi(\sigma) \quad & \propto \quad i_{[0,\infty)}(\sigma). \end{aligned}$$

We assume a uniform prior (from zero to infinity) for σ.

Question: What is the equivalent prior on σ^2?

Reminder: Let θ be a random variable with density $p(\theta)$ and let $\phi = h(\theta)$ be a one-to-one transformation. Then the density of ϕ satisfies

$$p_\phi(\phi) = p_\theta(\theta) \cdot |d\theta / d\phi| = p(\theta) \cdot |h'(\theta)|^{-1}.$$

We now apply the above change-of-variables formula to our problem: $h(\sigma) = \sigma^2$, $h'(\sigma) = 2\,\sigma$, yielding

$$\pi(\sigma^2) \propto \frac{1}{2\,\sigma}, \quad \sigma > 0$$

which is clearly *not uniform*. This implies that our prior belief is that the variance σ^2 is small.

Then, for a uniform prior on the variance σ^2, the equivalent prior on σ is

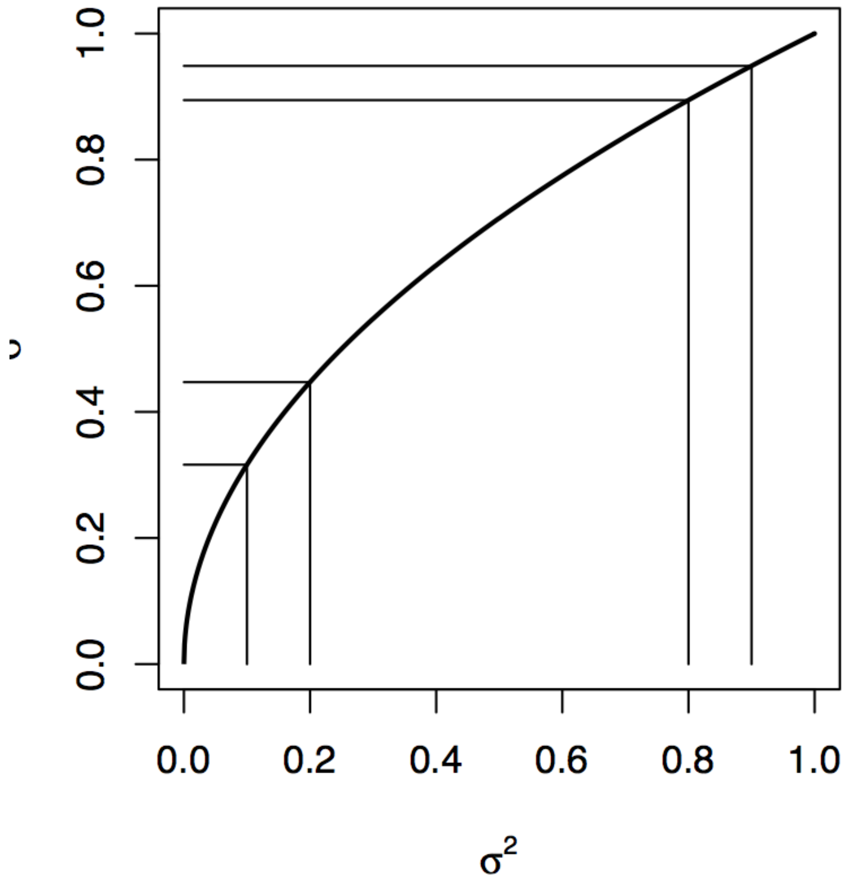
$$\pi(\sigma) \propto 2\,\sigma, \quad \sigma > 0$$

implying that we believe that the standard deviation σ is large.

(Problems of this kind are the main reasons why R. A. Fisher, the "father" of statistics, had a distaste for the Bayesian approach.)

One way to visualize what is happening is to look at what happens to intervals of equal measure.

In the case of σ^2 being uniform, an interval $[a, a + 0.1]$ must have the same prior measure as the interval $[0.1, 0.2]$. When we transform to σ, the corresponding prior measure must have intervals $[\sqrt{a}, \sqrt{a + 0.1}]$ having equal measure. But, the length of $[\sqrt{a}, \sqrt{a + 0.1}]$ is a decreasing function of a, which agrees with the increasing density in σ.



Therefore, when talking about non-informative priors, we need to think about scale.

Jeffreys' Priors

Can we pick a prior where the scale of the parameter does not matter?

Jeffreys' general principle states that any rule for determining the prior density $\pi_\theta(\theta)$ for parameter θ should yield an equivalent result if applied to the transformed parameter ($\phi = h(\theta)$, say). Therefore, applying

$$\pi_\phi(\phi) = \{\pi_\theta(\theta) \cdot |d\theta/d\phi|\}\big|_{\theta=h^{-1}(\phi)} = \{\pi_\theta(\theta) \cdot |h'(\theta)|^{-1}\}\big|_{\theta=h^{-1}(\phi)}$$

should give the same answer as dealing directly with the transformed model,

$$p(\boldsymbol{x}, \phi) = \pi_\phi(\phi)\, p(\boldsymbol{x} \,|\, \phi).$$

Jeffreys' suggestion:

$$\pi_\theta(\theta) \propto \sqrt{\mathcal{I}(\theta)}$$

where $\mathcal{I}(\theta)$ is the Fisher information. Why is this choice good?

We now compute $\mathcal{I}(\phi)$ for $\phi = h(\theta)$ and $h(\cdot)$ one-to-one:

$$\begin{aligned}
\mathcal{I}(\phi) &= \mathrm{E}_{X\,|\,\phi}\Big[\Big|\frac{d\log p(\boldsymbol{x}\,|\,\phi)}{d\phi}\Big|^2 \,\Big|\, \phi\Big] \\
&= \mathrm{E}_{X\,|\,\theta}\Big[\Big|\frac{d\log p(\boldsymbol{x}\,|\,\theta)}{d\theta}\Big|^2 \cdot \Big|\frac{d\theta}{d\phi}\Big|^2 \,\Big|\, \theta\Big]\Big|_{\theta=h^{-1}(\phi)} \\
&= \mathcal{I}(\theta)\,\Big|\frac{d\theta}{d\phi}\Big|^2\Big|_{\theta=h^{-1}(\phi)}
\end{aligned}$$

implying

$$\underbrace{\sqrt{\mathcal{I}(\phi)}}_{\text{computing Jeffrey's prior directly for } \phi} = \underbrace{\sqrt{\mathcal{I}(\theta)} \cdot \Big|\frac{d\theta}{d\phi}\Big|\Big|_{\theta=h^{-1}(\phi)}}_{\text{applying the usual Jacobian transformation}}$$

Reminder: the "usual" Jacobian transformation applied to the prior pdf/pmf is

$$\pi_\phi(\phi) = \{\pi_\theta(\theta) \cdot |d\theta/d\phi|\}\big|_{\theta=h^{-1}(\phi)}.$$

Example. Estimating the Variance of a Gaussian Distribution with Known Mean: Recall that the Fisher information for σ^2 is [see eq. (12) in handout # 2]

$$\mathcal{I}(\sigma^2) = \frac{N}{2\,\sigma^4}.$$

Therefore, the Jeffreys' prior for σ^2 is

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}$$

for $\sigma^2 > 0$. Alternative descriptions under different parameterizations for the variance parameter are (for $\sigma^2 > 0$)

$$\pi(\sigma) \propto \frac{1}{\sigma}, \quad \pi(\log \sigma^2) \propto 1 \quad \text{(uniform)}.$$

Example. Estimating the Mean of a Gaussian Distribution with Known Variance: $p(x_i \,|\, \theta) = \mathcal{N}(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$, where σ^2 is assumed known. Here

$$\mathcal{I}(\theta) = \frac{N}{\sigma^2} = \text{const}$$

and, therefore, the (clearly improper) Jeffreys' prior for θ is

$$\pi(\theta) \propto 1 \quad \text{(uniform)}.$$

Multiparameter Models

So far, we have discussed Bayesian estimation for toy scenarios with single parameters. In most real applications, we have multiple parameters that need to be estimated.

- Classical DC-signal-in-AWGN-noise model: given μ and σ^2, x_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$; here, both μ and σ^2 are *unknown*.

There are many more general examples: any signal-plus-noise model where we do not know a signal parameter (μ in the above example) and a noise parameter (σ^2 in the above example).

Note: The noise variance σ^2 is often considered a *nuisance parameter*. We are not interested in its value, but we need to take care of it because it is not known (and hence is a nuisance).

Consider the case with two parameters θ_1 and θ_2 and assume that only θ_1 is of interest. (Here, θ_1 and θ_2 could be vectors, but we describe the scalar case for simplicity.) An example of this would be the DC-signal-in-AWGN-noise model, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

We wish to base our inference on $p(\theta_1 \,|\, \boldsymbol{x})$, the marginal posterior pdf/pmf which accounts for the uncertainty due to the fact that θ_2 is unknown. First, we start with the joint posterior pdf/pmf:

$$p(\theta_1, \theta_2 \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta_1, \theta_2)\, \pi(\theta_1, \theta_2)$$

and then, in the continuous (pdf) case, *integrate out* the nuisance parameter (also discussed in Ch. 10.7 in Kay-I):

$$p(\theta_1 \,|\, \boldsymbol{x}) = \int p(\theta_1, \theta_2 \,|\, \boldsymbol{x})\, d\theta_2$$

or, equivalently,

$$p(\theta_1 \,|\, \boldsymbol{x}) = \int p(\theta_1 \,|\, \theta_2, \boldsymbol{x})\, p(\theta_2 \,|\, \boldsymbol{x})\, d\theta_2$$

which implies that the marginal posterior distribution of θ_1 can be viewed as its conditional posterior distribution (conditioned on the nuisance parameter, in addition to the data) *averaged over* the marginal posterior pdf/pmf of the nuisance parameter. Hence, the uncertainty due to the unknown θ_2 is taken into account.

Posterior Predictive Distribution: Suppose that we wish to predict data $x_\star$ coming from the model $p(\boldsymbol{x} \,|\, \theta_1, \theta_2)$. If $\boldsymbol{x}_\star$ and $\boldsymbol{x}$ are conditionally independent given θ_1 and θ_2, then, following analogous arguments as those used to derive (8), we obtain:

$$\begin{aligned}
p(\boldsymbol{x}_\star \,|\, \boldsymbol{x}) &= \int p(\boldsymbol{x}_\star \,|\, \theta_1, \theta_2) \cdot p(\theta_1, \theta_2 \,|\, \boldsymbol{x}) \, d\theta_1 \, d\theta_2 \\
&= \int p(\boldsymbol{x}_\star \,|\, \theta_1, \theta_2) \cdot p(\theta_1 \,|\, \theta_2, \boldsymbol{x}) \cdot p(\theta_2 \,|\, \boldsymbol{x}) \, d\theta_1 \, d\theta_2.
\end{aligned}$$

The above integrals are often difficult to evaluate analytically (and may be multidimensional if θ_2 and θ_1 are vectors). We typically need Monte Carlo methods to handle practical cases. An EM algorithm may suffice if we just need to find the mode of $p(\theta_1 \,|\, \boldsymbol{x})$. These approaches will be discussed in detail later.

The lack of analytical tractability is the reason why the Bayesian methodology had been considered obscure or impractical in the past. Sometimes, Bayesians made analytically tractable but meaningless (or hard to justify) constructs, which had made the Bayesian approach appear even more obscure.

But, analytical tractability is a double-edged sword. The advent of computers and development of Monte Carlo methods seem to have changed the balance in favor of the Bayesian approach, which has suddenly become practical and much more flexible than classical inference that still largely relies on analytical tractability or hard-to-justify asymptotic results.

We now consider one of the rare (and therefore classical) practical cases where analytical Bayesian computations are possible.

Example: DC-level Estimation in AWGN Noise with Unknown Variance

See also Ch. 3.2 in

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, Second ed. New York: Chapman & Hall, 2004

and Example 11.4 in Kay-I.

- We assume that μ and σ^2 are *a priori* independent and use the standard non-informative Jeffreys' priors for each:

$$\pi(\mu, \sigma^2) = \pi(\mu) \cdot \pi(\sigma^2) \propto 1 \cdot \frac{1}{\sigma^2} = \frac{1}{\sigma^2}$$

for $\sigma^2 > 0$. Recall that our data model corresponds to a DC level in additive white Gaussian noise, i.e. $x_i \,|\, \mu, \sigma^2$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ for $i = 1, 2, \ldots, N$.

The product of the above prior pdf and the likelihood is proportional to the posterior pdf:

$$p(\mu, \sigma^2 \,|\, \boldsymbol{x}) \propto \frac{1}{\sigma^2} \cdot (\sigma^2)^{-N/2} \cdot \exp\left\{ -\frac{1}{2\sigma^2}[N s^2 + N(\overline{x} - \mu)^2] \right\}$$

where

$$s^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \overline{x})^2, \quad \overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

which are the sufficient statistics. The above posterior pdf is proper provided that $N \geq 2$.

Conditional Posterior Pdf of μ

The conditional posterior density of μ given σ^2 is proportional to the joint posterior density with σ^2 held constant:

$$\mu \,|\, \sigma^2, \boldsymbol{x} \sim \mathcal{N}(\overline{x}, \frac{\sigma^2}{N})$$

which agrees (as it must) with (13) — the case of estimating the DC level in AWGN noise with known variance.

Marginal Posterior Pdf of σ^2

The marginal posterior density of σ^2 is a scaled inverted χ^2:

$$\sigma^2 \,|\, \boldsymbol{x} \sim \text{Inv-}\chi^2(N-1, s^2).$$

We now derive this result:

$$p(\sigma^2 \,|\, \boldsymbol{x}) = \frac{p(\mu, \sigma^2 \,|\, \boldsymbol{x})}{\underbrace{p(\mu \,|\, \sigma^2, \boldsymbol{x})}_{\mathcal{N}\left(\overline{x}, \frac{\sigma^2}{N}\right)}}$$

$$\underbrace{\propto}_{\substack{\text{keep track of the terms} \\ \text{containing } \mu \text{ and } \sigma^2}}$$

$$\frac{(\sigma^2)^{-N/2-1} \cdot \exp\left\{ -\frac{1}{2\sigma^2}[Ns^2 + N(\overline{x} - \mu)^2]\right\}}{(\sigma^2)^{-1/2} \cdot \exp[-\frac{(\mu - \overline{x})^2}{2\sigma^2/N}]}$$

$$= \text{Inv-}\chi^2(N - 1, \tfrac{N}{N-1}\, s^2)$$

$$= \text{Inv-}\chi^2\Big(N - 1, \tfrac{1}{N-1} \textstyle\sum_{i=1}^{N} (x_i - \overline{x})^2\Big)$$

i.e. the following holds for $\sigma^2 \,|\, \boldsymbol{x}$:

$$\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{\underbrace{\sigma^2}_{\text{coming from } p(\sigma^2 \,|\, \boldsymbol{x})}} \sim \chi^2_{N-1}.$$

Therefore, we have managed to compute

$$p(\sigma^2 \,|\, \boldsymbol{x}) = \int_{-\infty}^{\infty} p(\mu, \sigma^2 \,|\, \boldsymbol{x})\, d\mu$$

algebraically rather than by performing the actual integration.

Note: This is a Bayesian trick for "integrating" μ out without actually performing the integration! The key to this trick is that $p(\mu \,|\, \sigma^2, \boldsymbol{x})$ is known exactly (e.g. it belongs to a family of pdfs/pmfs that appear in our table of distributions).

Marginal Posterior Pdf of μ

$$p(\mu \,|\, \boldsymbol{x}) = \int_0^\infty p(\mu, \sigma^2 \,|\, \boldsymbol{x}) \, d\sigma^2$$

$$\underbrace{\propto}_{\substack{\text{keep track of the terms} \\ \text{containing } \mu \text{ and } \sigma^2}} \int_0^\infty \left((\sigma^2)^{-N/2-1} \exp\left\{ -\frac{1}{2\sigma^2} [N s^2 + N(\overline{x} - \mu)^2] \right\} \right) d\sigma^2.$$

Make a substitution:

$$z = \frac{A(\mu)}{2\sigma^2}$$

where $A(\mu) = N s^2 + N(\overline{x} - \mu)^2$. Then

$$\frac{d\sigma^2}{dz} = -\frac{A(\mu)}{2z^2}$$

and

$$p(\mu \,|\, \boldsymbol{x}) \propto \int_0^\infty \Big(\frac{z}{A(\mu)}\Big)^{N/2+1} \cdot \frac{A(\mu)}{z^2} \exp(-z)\, dz$$

$$p(\mu \,|\, \boldsymbol{x}) \propto A(\mu)^{-N/2} \cdot \underbrace{\int_0^\infty z^{N/2-1} \exp(-z)\, dz}_{\text{unnormalized gamma integral}}$$

$$\propto [Ns^2 + N(\overline{x} - \mu)^2]^{-N/2} \propto \Big[1 + \frac{1}{N-1} \cdot \frac{(\mu - \overline{x})^2}{s^2/(N-1)}\Big]^{-N/2}$$

$$= t_{N-1}\Big(\mu \,\Big|\, \overline{x}, \frac{s^2}{N-1}\Big)$$

i.e. this is the pdf of the *(scaled) t distribution* with $N-1$ degrees of freedom and parameters

$$\overline{x} \quad \text{(mean)} \quad \text{and} \quad \frac{s^2}{N-1} = \frac{\sum_{i=1}^N (x_i - \overline{x})^2}{N(N-1)} \quad \text{(scale)}.$$

Equivalently,

$$\mu \,|\, \boldsymbol{x} \overset{\mathrm{d}}{\equiv} \overline{x} + \underbrace{T}_{\substack{\text{standard } t \text{ RV with} \\ N-1 \text{ degrees of freedom}}} \cdot \sqrt{\frac{\sum_{i=1}^N (x_i - \overline{x})^2}{N(N-1)}}.$$

For HW, apply our Bayesian trick here and compute $p(\mu \,|\, \boldsymbol{x})$ without performing the integral $\int_0^\infty p(\mu, \sigma^2 \,|\, \boldsymbol{x})\, d\sigma^2$.

How do We Summarize the Obtained Posterior Distributions?

We can compute moments: means, variances (covariance matrices) of the posterior distributions, or perhaps its mode, median etc.

We can try to make “interval inferences” based on the posterior distributions $\Longrightarrow$ *credible sets*, also known as *Bayesian confidence intervals*.

Let A be some subset of the parameter space for θ. Then, A is a $100\,c\,\%$ credible set for θ if

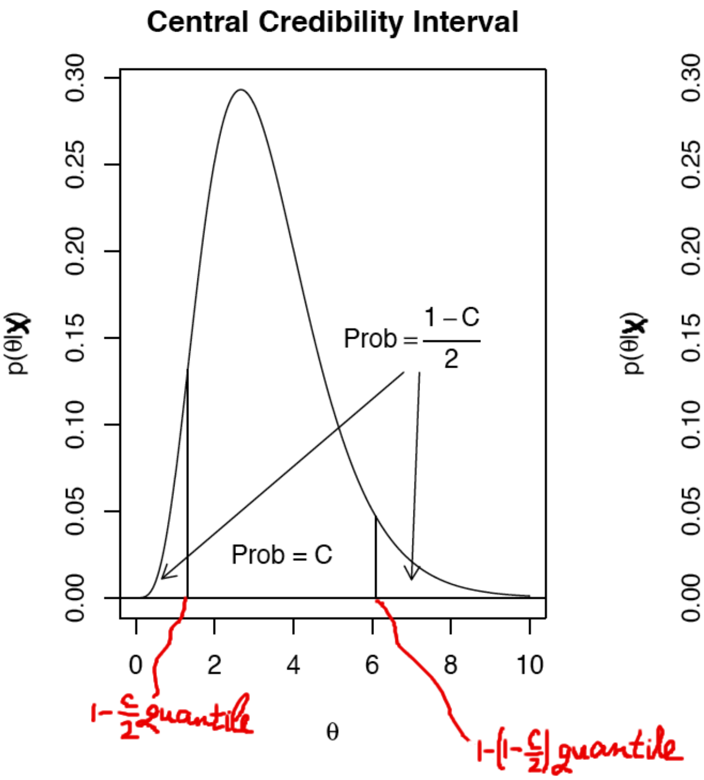
$$P\{\theta \in A \,|\, \boldsymbol{x}\} = c.$$

The most common approach to credible intervals (i.e. scalar credible sets) are *central credible intervals*. A central credible interval $[\tau_{\mathrm{l}}, \tau_{\mathrm{u}}]$ satisfies

$$\frac{1-c}{2} = \int_{-\infty}^{\tau_l} p(\theta \,|\, \boldsymbol{x})\, d\theta, \quad \frac{1-c}{2} = \int_{\tau_u}^{\infty} p(\theta \,|\, \boldsymbol{x})\, d\theta$$

where τ_{l} and τ_{u} are $\frac{1-c}{2}$ and $1 - \frac{1-c}{2}$ quantiles of the posterior pdf, see also the figure in the example to follow. An alternative is the $100\,c\,\%$ *highest posterior density* (HPD) region, which is defined as the smallest region of the parameter space with probability c.

Example:



Central interval: (1.313, 6.102); length = 4.789.

HPD interval: (1.006, 5.571); length = 4.565.

The central interval is usually easier to determine as it involves only finding quantiles of the posterior distribution.

Example: Go back to the example on pp. 39–43. A 95% (say)

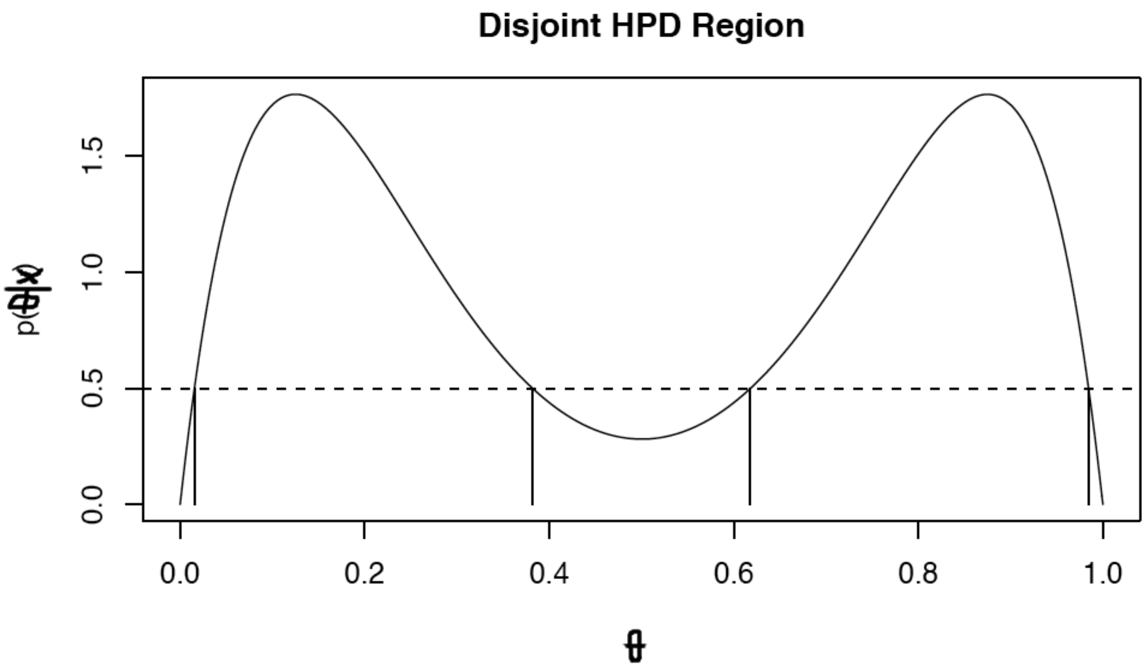
HPD credible region for μ based on

$$p(\mu \,|\, \boldsymbol{x}) = t_{N-1}\Big(\mu \,\Big|\, \overline{x}, \frac{s^2}{N-1}\Big)$$

coincides with the common 95% confidence interval based on the classical t test, taught in STAT 101. However, the Bayesian interpretation is very different.

There are a couple of problems with the HPD approach to constructing credible sets:

- the HPD approach may yield multiple disconnected regions (if the posterior distribution is not unimodal), e.g.



- the HPD approach is *not* invariant to parameter transformations. A related point is that what may look like a “flat”/noninformative distribution for model parameters under one parametrization may look “non-flat” under another parametrization. We have seen such an example earlier in this handout, but let us give another example with focus on HPD credible-region computation.

Suppose that $0 < \theta < 1$ is a parameter of interest, but that we are equally interested in $\gamma = 1/(1-\theta)$. Now, if a posterior density $p(\theta \,|\, \boldsymbol{x})$ is, say,

$$p(\theta \,|\, \boldsymbol{x}) = \begin{cases} 2\,\theta, & 0 < \theta < 1 \\ 0, & \text{otherwise} \end{cases}$$

the corresponding cdf is

$$P(\theta \,|\, \boldsymbol{x}) = \begin{cases} 0, & \theta < 0 \\ \theta^2, & 0 < \theta < 1 \\ 1, & \theta > 1 \end{cases}$$

implying that, for example, a 95% HPD credible set for θ is $(\sqrt{0.05}, 1)$. Now, despite the fact that $\gamma = 1/(1-\theta)$ is a monotone function of θ, the interval

$$\Big(\frac{1}{1-\sqrt{0.05}}, \infty\Big) \qquad (17)$$

is not an HPD credible set for γ. Clearly, (17) is a 95% credible set for γ, but it is not HPD. To see this, we find the cdf $P(\gamma \,|\, \boldsymbol{x})$: for $t \geq 1$,

$$\begin{aligned} P(t \,|\, \boldsymbol{x}) &= P[\gamma \leq t] = P\Big[\frac{1}{1-\theta} \leq t\Big] \\ &= P\Big[\theta \leq 1 - \frac{1}{t}\Big] = \Big(1 - \frac{1}{t}\Big)^2. \end{aligned}$$

Therefore, $\gamma \,|\, \boldsymbol{x}$ has the pdf

$$p(\gamma \,|\, \boldsymbol{x}) = \begin{cases} 2\left(1 - \frac{1}{\gamma}\right)\frac{1}{\gamma^2}, & \text{for } \gamma \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

and HPD intervals for $\gamma \,|\, \boldsymbol{x}$ must be two-sided, which is in contrast with (17).

Bayesian MMSE Estimation

Suppose that we need to provide a point estimate of the parameter of interest. How do we do that in the Bayesian setting? Here, we first consider the most popular *squared-error loss scenario* and later, we will discuss the general scenario (for an arbitrary loss). For simplicity, we focus on the scalar parameter case.

If we are *true Bayesians*, we should construct our estimator $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x})$ based on the posterior distribution $p(\theta \,|\, \boldsymbol{x})$. Hence, a *truly Bayesian approach* to solving the above problem is, say, to obtain $\widehat{\theta}$ by minimizing the *posterior expected squared loss*:

$$\rho(\widehat{\theta} \,|\, \boldsymbol{x}) = \int \underbrace{(\widehat{\theta} - \theta)^2}_{\text{squared-error loss}} p(\theta \,|\, \boldsymbol{x}) \, d\theta$$

with respect to $\widehat{\theta}$. This is easy to do: decompose $\rho(\widehat{\theta} \,|\, \boldsymbol{x})$ as

$$\begin{aligned}
\rho(\widehat{\theta} \,|\, \boldsymbol{x}) &= \int \Big(\widehat{\theta} - \mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}] + \mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}] - \theta\Big)^2 p(\theta \,|\, \boldsymbol{x}) \, d\theta \\
&= \underbrace{\int \Big(\widehat{\theta} - \mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}]\Big)^2 p(\theta \,|\, \boldsymbol{x}) \, d\theta}_{(\widehat{\theta} - \mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}])^2} \\
&\quad + \int \Big(\theta - \mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}]\Big)^2 p(\theta \,|\, \boldsymbol{x}) \, d\theta
\end{aligned}$$

and the optimal $\widehat{\theta}$ follows:

$$\mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\theta \,|\, \boldsymbol{x}] = \arg\min_{\widehat{\theta}} \rho(\widehat{\theta} \,|\, \boldsymbol{x}).$$

Hence, the posterior mean of the parameter θ minimizes its posterior expected squared loss.

Mean-square error measures.

1. **Classical:** $\mathrm{MSE}(\widehat{\theta}) = \int (\widehat{\theta} - \theta)^2 p(x; \theta) dx$.

2. **"Bayesian" MSE (Preposterior, not truly Bayesian):** $\mathrm{BMSE}(\widehat{\theta}) = \int \int (\widehat{\theta} - \theta)^2 p(x|\theta)\pi(\theta) dx d\theta = \mathrm{E}_{\theta}[\mathrm{MSE}(\widehat{\theta})]$.

The preposterior MSE (BMSE) is obtained by averaging the squared-error loss over both the noise *and* parameter realizations. It is computable before the data has been collected, hence the name preposterior.

- The classical MSE generally depends on true θ. Therefore, classical MMSE "estimates" usually depend on θ $\Longrightarrow$ classical MMSE estimates do not exist.

- The preposterior BMSE *does not* depend on θ $\Longrightarrow$ Bayesian MMSE estimates exist.

Which $\widehat{\theta}$ minimizes BMSE? Since

$$\mathrm{BMSE}(\widehat{\theta}) = \mathrm{E}_{x,\theta}[(\widehat{\theta} - \theta)^2] = \mathrm{E}_{\boldsymbol{x}}\Big\{ \underbrace{\mathrm{E}_{\theta|\boldsymbol{x}}[(\theta - \widehat{\theta})^2 | \boldsymbol{x}]}_{\rho(\widehat{\theta}\,|\,\boldsymbol{x})} \Big\}$$

and, for every given $\boldsymbol{x}$, we know that $\widehat{\theta} = \mathrm{E}_{\theta\,|\,\boldsymbol{x}}[\theta\,|\,\boldsymbol{x}]$ minimizes $\rho(\widehat{\theta}\,|\,\boldsymbol{x})$; then, clearly, the MMSE estimator is the *posterior mean of θ*:

$$\boxed{\widehat{\theta} = \mathrm{E}_{\theta\,|\,\boldsymbol{x}}[\theta\,|\,\boldsymbol{x}]}\,.$$

Linear MMSE Estimation: Gaussian Vector Case (Theorem 10.2 in Kay-I)

The MMSE estimate for a parameter vector:

$$\widehat{\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}}[\boldsymbol{\theta} \,|\, \boldsymbol{x}].$$

The Bayesian MMSE estimator is tractable if $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are jointly Gaussian.

Theorem 2. *Assume that*

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\theta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \boldsymbol{\mu}_{\theta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{x}} & \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{\theta}} \\ \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{x}} & \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{bmatrix} \right).$$

Then, the posterior pdf $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ *is also Gaussian, with the first two moments given by*

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{\theta} | \boldsymbol{X} = \boldsymbol{x}\right] &= \boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{x}} \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}) \\ \boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} &= \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{x}} \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{\theta}}. \end{aligned}$$

Proof. See Appendix 10 A in Kay-I and p. 26 in handout # 1.

Gaussian Linear Model (Theorem 10.3 in Kay-I)

Theorem 3. *Assume that*

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{w}$$

where $\boldsymbol{H}$ is a known matrix and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{w}})$, independent of the parameter $\boldsymbol{\theta}$. Here, we assume that $\boldsymbol{C}_{\boldsymbol{w}}$ is known. If the prior pdf for $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{C}_{\boldsymbol{\theta}})$ [where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{C}_{\boldsymbol{\theta}}$ are known], then the posterior pdf $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is also Gaussian, with

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{\theta} | \boldsymbol{X} = \boldsymbol{x}\right] &= \boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T + \boldsymbol{C}_{\boldsymbol{w}})^{-1}(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\mu}_{\boldsymbol{\theta}}) \\ \boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} &= \boldsymbol{C}_{\boldsymbol{\theta}} - \boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T + \boldsymbol{C}_{\boldsymbol{w}})^{-1}\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}. \end{aligned}$$

Proof. Insert

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{x}\boldsymbol{x}} &= \boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T + \boldsymbol{C}_{\boldsymbol{w}} \\ \boldsymbol{C}_{\boldsymbol{\theta}\boldsymbol{x}} &= \boldsymbol{C}_{\boldsymbol{\theta}}\boldsymbol{H}^T \end{aligned}$$

into Theorem 2. □

Recall the matrix inversion lemma:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

and use it as follows:

$$
\begin{aligned}
(A+BCD)^{-1}BC &= A^{-1}BC - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}BC \\
&= A^{-1}B(C^{-1}+DA^{-1}B)^{-1}(C^{-1}+DA^{-1}B)C \\
&\quad -A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}BC \\
&= A^{-1}B(C^{-1}+DA^{-1}B)^{-1}
\end{aligned}
$$

implying $(\boldsymbol{C_w} + \boldsymbol{H}\boldsymbol{C_\theta}\boldsymbol{H}^T)^{-1}\,\boldsymbol{H}\boldsymbol{C}_{\boldsymbol{\theta}}^T = \boldsymbol{C}_{\boldsymbol{w}}^{-1}\,\boldsymbol{H}\,(\boldsymbol{C}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}$ or, equivalently,

$$\boldsymbol{C_\theta}\boldsymbol{H}^T(\boldsymbol{C_w}+\boldsymbol{H}\boldsymbol{C_\theta}\boldsymbol{H}^T)^{-1} = (\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}+\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}.$$

We can use this formula and the matrix inversion lemma to rewrite the results of Theorem 3:

$$
\begin{aligned}
\mathrm{E}\,[\boldsymbol{\theta}|\boldsymbol{X}=\boldsymbol{x}] &= \boldsymbol{\mu_\theta} \\
&\quad +(\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}+\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}(\boldsymbol{x}-\boldsymbol{H}\boldsymbol{\mu_\theta}) \\
&= (\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}+\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{x} \\
&\quad +(\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}+\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H})^{-1}\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu_\theta} \\
&= (\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}+\boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}\,(\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{x}+\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu_\theta}) \\
\boldsymbol{C}_{\boldsymbol{\theta}|\boldsymbol{x}} &= (\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}+\boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}.
\end{aligned}
$$

These expressions are computationally simpler than those in Theorem 3.

Gaussian Linear Model (cont.)

Let us derive the posterior pdf $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ in Theorem 3 using our "$\propto$ approach":

$$\begin{aligned}
p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) &\propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) \\
&\propto \exp[-\tfrac{1}{2}\,(\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta})^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} (\boldsymbol{x} - \boldsymbol{H}\boldsymbol{\theta})] \\
&\quad \cdot \exp[-\tfrac{1}{2}\,(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})] \\
&\propto \exp(-\tfrac{1}{2}\,\boldsymbol{\theta}^T \boldsymbol{H}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H}\boldsymbol{\theta} + \boldsymbol{x}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H}\boldsymbol{\theta}) \\
&\quad \cdot \exp(-\tfrac{1}{2}\,\boldsymbol{\theta}^T \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta} + \boldsymbol{\mu}_{\boldsymbol{\theta}}^T \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}) \\
&= \exp[-\tfrac{1}{2}\,\boldsymbol{\theta}^T (\boldsymbol{H}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})\, \boldsymbol{\theta} \\
&\quad + (\boldsymbol{x}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H} + \boldsymbol{\mu}_{\boldsymbol{\theta}}^T \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})\, \boldsymbol{\theta}] \\
&\propto \mathcal{N}\Big((\boldsymbol{H}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}\, (\boldsymbol{H}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{x} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\theta}}), \\
&\quad (\boldsymbol{H}^T \boldsymbol{C}_{\boldsymbol{w}}^{-1} \boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}\Big).
\end{aligned}$$

Remarkably simple!

Comments:

- DC-level estimation in AWGN with known variance introduced on p. 11 is a special case of this result, see also Example 10.2 in Kay-I and the discussion on p. 58.

- Let us examine the posterior mean:

$$\begin{aligned} \mathrm{E}\,[\boldsymbol{\theta}\,|\,\boldsymbol{x}] &= (\underbrace{\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H}}_{\text{likelihood precision}} + \underbrace{\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}}_{\text{prior precision}})^{-1} \\ &\quad \cdot \underbrace{(\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{x}}_{\text{data-dependent term}} + \underbrace{\boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}}_{\text{prior-dependent term}}). \end{aligned}$$

- **Noninformative (flat) prior on θ and white noise.** Consider the Jeffreys' noninformative (flat) prior on $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) \propto 1 \quad \Longleftrightarrow \quad \boldsymbol{C}_{\boldsymbol{\theta}}^{-1} = \boldsymbol{0}$$

  and white noise:

$$\boldsymbol{C}_{\boldsymbol{w}} = \sigma_{\boldsymbol{w}}^2 \cdot \boldsymbol{I}.$$

  Then, $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ simplifies to

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \mathcal{N}\Big(\overbrace{(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}}^{\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}},\; \sigma_{\boldsymbol{w}}^2\,(\boldsymbol{H}^T\boldsymbol{H})^{-1} \Big)$$

$\Longrightarrow$ (central, typically) credible sets constructed based on this posterior pdf are *exactly the same* as classical confidence regions for linear least squares, which are based on

$$(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{x}\,|\,\boldsymbol{\theta} \sim p(\boldsymbol{\theta}, \sigma^2_{\boldsymbol{w}}\,(\boldsymbol{H}^T\boldsymbol{H})^{-1}).$$

(The classical confidence intervals will be discussed later, in the detection part.)

- **Prediction:** Now, let us practice prediction for this model. Say we wish to predict a $x_\star$ coming from the following model:

$$x_\star = \boldsymbol{h}_\star^T\boldsymbol{\theta} + w_\star.$$

Suppose that $w_\star \sim \mathcal{N}(0, \sigma^2)$ independent from $\boldsymbol{w}$, implying that $p(x_\star\,|\,\boldsymbol{\theta}, \boldsymbol{x}) = p(x_\star\,|\,\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{h}_\star^T\boldsymbol{\theta}, \sigma^2)$. Then, our posterior predictive pdf is

$$p(x_\star\,|\,\boldsymbol{x}) = \int \underbrace{p(x_\star\,|\,\boldsymbol{\theta}, \boldsymbol{x})}_{\mathcal{N}(\boldsymbol{h}_\star^T\boldsymbol{\theta},\sigma^2)} \cdot \underbrace{p(\boldsymbol{\theta}\,|\,\boldsymbol{x})}_{\mathcal{N}(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}, C_{\mathrm{MAP}})}\, d\boldsymbol{\theta}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= (\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}\,(\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{x} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}}) \\
C_{\mathrm{MAP}} &= (\boldsymbol{H}^T\boldsymbol{C}_{\boldsymbol{w}}^{-1}\boldsymbol{H} + \boldsymbol{C}_{\boldsymbol{\theta}}^{-1})^{-1}
\end{aligned}$$

implying

$$p(x_\star\,|\,\boldsymbol{x}) = \mathcal{N}(\boldsymbol{h}_\star^T\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}, \boldsymbol{h}_\star^T C_{\mathrm{MAP}}\boldsymbol{h}_\star + \sigma^2).$$

Gaussian Linear Model – Example

(Example 10.2 in Kay-I). DC-level estimation in white Gaussian noise with known variance — the same as the example on pp. 16–19 of this handout (except for the notation, which is now the same as in Kay-I).

$$\boldsymbol{x} = \mathbf{1}\,A + \boldsymbol{w}.$$

The additive noise $\boldsymbol{w}$ follows a $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ distribution and the *prior* on A is chosen to be $\mathcal{N}(\mu_A, \sigma_A^2)$. Then, applying Theorem 3 yields the MMSE estimate of A:

$$\widehat{A} = \mathrm{E}\,[A|\boldsymbol{X} = \boldsymbol{x}] = (\mathbf{1}^T\mathbf{1}/\sigma^2 + 1/\sigma_A^2)^{-1}\,(\mathbf{1}^T\boldsymbol{x}/\sigma^2 + \mu_A/\sigma_A^2)$$

i.e.

$$\widehat{A} = \underbrace{\frac{\frac{N}{\sigma^2}\overline{x} + \frac{1}{\sigma_A^2}\mu_A}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}}}_{\text{same as (11)}} = \alpha\,\overline{x} + (1-\alpha)\,\mu_A$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x[i]$ is the sample mean and

$$\alpha = \frac{1}{1 + \sigma^2/(\sigma_A^2 N)}.$$

Application: Signal Waveform Estimation

Sensor array signal processing model:

$$\boldsymbol{x}(t) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{s}(t) + \boldsymbol{w}(t)$$

where the dimension of $\boldsymbol{x}(t)$ is the number of sensors, t is time, $\boldsymbol{\theta}$ is direction, $\boldsymbol{s}(t)$ is the vector of signal waveforms, and $\boldsymbol{e}(t)$ is noise.

Suppose we wish to estimate $\boldsymbol{s}(t)$, assuming everything else is known! For notational simplicity, we will use $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta})$.

Let $\boldsymbol{w}(t) \sim \mathcal{N}_{\rm c}(\mathbf{0}, \sigma^2 I)$ and let us assign a prior pdf on the signals $\boldsymbol{s}(t) \sim \mathcal{N}_{\rm c}(\mathbf{0}, \boldsymbol{P})$. Then, the MMSE estimates of $\boldsymbol{s}(t)$ are:

$$\widehat{\boldsymbol{s}}(t) = \boldsymbol{P}\boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^H + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{x}(t).$$

The MMSE estimate outperforms (in the BMSE sense) the “usual” LS estimate:

$$\widehat{\boldsymbol{s}}_{\rm LS}(t) = (\boldsymbol{A}^H\boldsymbol{A})^{-1}\boldsymbol{A}^H\boldsymbol{x}(t)$$

since it exploits additional information provided by the prior distribution. Of course, this holds only if the assumed model is correct.

Cost Functions for Bayesian Estimation and the Corresponding "Optimal" Estimators

Define the estimation error $\epsilon = \theta - \widehat{\theta}$ and assign a loss (cost) $\mathrm{L}(\epsilon)$. We may choose $\widehat{\theta}$ to minimize the *Bayes (preposterior) risk*:

$$\mathrm{E}_{\boldsymbol{x},\theta}[\mathrm{L}(\epsilon)] = \mathrm{E}_{\boldsymbol{x},\theta}[\mathrm{L}(\theta - \widehat{\theta})]$$

but this is equivalent to minimizing the *posterior expected loss*:

$$\rho(\widehat{\theta} \,|\, \boldsymbol{x}) = \int \mathrm{L}(\theta - \widehat{\theta})\, p(\theta \,|\, \boldsymbol{x})\, d\theta$$

for each $\boldsymbol{x}$, which is what *true Bayesians* would do. The proof is the same as before (trivially extending the case of the squared-error loss):

$$\mathrm{E}_{\boldsymbol{x},\theta}[\mathrm{L}(\theta - \widehat{\theta})] = \mathrm{E}_{\boldsymbol{x}}\{\underbrace{\mathrm{E}_{\theta \,|\, \boldsymbol{x}}[\mathrm{L}(\theta - \widehat{\theta})]}_{\rho(\widehat{\theta} \,|\, \boldsymbol{x})}\}.$$

Here are a few popular loss functions:

1. $\mathrm{L}(\epsilon) = \epsilon^2$ (the popular squared-error loss, accurate),

2. $\mathrm{L}(\epsilon) = |\epsilon|$ (robust to outliers),

3. $\mathrm{L}(\epsilon) = \begin{cases} 0, & |\epsilon| \le \Delta/2 \\ 1, & |\epsilon| > \Delta/2 \end{cases}$ (0-1 loss, tractable).

corresponding to:

1. **MMSE.** $\mathrm{E}\left[\theta|\boldsymbol{x}\right]$, the *posterior mean* (which we proved earlier in this handout).

2. **Posterior median.** the optimal $\widehat{\theta}$ satisfies:

$$\int_{-\infty}^{\widehat{\theta}} p(\theta \,|\, \boldsymbol{x}) \, d\theta = \int_{\widehat{\theta}}^{\infty} p(\theta \,|\, \boldsymbol{x}) \, d\theta$$

   (HW: check this).

3. **MAP (maximum *a posteriori*) estimator.** the optimal $\widehat{\theta}$ satisfies:

$$\arg\max_{\widehat{\theta}} p_{\theta \,|\, \boldsymbol{x}}(\widehat{\theta} \,|\, \boldsymbol{x})$$

   also known as the *posterior mode*.

As an example, we now show the result in 3.

MAP Estimation:

$$\mathrm{E}_{\,x}\Big\{\mathrm{E}_{\,\theta|x}[\mathrm{L}(\epsilon)]\Big\} = \mathrm{E}_{\,x}\Big\{1 - \int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} p_{\theta \,|\, \boldsymbol{x}}(\theta \,|\, x) \, d\theta\Big\}.$$

To minimize this expression, we maximize $\int_{\widehat{\theta}-\Delta/2}^{\widehat{\theta}+\Delta/2} p_{\theta \,|\, \boldsymbol{x}}(\theta \,|\, \boldsymbol{x}) \, d\theta$ with respect to $\widehat{\theta}$. For very small Δ, maximizing the above expression is equivalent to maximizing $p_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}(\widehat{\theta}\,|\,\boldsymbol{x})$ with respect to $\widehat{\theta}$.

The loss function for the vector case:

$$\mathrm{L}(\boldsymbol{\epsilon}) = \begin{cases} 0, & \|\boldsymbol{\epsilon}\| \le \Delta/2 \\ 1, & \|\boldsymbol{\epsilon}\| > \Delta/2 \end{cases} .$$

so (as usual)

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})$$

and, consequently,

$$\boxed{\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}}[\log p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})]} .$$

Note that $\log p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ is the log-likelihood function. Thus, for a *flat prior*

$$\pi(\boldsymbol{\theta}) \propto 1$$

we have MAP $\equiv$ ML.

Example: Wiener Filter

Estimate noise-corrupted signal $s[n]$:

$$x[n] = \theta[n] + w[n].$$

Given $x[0], x[1], \ldots, x[N-1]$, write

$$\boldsymbol{x} = \boldsymbol{\theta} + \boldsymbol{w}$$

where

$$\boldsymbol{x} = \left[\begin{array}{c} x[0] \\ x[1] \\ \vdots \\ x[N-1]]^T \end{array} \right], \; \boldsymbol{\theta} = \left[\begin{array}{c} \theta[0] \\ \theta[1] \\ \vdots \\ \theta[N-1] \end{array} \right], \; \boldsymbol{w} = \left[\begin{array}{c} w[0] \\ w[1] \\ \vdots \\ w[N-1] \end{array} \right].$$

Assume

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_\theta) \text{ (prior pdf)}, \quad \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}_w).$$

Then, the MAP (posterior mode) $=$ MMSE (posterior mean) $=$ median Bayesian estimator is

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{C}_w^{-1} + \boldsymbol{C}_\theta^{-1})^{-1}(\boldsymbol{C}_w^{-1}\boldsymbol{x} + \boldsymbol{C}_\theta^{-1}\mathbf{0}) = \boldsymbol{C}_\theta(\boldsymbol{C}_\theta + \boldsymbol{C}_w)^{-1}\boldsymbol{x} \tag{18}$$

which follows from our general Gaussian linear model results.

If $x[n]$ is a wide-sense stationary (WSS) process, we obtain the well-known frequency-domain result:

$$\widehat{\Theta}(\omega) = \frac{P_\theta(\omega)}{P_\theta(\omega) + P_w(\omega)} X(\omega) \tag{19}$$

where $P_\theta(\omega)$ and $P_w(\omega)$, $\omega \in (-\pi, \pi)$ are the PSDs of the random processes $\theta[n]$ and $w[n]$, $n = 0, 1, \ldots$, and $X(\omega)$ is the DTFT of $x[n]$, $n = 0, 1, \ldots, N-1$.

Proof. The proof is standard: $\boldsymbol{C}_\theta, \boldsymbol{C}_w$ are Töplitz, asymptotically circulant, eigenvalues are PSDs etc., see e.g. Ch. 12.7 in Kay-I or Ch. 2.4 in Kay-II. In particular, for large N (i.e. asymptotically), we can approximate $\boldsymbol{C}_\theta$ and $\boldsymbol{C}_w$ with circulant matrices $\widetilde{\boldsymbol{C}}_\theta$ and $\widetilde{\boldsymbol{C}}_w$. The eigenvectors of circulant matrices of size N are the following DFT vectors:

$$\boldsymbol{u}_i = N^{-1/2} \left[1, \exp(j\omega_i), \exp(j2\omega_i), \ldots, \exp(j(N-1)\omega_i)\right]^T$$

where $i = 0, 1, \ldots, N-1$, $\omega_i = 2\pi i/N$, and the corresponding eigenvalues are $P_\theta(\omega_i)$ (for $\widetilde{\boldsymbol{C}}_\theta$) and $P_w(\omega_i)$ (for $\widetilde{\boldsymbol{C}}_w$). Let us construct the eigenvector matrix:

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N]$$

and

$$\begin{aligned}\boldsymbol{P}_\theta &= \mathrm{diag}\{P_\theta(\omega_0), P_\theta(\omega_1), \ldots, P_\theta(\omega_{N-1})\}\\ \boldsymbol{P}_w &= \mathrm{diag}\{P_w(\omega_0), P_w(\omega_1), \ldots, P_w(\omega_{N-1})\}.\end{aligned}$$

Note that $\boldsymbol{U}$ is orthonormal: $\boldsymbol{U}\boldsymbol{U}^H = \boldsymbol{U}^H\boldsymbol{U} = \boldsymbol{I}$; it is also the DFT matrix, i.e. $\boldsymbol{U}^H\boldsymbol{x} = \mathrm{DFT}(\boldsymbol{x})$. Finally, we can approximate (18) as

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &\approx \boldsymbol{U}\boldsymbol{P}_\theta\boldsymbol{U}^H(\boldsymbol{U}\boldsymbol{P}_\theta\boldsymbol{U}^H + \boldsymbol{U}\boldsymbol{P}_w\boldsymbol{U}^H)^{-1}\boldsymbol{x} \quad \Longrightarrow\\ \mathrm{DFT}(\widehat{\boldsymbol{\theta}}) &= \boldsymbol{U}^H\widehat{\boldsymbol{\theta}} = \underbrace{\boldsymbol{P}_\theta(\boldsymbol{P}_\theta + \boldsymbol{P}_w)^{-1}}_{\text{diagonal matrix}}\mathrm{DFT}(\boldsymbol{x})\end{aligned}$$

and (19) follows (at least for $\omega = \omega_i = 2\pi i/N$, $i = 0, 1, \ldots, N-1$). $\square$

Example: MAP Denoising for Eddy-Current Data

Note: Continuation of the example on p. 53 of handout # 3.

For fixed $\boldsymbol{\lambda} = [a, b, c, d]^T$, we compute the joint MAP estimates of the signal amplitudes $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_{K-1}]^T$ and phases $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_{K-1}]^T$ by maximizing

$$\sum_{k=1}^{K} \log[p_{x \,|\, \theta}(x_k \,|\, \boldsymbol{\theta}_k) \cdot p_\alpha(\alpha_k; a, b) \cdot p_\beta(\beta_k; c, d)].$$

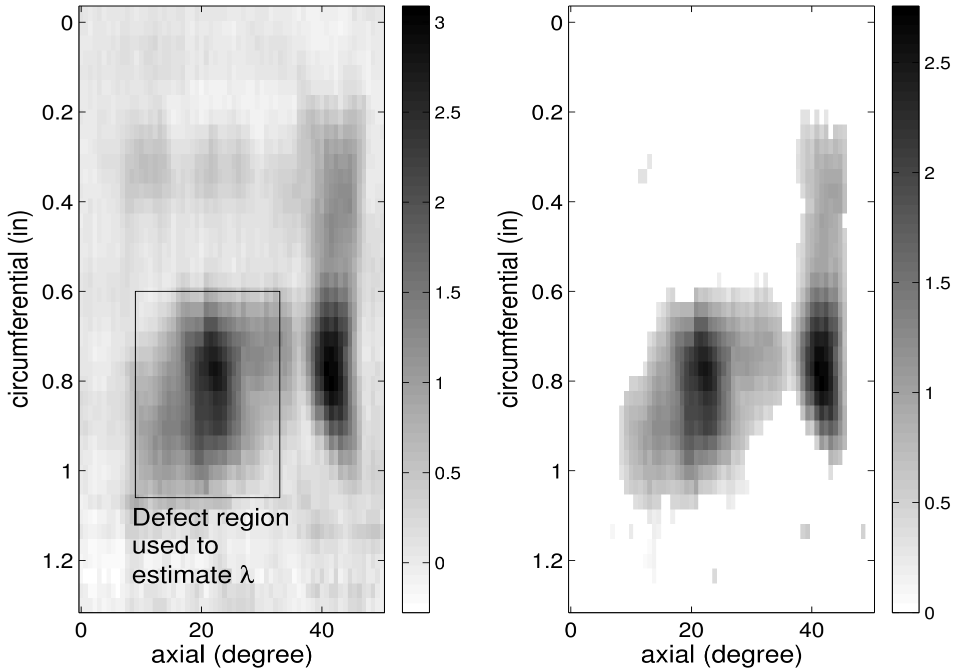
Here are the MAP "denoising" steps:

- Estimate the model parameter vector $\boldsymbol{\lambda}$ from a "training" defect region (using the ML method, say),
- Apply the MAP algorithm to the whole image by replacing $\boldsymbol{\lambda}$ with its ML estimate obtained from the training region.

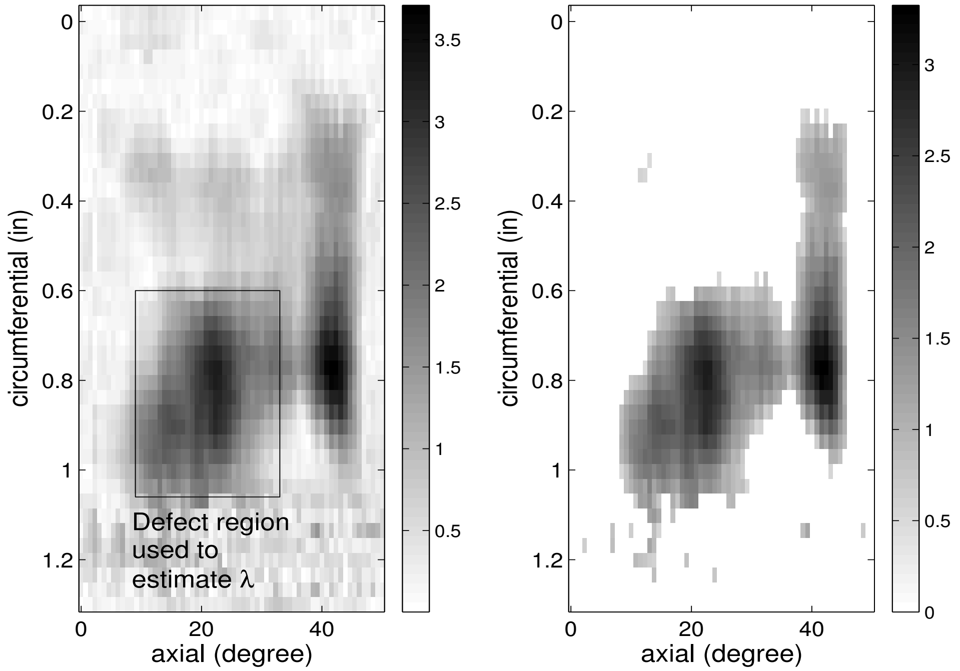
The MAP filter enhances the signals that have similar amplitude and phase distributions to the "training" defect signal and suppresses other signals.

The above procedure is an example of *empirical Bayesian estimation* $\Longrightarrow$ a somewhat successful combination of classical and Bayesian methods:

- model parameters $\boldsymbol{\lambda}$ estimated using classical (ML) method,
- signal amplitudes and phases $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ estimated using Bayesian (MAP) method.



Imaginary-component plots of original eddy-current data (left) and corresponding empirical MAP estimates (right).



Magnitude plots of original eddy-current data (left) and corresponding empirical MAP estimates (right).

MAP Computation

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$$

where

$$V(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x}).$$

Newton-Raphson Iteration:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \boldsymbol{H}_i^{-1} \boldsymbol{g}_i$$

where

$$\begin{aligned} \boldsymbol{g}_i &= \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}} \\ \boldsymbol{H}_i &= \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}. \end{aligned}$$

Comments:

- Newton Raphson is not guaranteed to converge but
- its convergence is very fast in the neighborhood of the MAP estimate.

Upon convergence (which we denote as $i \to \infty$), we have

$$\boldsymbol{g}_\infty = \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\infty)}=\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}} = \mathbf{0}. \tag{20}$$

Posterior Approximation Around the MAP Estimate

Expanding (in Taylor series) the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ around the MAP estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ and keeping the first three terms yields:

$$\begin{aligned}\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) &\approx \log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} \,|\, \boldsymbol{x}) \\ &+ \tfrac{1}{2}\, (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}})^T \, \frac{\partial^2 \log p(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}). \qquad (21)\end{aligned}$$

The second term in the Taylor-series expansion vanishes because the log posterior pdf/pmf has zero derivative at the MAP estimate, see (20).

If the number of observations is large, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ will be *unimodal*. Furthermore, if $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is in the interior of the parameter space $\boldsymbol{\Theta}$ (preferably *far from the boundary* of $\boldsymbol{\Theta}$), we can use the following approximation for the posterior distribution:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \approx \mathcal{N}\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}, -\Big[\frac{\partial^2 \log p(\boldsymbol{\theta} \,|\, \boldsymbol{x})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}}\Big]^{-1}\Big).$$

This approximation follows by looking at (21) as a function of $\boldsymbol{\theta}$ and observing that $\log p(\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} \,|\, \boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$.

We may obtain another (simpler, but typically poorer) approximation by replacing the covariance matrix of the above Gaussian distribution with the (classical, non-Bayesian) CRB evaluated at $\widehat{\boldsymbol{\theta}}_{\rm MAP}$.

If $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ has multiple modes (and we can find them all), we can use a Gaussian mixture (or perhaps a more general t-distribution mixture) to approximate $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$.

A Bit More on Asymptotic Normality and Consistency for Bayesian Models

Assume that the observations $x_1, x_2, \ldots, x_n$ are conditionally i.i.d. (given θ), following

$$x_1, x_2, \ldots, x_n \quad \sim \quad \underbrace{p_{\text{TRUE}}(x)}_{\text{the true distribution of the data}} \quad .$$

As before, we also have

$p(x \,|\, \theta)$ – data model, likelihood;

$\pi(\theta)$ – prior distribution on θ.

Define

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}.$$

First, note that we may not be modeling the data correctly. Here, modeling the data correctly means that $p_{\text{TRUE}}(x) = p(x \,|\, \theta_0)$ for some θ_0. For simplicity, we consider the case of a scalar parameter θ, but the results can be generalized.

Recall the definition of the *Kullback-Leibler distance* $D(\boldsymbol{p} \,\|\, \boldsymbol{q})$ from one pmf ($\boldsymbol{p}$) to another ($\boldsymbol{p}$):

$$D(\boldsymbol{p} \,\|\, \boldsymbol{q}) = \sum_k p_k \log \frac{p_k}{q_k}$$

and apply it to measure the distance between $p_{\text{TRUE}}(x)$ and $p(x \,|\, \theta)$ (for the case where these distributions are discrete, i.e. pmfs):

$$H(\theta) = \sum_x p_{\text{TRUE}}(x) \log \frac{p_{\text{TRUE}}(x)}{p(x \,|\, \theta)}.$$

If the data is modeled correctly, then $p_{\text{TRUE}}(x) = p(x \,|\, \theta_0)$ and, consequently, $H(\theta)$ is minimized at θ_0, yielding $H(\theta_0) = 0$. **In the following discussion, we assume that the data are modeled correctly.**

Theorem 4. **[Convergence in discrete parameter space.]** *If the parameter space* Θ *is finite and* $P\{\theta = \theta_0\} > 0$, *then*

$$P\{\theta = \theta_0 \,|\, \boldsymbol{x}\} \to 1 \quad \textit{as} \quad n \to \infty.$$

Proof. Consider the *log posterior odds*:

$$\log\left(\frac{p(\theta \,|\, \boldsymbol{x})}{p(\theta_0 \,|\, \boldsymbol{x})}\right) = \log\left(\frac{\pi(\theta)}{\pi(\theta_0)}\right) + \sum_{i=1}^{n} \log\left(\frac{p(x_i \,|\, \theta)}{p(x_i \,|\, \theta_0)}\right). \quad (22)$$

The second term in this expression is the sum of n conditionally i.i.d. random variables (given θ and θ_0). Recall that $x_1, x_2, \ldots, x_n$ are coming from $p_{\mathrm{TRUE}}(\cdot) = p(\cdot \,|\, \theta_0)$.

Then

$$\mathrm{E}_{\,x|\theta_0}\Big[\log\Big(\frac{p(X_i \,|\, \theta)}{p(X_i \,|\, \theta_0)}\Big)\Big] = -H(\theta) \le 0.$$

If $\theta \neq \theta_0$, the second term in (22) is the sum of n i.i.d. random variables with negative mean, which should diverge to $-\infty$ as $n \to \infty$. As long as $P\{\theta = \theta_0\} = \pi(\theta_0) > 0$, making the first term in (22) finite, the log posterior odds $\to -\infty$ as $n \to \infty$. Thus, if $\theta \neq \theta_0$, the posterior odds go to zero:

$$\frac{p(\theta \,|\, \boldsymbol{x})}{p(\theta_0 \,|\, \boldsymbol{x})} \to 0$$

which implies $p(\theta \,|\, \boldsymbol{x}) \to 0$. As all the probabilities summed over all values of θ must add to one, we have

$$p(\theta_0 \,|\, \boldsymbol{x}) \to 1.$$

□

Theorem 5. **[Convergence in continuous parameter space.]** *If θ is defined on a compact set (i.e. closed and bounded) and A is a neighborhood of θ_0 (more precisely, A is an open subset of the parameter space containing θ_0) with* *prior probability $\pi(\theta)$ satisfying $\int_{\theta \in A} \pi(\theta)\, d\theta > 0$, then*

$$P\{\theta \in A \,|\, \boldsymbol{x}\} \to 1 \quad \textit{as} \quad n \to \infty.$$

Proof. The proof is similar in spirit to the proof for the discrete case. □

Technical details:

- In many popular continuous-parameter scenarios, the parameter space is not a compact set: e.g. the parameter space for the mean of a Gaussian random variable is $(-\infty, \infty)$. Luckily, for most problems of interest, the compact-set assumption of Theorem 5 can be relaxed.

- Similarly, Theorem 4 can often be extended to allow for an infinite discrete parameter space.

Theorem 6. **[Asymptotic Normality of $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$]** *Under some regularity conditions (particularly that θ_0 is not on the boundary of the parameter space) and under the conditional i.i.d. measurement model from p. 71, as $n \to \infty$,*

$$\sqrt{n}\,(\widehat{\theta}_{\mathrm{MAP}} - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{I}_1(\theta_0)^{-1})$$

where $\mathcal{I}_1(\theta_0)$ is the (classical, non-Bayesian) Fisher information *of a single measurement (say x_1):*

$$\begin{aligned}\mathcal{I}_1(\theta) &= \mathrm{E}_{p(x_1 \,|\, \boldsymbol{\theta})}\Big[\Big(\frac{d \log p(X_1 \,|\, \theta)}{d\theta}\Big)^2 \,\Big|\, \theta\Big] \\ &= -\mathrm{E}_{p(x_1 \,|\, \boldsymbol{\theta})}\Big[\frac{d^2 \log p(X_1 \,|\, \theta)}{d\theta^2} \,\Big|\, \theta\Big].\end{aligned}$$

Proof. See Appendix B in Gelman, Carlin, Stern, and Rubin. □

Here are some useful observations to help justify Theorem 6. Consider (scalar version of) the Taylor-series expansion in (21):

$$\begin{aligned}\log p(\theta \,|\, \boldsymbol{x}) &\approx \log p(\widehat{\theta}_{\mathrm{MAP}} \,|\, \boldsymbol{x}) \\ &\quad + \tfrac{1}{2}\,(\theta - \widehat{\theta}_{\mathrm{MAP}})^2\, \frac{d^2}{d\theta^2}[\log p(\widehat{\theta}_{\mathrm{MAP}} \,|\, \boldsymbol{x})].\end{aligned}$$

Now, study the behavior of

negative Hessian of the log posterior at $\theta = -\dfrac{d^2 \log p(\theta \,|\, \boldsymbol{x})}{d\theta^2}$

$$\begin{aligned}&= -\frac{d^2 \log \pi(\theta)}{d\theta^2} - \frac{d^2 \log p(\boldsymbol{x} \,|\, \theta)}{d\theta^2} \\ &= -\frac{d^2 \log \pi(\theta)}{d\theta^2} - \sum_{i=1}^{n} \frac{d^2 \log p(x_i \,|\, \theta)}{d\theta^2}\end{aligned}$$

and, therefore,

$$\mathrm{E}_{\boldsymbol{x}\,|\,\theta}\Big[-\frac{d^2 \log p(\theta\,|\,\boldsymbol{x})}{d\theta^2}\,\Big|\,\theta\Big] = -\frac{d^2 \log \pi(\theta)}{d\theta^2} + n\,\mathcal{I}_1(\theta)$$

implying that, as n grows,

negative Hessian of log posterior at $\theta \approx n\,\mathcal{I}_1(\theta)$.

To summarize: Asymptotically (i.e. for large number of samples n), Bayesian (MAP, in particular) and classical approaches give equivalent answers.

MAP Estimator Computation for Multiple (Subsets of) Parameters

Consider again the case of multiple (subsets of) parameters, denoted by $\boldsymbol{\theta}$ and $\boldsymbol{u}$, i.e. the (overall) vector of the unknown parameters is $[\boldsymbol{\theta}^T, \boldsymbol{u}^T]^T$ The joint posterior pdf for $\boldsymbol{\theta}$ and $\boldsymbol{u}$ is

$$p(\boldsymbol{\theta}, \boldsymbol{u} \,|\, \boldsymbol{x}) = p(\boldsymbol{u} \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \cdot p(\boldsymbol{\theta} \,|\, \boldsymbol{x}).$$

We wish to estimate both $\boldsymbol{\theta}$ and $\boldsymbol{u}$.

Here is our **first attempt at estimating θ and u:** maximize the marginal posterior pdfs/pmfs

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

and

$$\widehat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} p(\boldsymbol{u} \,|\, \boldsymbol{x})$$

which take into account the uncertainties about the other parameter. This is perhaps the most desirable approach for estimating $\boldsymbol{\theta}$ and $\boldsymbol{u}$. Note that we can do the two optimizations (with respect to $\boldsymbol{\theta}$ and $\boldsymbol{u}$) separately.

But, what if we cannot easily obtain these two marginal posterior pdfs/pmfs? Suppose now that we can obtain $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$

but $p(\boldsymbol{u} \,|\, \boldsymbol{x})$ is not easy to handle. Here is our **second attempt at estimating θ and u:**

1. First, find the marginal MAP estimate of $\boldsymbol{\theta}$:

   $$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

   which, as desired, takes into account the uncertainty about $\boldsymbol{u}$ (by integrating $\boldsymbol{u}$ out from the joint posterior).

2. Then, find the conditional MAP estimate of $\boldsymbol{u}$ by maximizing $p(\boldsymbol{u} \,|\, \boldsymbol{\theta}, \boldsymbol{x})$:

   $$\widehat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} p(\boldsymbol{u} \,|\, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \boldsymbol{x}).$$

Finally, what if we cannot easily obtain either of the two marginal posterior pdfs/pmfs? Here is our **third attempt at estimating θ and u:** find the joint MAP estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{u}})$,

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{u}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{u}} p(\boldsymbol{\theta}, \boldsymbol{u} \,|\, \boldsymbol{x}).$$

This estimation is sometimes done as follows: iterate between

1. finding the conditional MAP estimate of $\boldsymbol{\theta}$ by maximizing

$p(\boldsymbol{\theta} \,|\, \boldsymbol{u}_p, \boldsymbol{x})$:

$$\boldsymbol{\theta}_{p+1} = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta} \,|\, \boldsymbol{u}}(\boldsymbol{\theta} \,|\, \boldsymbol{u}_p, \boldsymbol{x})$$

and

2. finding the conditional MAP estimate of $\boldsymbol{u}$ by maximizing $p(\boldsymbol{u} \,|\, \boldsymbol{\theta}_{p+1}, \boldsymbol{x})$:

$$\boldsymbol{u}_{p+1} = \arg\max_{\boldsymbol{u}} p_{\boldsymbol{u} \,|\, \boldsymbol{\theta}}(\boldsymbol{u} \,|\, \boldsymbol{\theta}_{p+1}, \boldsymbol{x})$$

known as the *iterated conditional modes (ICM) algorithm*. This is just an application of the *stepwise-ascent approach to optimization* — we have seen it before, e.g. in the ECM algorithm.

EM Algorithm for Computing the Marginal MAP Estimator

Let us continue with the multiparameter model from the previous page and assume that we wish to find the marginal MAP estimate of $\boldsymbol{\theta}$. Denote the observed data by $\boldsymbol{x}$. We wish to maximize

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{u} \,|\, \boldsymbol{x})\, d\boldsymbol{u}.$$

but this may be difficult to do directly. However, if maximizing

$$p(\boldsymbol{\theta}, \boldsymbol{u} \,|\, \boldsymbol{x}) = p(\boldsymbol{u} \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \cdot p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

is easy, we can treat $\boldsymbol{u}$ as the *missing data* and apply the EM algorithm!

Now

$$\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \log p(\boldsymbol{\theta}, \boldsymbol{u} \,|\, \boldsymbol{x}) - \log p(\boldsymbol{u} \,|\, \boldsymbol{\theta}, \boldsymbol{x}).$$

and let us take the expectation of this expression with respect to $p(\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x})$:

$$\begin{aligned}\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \quad &= \quad \underbrace{\mathrm{E}_{\,p(\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x})}[\log p(\boldsymbol{\theta}, \boldsymbol{U} \,|\, \boldsymbol{x}) \,|\, \boldsymbol{\theta}_p, \boldsymbol{x}]}_{Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p)} \\ &\qquad - \underbrace{\mathrm{E}_{\,p(\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x})}[\log p(\boldsymbol{U} \,|\, \boldsymbol{\theta}, \boldsymbol{x}) \,|\, \boldsymbol{\theta}_p, \boldsymbol{x}]}_{H(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p)}.\end{aligned}$$

Recall that our goal is to maximize $\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ with respect to $\boldsymbol{\theta}$. The key to the missing information principle is that $H(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p)$ is maximized (with respect to $\boldsymbol{\theta}$) by $\boldsymbol{\theta} = \boldsymbol{\theta}_p$, which we showed in handout `emlecture`, see (2) in handout `emlecture`. Hence, finding a $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p)$ will increase $\log p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$:

$$\begin{aligned} &\log p(\boldsymbol{\theta}_{p+1} \,|\, \boldsymbol{x}) - \log p(\boldsymbol{\theta}_p \,|\, \boldsymbol{x}) \\ &\quad = \underbrace{Q(\boldsymbol{\theta}_{p+1} \,|\, \boldsymbol{\theta}_p) - Q(\boldsymbol{\theta}_p \,|\, \boldsymbol{\theta}_p)}_{\geq 0, \text{ since } Q \text{ is increased}} \\ &\quad + \underbrace{H(\boldsymbol{\theta}_p \,|\, \boldsymbol{\theta}_p) - H(\boldsymbol{\theta}_{p+1} \,|\, \boldsymbol{\theta}_p)}_{\geq 0, \text{ by the fact that } H(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p) \leq H(\boldsymbol{\theta}_p \,|\, \boldsymbol{\theta}_p)} \geq 0. \end{aligned}$$

EM Algorithm:

- Denote the estimate at the pth step by $\boldsymbol{\theta}_p$.
- **E Step:** Compute

$$\begin{aligned} Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_p) &= \mathrm{E}_{p(\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x})}[\log p(\boldsymbol{U}, \boldsymbol{\theta} \,|\, \boldsymbol{x}) \,|\, \boldsymbol{\theta}_p, \boldsymbol{x}] \\ &= \int \log p(\boldsymbol{u}, \boldsymbol{\theta} \,|\, \boldsymbol{x}) \, p_{\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x}}(\boldsymbol{u} \,|\, \boldsymbol{\theta}_p, \boldsymbol{x}) \, d\boldsymbol{u}. \end{aligned}$$

We need to average the *complete-data log-posterior function* over the *conditional posterior distribution of $\boldsymbol{u}$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_p$*.

- **M step:** Maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_p)$ with respect to $\boldsymbol{\theta}$, yielding $\boldsymbol{\theta}_{p+1}$.

Using similar arguments as in likelihood maximization, we can show that $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ increases in each EM iteration step.

Example (from Ch. 12.3 in Gelman, Carlin, Stern, and Rubin): Consider the classical DC-level estimation problem where the measurements $x_i \mid \mu, \sigma^2$, $i = 1, 2, \ldots, N$ follow the $\mathcal{N}(\mu, \sigma^2)$ distribution with both μ and σ^2 *unknown*.

Choose *semi-conjugate* priors, i.e. μ and σ^2 are independent *a priori*:

$$\underbrace{\pi(\mu, \sigma^2) = \pi(\mu) \cdot \pi(\sigma^2)}_{\text{not conjugate for } \mu \text{ and } \sigma^2}$$

and

$$\begin{aligned} \pi(\mu) &= \mathcal{N}(\mu_0, \tau_0^2) \\ \pi(\sigma^2) &\propto \underbrace{1/\sigma^2}_{\text{Jeffreys' prior}} \end{aligned} .$$

Comments

- **(A "definition" of a semi-conjugate prior):** If σ^2 were known, the above $\pi(\mu)$ would be a conjugate prior for μ. Similarly, if μ were known, the above $\pi(\sigma^2)$ would be a conjugate prior for μ. However, $\pi(\mu, \sigma^2) = \pi(\mu) \cdot \pi(\sigma^2)$ is *not* a conjugate prior for both μ and σ^2, hence the prefix "semi."

(Conjugate prior is obscure): A conjugate prior exists for μ and σ^2 under the above model, but it falls into the "obscure" category that we mentioned earlier. For example, this conjugate prior $\pi(\mu, \sigma^2)$ does not allow *a priori* independence of μ and σ^2.

If σ^2 were known, our job would be really easy — the MAP estimate of μ for this case is given by

$$\mu_{\text{MAP}} = \frac{\frac{1}{\tau_0^2}\,\mu_0 + \frac{N}{\sigma^2}\,\overline{x}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}} \tag{23}$$

see equation (11) and Example 10.2 in Kay-I.

Since we assume that μ and σ^2 are independent *a priori*, this problem does not have a closed-form solution for the MAP estimate of μ. In most applications that come to mind, it appears intuitively appealing to have μ and σ^2 independent *a priori*.

We wish to find the marginal posterior mode (MAP estimate) of μ. Hence, μ corresponds to θ and σ^2 corresponds to the missing data u.

We now derive the EM algorithm for the above problem. The joint posterior pdf is

$$p(\mu, \sigma^2 \,|\, \boldsymbol{x}) \propto \overbrace{\exp\left[-\frac{1}{2\,\tau_0^2}(\mu-\mu_0)^2\right]\cdot(\sigma^2)^{-1}}^{\text{prior pdf}}$$

$$\underbrace{\cdot\,(\sigma^2)^{-N/2}\cdot\exp\left[-\frac{1}{2\,\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right]}_{\text{likelihood}}. \tag{24}$$

This joint posterior pdf is key to all steps of our EM algorithm derivation. First, write down the log of (24) as follows:

$$\log p(\mu, \sigma^2 \,|\, \boldsymbol{x}) = \underbrace{\text{const}}_{\text{not a function of } \mu} -\frac{1}{2\,\tau_0^2}(\mu-\mu_0)^2$$

$$-\frac{1}{2\,\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 + \text{const}.$$

Why can we ignore terms in the above expression that do not contain μ? Because the maximization in the M step will be with respect to μ.

Now, we need the find the *conditional posterior pdf of the* *"missing" σ^2 given μ and $\boldsymbol{x}$, evaluated at $\mu = \mu_p$*:

$$p(\sigma^2 \,|\, \mu_p, \boldsymbol{x}) \propto (\sigma^2)^{-N/2-1} \cdot \exp\Big[-\frac{1}{2\,\sigma^2} \sum_{i=1}^{N} (x_i - \mu_p)^2 \Big]$$

(look up the table of distributions)

is the kernel of $\quad \text{Inv-}\chi^2\Big(N, \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2\Big)$. (25)

We are now ready to derive the EM algorithm for this problem.

E Step: Conditional on μ_p and $\boldsymbol{x}$, find the expectation of $\log p(\mu, \sigma^2|\boldsymbol{x})$ by averaging over σ^2:

$$\begin{aligned} Q(\mu \,|\, \mu_p) &= \mathrm{E}_{\,p(\sigma^2|\mu_p, \boldsymbol{x})}\Big[\log p(\mu, \sigma^2|\boldsymbol{x}) \,|\, \mu_p, \boldsymbol{x}\Big] \\ &= \underbrace{\text{const}}_{\text{not a function of } \mu} \qquad \underbrace{-\frac{1}{2\,\tau_0^2}\,(\mu - \mu_0)^2}_{\text{no } \sigma^2\text{, expectation disappears}} \\ &\quad -\tfrac{1}{2} \cdot \mathrm{E}_{\,p(\sigma^2|\mu_p, \boldsymbol{x})}\Big[\frac{1}{\sigma^2} \,\Big|\, \mu_p, \boldsymbol{x}\Big] \cdot \sum_{i=1}^{N}(x_i - \mu)^2. \end{aligned}$$

Comments:

- We need to maximize the above expression with respect to μ.

- Now, we need to evaluate $\mathrm{E}_{p(\sigma^2|\mu_p,\boldsymbol{x})}\left[\frac{1}{\sigma^2}\,\middle|\,\mu_p,\boldsymbol{x}\right]$. But, (25) implies that $\sigma^2|\mu_p,\boldsymbol{x}$ is distributed as

$$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_p)^2\cdot N}{Z}$$

where Z is a (central) χ^2_N random variable, see p. 27 of this handout. Since the mean of any χ^2_N random variable is N (see the distribution table), we have:

$$\mathrm{E}_{p(\sigma^2|\mu_p,\boldsymbol{x})}\left[\frac{1}{\sigma^2}\,\middle|\,\mu_p,\boldsymbol{x}\right]=\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_p)^2\right]^{-1}$$

which is intuitively appealing — this expression is simply an inverse of the sample estimate of σ^2 (with the *known* mean

μ replaced by its latest estimate μ_p). Finally

$$
\begin{aligned}
& Q(\mu \,|\, \mu_p) \\
& = -\frac{1}{2\tau_0^2}\,(\mu - \mu_0)^2 \\
& -\tfrac{1}{2} \cdot \Big[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2\Big]^{-1} \cdot \sum_{j=1}^{N}(x_j - \mu)^2 + \underbrace{\text{const}}_{\text{not a function of } \mu} \\
& = \underbrace{\text{const}}_{\text{not a function of } \mu} - \frac{1}{2\tau_0^2}\,(\mu^2 - 2\,\mu\,\mu_0) \\
& -\tfrac{1}{2} \cdot \Big[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2\Big]^{-1} \cdot \Big[N\,\mu^2 - 2\,\mu\,(\sum_{j=1}^{N} x_j)\Big] \\
& = \text{const} - \tfrac{1}{2}\Big[\frac{1}{\tau_0^2} + \frac{N}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2}\Big]\,\mu^2 \\
& \qquad + \Big(\frac{\mu_0}{\tau_0^2} + \frac{N\,\overline{x}}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2}\Big) \cdot \mu. \qquad (26)
\end{aligned}
$$

M Step: Find μ that maximizes $Q(\mu \,|\, \mu_p)$ and choose it to be μ_{p+1}:

$$
\mu_{p+1} = \frac{\frac{1}{\tau_0^2} \cdot \mu_0 + \frac{N}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2} \cdot \overline{x}}{\frac{1}{\tau_0^2} + \frac{N}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2}} \qquad (27)
$$

which is very simple and intuitive. Compare (27) with (23) and note that $\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_p)^2$ estimates σ^2 based on μ_p. Our iteration (27) should converge to the marginal posterior mode of $p(\mu \,|\, \boldsymbol{x})$.

Simulation from Hierarchical Models: Composition Sampling

Consider the basic model with likelihood $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ and prior $\pi(\boldsymbol{\theta})$. To simulate N observations $\boldsymbol{x}_i$, $i = 1, 2, \ldots, N$ from the marginal pdf/pmf of $\boldsymbol{x}$, do the following:

(i) draw $\boldsymbol{\theta}_i$ from $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and

(ii) draw $\boldsymbol{x}_i$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ from $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta}_i)$.

Then, these $\boldsymbol{x}_i$, $i = 1, 2, \ldots, N$ correspond to samples from the marginal (pdf in this case):

$$p(\boldsymbol{x}) = \int p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})\,\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

and

$$\left[\begin{array}{c} \boldsymbol{x}_i \\ \boldsymbol{\theta}_i \end{array}\right] \quad \text{are draws from} \quad p_{\boldsymbol{x},\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta}) = p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

Similarly, to generate observations from the marginal pdf/pmf of $\boldsymbol{x}$, where the model is described by $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, $p_{\boldsymbol{\theta}\,|\,\boldsymbol{\lambda}}(\boldsymbol{\theta}\,|\,\boldsymbol{\lambda})$, and $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ [in this model, we assume that $\boldsymbol{x}$ depends on $\boldsymbol{\lambda}$ only through $\boldsymbol{\theta}$ and hence $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}) = p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ — *hierarchical structure*]:

(i) draw $\boldsymbol{\lambda}_i$ from $\pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$,

(ii) draw $\boldsymbol{\theta}_i$ given $\boldsymbol{\lambda} = \boldsymbol{\lambda}_i$ from $p_{\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}_i)$,

(iii) draw $\boldsymbol{x}_i$ given $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ from $p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}_i)$.

Then, $\boldsymbol{x}_i, \; i = 1, 2, \ldots, N$ are samples from

$$p(\boldsymbol{x}) = \int\!\!\int \underbrace{p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \, p_{\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \, \pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}_{p_{\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\lambda})} \, d\boldsymbol{\lambda} \, d\boldsymbol{\theta}$$

whereas $\boldsymbol{\theta}_i, \; i = 1, 2, \ldots, N$ are samples from

$$p(\boldsymbol{\theta}) \quad = \quad \int \underbrace{p_{\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}}(\boldsymbol{\theta} \,|\, \boldsymbol{\lambda}) \, \pi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}_{p_{\boldsymbol{\theta},\boldsymbol{\lambda}}(\boldsymbol{\theta},\boldsymbol{\lambda})} \, d\boldsymbol{\lambda}.$$

A BRIEF INTRODUCTION TO MONTE CARLO METHODS

Monte Carlo methods are useful when inference cannot be performed analytically.

Basic idea: Draw a large number of samples distributed according to some probability distribution and use them to obtain estimates, confidence regions etc. of unknown parameters.This part is based on STAT 515/601 notes by Prof. Kaiser.

We first describe three important problems where Monte Carlo methods are employed:

(*) Simulation-based inference,

(**) Monte Carlo integration, and

(***) Simulation from complex models

and then try to relate (*), (**) and (***).

Simulation-Based Inference (*)

By simulation, we mean drawing "samples" from probability distributions. Simulation-based inference is largely confined to the Bayesian analysis of models:

$p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ — data model;

$\boldsymbol{\theta}$ — "true state of nature";

$\pi(\boldsymbol{\theta})$ — prior distribution on $\boldsymbol{\theta}$.

Inference is based on the posterior distribution of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})\, \pi(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}}.$$

Note that $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ is in the form of a probability density or mass function.

For inference about $\boldsymbol{\theta}$, we need to be able to compute the posterior pdf or pmf $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. Often, however, the integral (in the continuous case)

$$\int p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}(\boldsymbol{x} \,|\, \boldsymbol{\vartheta})\, \pi_{\boldsymbol{\theta}}(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}$$

cannot be computed in closed form. This makes analytical analysis impossible or difficult. (We could resort to tractable

MAP estimation, which does not require integration, as discussed before. However, MAP approach provides only a point estimate of $\boldsymbol{\theta}$.)

What if we could simulate M "samples" $\boldsymbol{\vartheta}_m$, $m = 1, 2, \ldots, M$ of $\boldsymbol{\theta}|\boldsymbol{x}$ from the distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$? We could then estimate $\mathrm{E}_{\,\theta|x}(\boldsymbol{\theta}|\boldsymbol{x})$ as

$$\widehat{\boldsymbol{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\vartheta}_m, \qquad \boldsymbol{\vartheta}_m \quad \text{are samples from} \quad p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$$

and then similarly estimate $\mathrm{var}(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ and other desired quantities (moments, confidence regions, etc.).

Comments

- Once we have draws from $p(\boldsymbol{\theta}|\boldsymbol{x})$, we can easily estimate the marginal posterior pdf/pmf $p_{\theta_i|\boldsymbol{x}}(\theta_i|\boldsymbol{x})$ [of θ_i, the ith coordinate of $\boldsymbol{\theta}$] as shown below. If we have samples $\boldsymbol{\vartheta}_m \sim p(\boldsymbol{\theta}|\boldsymbol{x})$ where

$$\boldsymbol{\vartheta}_1 = \begin{bmatrix} \vartheta_{1,1} \\ \vdots \\ \vartheta_{1,p} \end{bmatrix}, \quad \boldsymbol{\vartheta}_2 = \begin{bmatrix} \vartheta_{2,1} \\ \vdots \\ \vartheta_{2,p} \end{bmatrix}, \quad \boldsymbol{\vartheta}_M = \begin{bmatrix} \vartheta_{M,1} \\ \vdots \\ \vartheta_{M,p} \end{bmatrix}$$

  then $\{\vartheta_{m,i}, \; m = 1, 2, \ldots, M\}$ are samples from the marginal pdfs/pmfs $p_{\theta_i\,|\,\boldsymbol{x}}(\theta_i\,|\,\boldsymbol{x})$, $i = 1, 2, \ldots, p$. We can then obtain histogram (or fancier kernel density estimates) of $p_{\theta_i|\boldsymbol{x}}(\theta_i|\boldsymbol{x})$, $i = 1, 2, \ldots, p$.

- If $\{\boldsymbol{\vartheta}_m, \; m = 1, 2, \ldots, M\}$ are i.i.d., then

$$\widehat{\theta}_i = \frac{1}{M} \sum_{m=1}^{M} \vartheta_{m,i} \xrightarrow{\text{w.p. } 1} \int \theta_i \, p_{\theta_i|\boldsymbol{x}}(\theta_i|\boldsymbol{x}) \, d\theta_i.$$

  This is also true for non-independent $\boldsymbol{\vartheta}_k$, as long as the sequence $\{\boldsymbol{\vartheta}_m, \; m = 1, 2, \ldots\}$ is *ergodic*.

Monte Carlo Integration (**)

Monte Carlo integration is useful for classical inference and for computing Bayes factors in Bayesian settings. In these applications, we wish to compute integrals of the form:

$$L = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \, \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

We could generate samples $\boldsymbol{\vartheta}_k$ of $\boldsymbol{\theta}$ from $\pi(\boldsymbol{\theta})$, and compute a Monte-Carlo estimate of this integral:

$$\widehat{L} = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}_m), \quad \boldsymbol{\vartheta}_m \equiv \text{samples from } \pi(\boldsymbol{\theta}).$$

Key observation: The accuracy of $\widehat{L}$ depends on M, but is independent of $\dim(\boldsymbol{\theta})$ (i.e. the dimensionality of integration). And, by the law of large numbers

$$\lim_{M \to \infty} \widehat{L} = L.$$

We can extend the above idea and modify the Monte Carlo estimate of L to

(i) improve $\operatorname{var}(\widehat{L})$ or

(ii) simplify implementation

or both. For example, for any distribution $\widetilde{p}(\boldsymbol{\theta})$ having the same support as $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ (i.e. is nonzero wherever $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ is nonzero), we have

$$L = \int p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \frac{\widetilde{p}(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})} \, d\boldsymbol{\theta}.$$

Now, we can generate samples $\boldsymbol{\vartheta}_k$ from $\widetilde{p}(\boldsymbol{\theta})$ and use the following Monte-Carlo estimate:

$$\widehat{L}(\boldsymbol{\lambda}) = \frac{1}{M} \sum_{m=1}^{M} \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\vartheta}_m)\, \pi(\boldsymbol{\vartheta}_m)}{\widetilde{p}(\boldsymbol{\vartheta}_m)}$$

where $\boldsymbol{\vartheta}_m$, $m = 1, 2, \ldots, M$ are samples from $\widetilde{p}(\boldsymbol{\theta})$. This is called *importance sampling*. The trick is to find a good proposal distribution $\widetilde{p}(\boldsymbol{\theta})$. Good $\widetilde{p}(\boldsymbol{\theta})$ are those that match the behavior of $p(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$, see more details later. We should choose $\widetilde{p}(\boldsymbol{\vartheta}_k)$ to have heavy tails so that it serves as an "envelope" to $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$.

Simulation from Complex Models (***)

Suppose that we have a model for dependent random variables $X_1, X_2, \ldots X_n$ that is written only in terms of the conditional density (or mass) functions

$$
\begin{gathered}
p(x_1 \,|\, x_2, \ldots, x_n) \\
p(x_2 \,|\, x_1, x_3, \ldots, x_n) \\
\vdots \\
p(x_i \,|\, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \\
\vdots \\
p(x_n \,|\, x_1, \ldots, x_{n-1}).
\end{gathered}
$$

This is called a *conditionally-specified model*, and such models are common, particularly in image analysis and spatial statistics. For example, Markov fields and their variations (hidden Markov models, Markov processes etc) are conditionally specified.

Our goal might be to draw samples from the joint pdf or pmf $p(x_1, \ldots, x_n|\boldsymbol{\theta})$, given that we know only the above conditional forms.

An Example of Conditionally-Specified Complex Models: Markov Random Field Models

Most people are familiar with the standard Markov assumption in time — given the entire past, the present depends only on the most immediate past. What does this imply about the joint distribution of variables in a one-dimensional random field? How would we extend this concept to two (or more) dimensions?

One-Dimensional Example: Consider a one-dimensional random field (process) $\{X_1, X_2, \ldots, X_N\}$ and denote

- $p(x) \equiv$ marginal density (or mass function) of X,
- $p(x_1, x_2) \equiv$ joint density (or mass function) of X_1 and X_2,
- $p(x_1|x_2) \equiv$ conditional density (or mass function) of X_1 given X_2.

The following is always true:

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2|x_1) \, \cdots p(x_n \,|\, x_1, x_2, \ldots, x_{N-1})$$

which becomes (using the Markov property)

$$p(x_1, x_2, \ldots, x_N) = p(x_1)\, p(x_2|x_1) \, \cdots p(x_N \,|\, x_{N-1}).$$

Note also that the Markov property implies that

$$p(x_i \,|\, \{x_j,\, j = 1, 2, \ldots, i-2\}) = p(x_i \,|\, x_{i-2}).$$

Now

$$\begin{aligned}
& p(x_i \,|\, \{x_j,\, j \neq i\}) \\
&= \frac{p(x_1, x_2, \ldots, x_N)}{p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)} \\
&= \frac{p(x_1)\, p(x_2|x_1)\, \cdots p(x_i \,|\, x_{i-1})}{p(x_1)\, p(x_2|x_1)\, \cdots p(x_{i-1} \,|\, x_{i-2})} \\
&\quad \cdot \frac{p(x_{i+1}|x_i)\, p(x_{i+2}|x_{i+1})\, \cdots p(x_N \,|\, x_{N-1})}{p(x_{i+1}|x_{i-1})\, p(x_{i+2}|x_{i+1})\, \cdots p(x_N \,|\, x_{N-1})} \\
&= \frac{p(x_i \,|\, x_{i-1})\, p(x_{i+1} \,|\, x_i)}{p(x_{i+1}|x_{i-1})} \\
&= \frac{p(x_i \,|\, x_{i-1})\, \overbrace{p(x_{i+1} \,|\, x_i, x_{i-1})}^{p(x_{i+1} \,|\, x_i)}}{p(x_{i+1}|x_{i-1})} \\
&= \underbrace{\frac{p(x_i, x_{i-1})}{p(x_{i-1})}}_{p(x_i \,|\, x_{i-1})} \cdot \underbrace{\frac{p(x_{i-1}, x_i, x_{i+1})}{p(x_{i-1}, x_i)}}_{p(x_{i+1} \,|\, x_i, x_{i-1})} \cdot \frac{1}{p(x_{i+1}|x_{i-1})} \\
&= \frac{p(x_{i-1}, x_i, x_{i+1})}{p(x_{i-1}, x_{i+1})} = p(x_i|x_{i-1}, x_{i+1})
\end{aligned}$$

which will become immediately clear once we master graphical models.

Thus, the typical Markov property in one dimension (e.g. time) implies that the conditional distribution of x_i given *all other* random-field variables (i.e. x_j, $j = 1, 2, \ldots, i-1, i+1, \ldots, N$ in the above example) depends only on the adjacent values x_{i-1} and x_{i+1}.

It is the structure of such *full conditional* distributions that are of concern in studying Markov random fields. Consider a collection of random observations $\{X(\boldsymbol{s}_i) : i = 1, 2, \ldots N\}$ where $\boldsymbol{s}_i$, $i = 1, 2, \ldots N$ denote (generally known) spatial locations at which these observations have been collected. The collection $\{X(\boldsymbol{s}_i) : i = 1, 2, \ldots N\}$ constitutes a Markov random field if, for each $i = 1, 2 \ldots, N$, the full conditional density or mass functions satisfy

$$p\big(x(\boldsymbol{s}_i) \,|\, \{x(\boldsymbol{s}_j) : j \neq i\}\big) = p\big(x(\boldsymbol{s}_i) \,|\, \{x(\boldsymbol{s}_j) : j \in \mathcal{N}_i\}\big) \quad (28)$$

where $\{\mathcal{N}_i : i = 1, 2, \ldots, N\}$ are neighborhoods. For example, we can choose $\mathcal{N}_i$ to be the set of measurements collected at locations within a specified distance from the ith location $\boldsymbol{s}_i$ (excluding $\boldsymbol{s}_i$).

Comments:

- Equation (28) *does not* imply that there is no spatial dependence between the random-field values at $\boldsymbol{s}_i$ and other locations that are outside its neighborhood; rather, it implies that, given the random-field values from its neighborhood, there is *no functional dependence* on the values at other locations outside the neighborhood.

- Neighborhoods must be symmetric, i.e. if $\boldsymbol{s}_j \in \mathcal{N}_i$ then $\boldsymbol{s}_i \in \mathcal{N}_j$.

A Common Framework for (*), (**), and (***)

The problem (*) is to sample from the joint posterior distribution

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\theta_1, \theta_2, \ldots, \theta_p \,|\, \boldsymbol{x}).$$

The problem (**) is to evaluate an integral

$$L(\boldsymbol{\lambda}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \, \frac{\widetilde{p}(\boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\theta})} \, d\boldsymbol{\theta}.$$

The problem (***) is to simulate data from a joint pdf/pmf

$$p(x_1, \ldots, x_n \,|\, \boldsymbol{\theta})$$

which is *unknown* but the full conditional pdfs/pmfs are *known*.

First, problems (*) and (**) are often similar since

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}) \, \pi(\boldsymbol{\theta})}{\underbrace{\int p(\boldsymbol{x}|\boldsymbol{\vartheta}) \, \pi(\boldsymbol{\vartheta}) \, d\boldsymbol{\vartheta}}_{\text{looks like } L(\boldsymbol{\lambda})}}$$

and the difficulty may arise from an unknown denominator.

The problems in (*) and (**) can sometimes be put into (or related to) the form (***). For example, in many Bayesian models, it may be impossible to obtain a closed-form expression for the posterior pdf/pmf $p(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, but is possible to get the full conditional posterior pdfs/pmfs:

$$p(\theta_k \,|\, \boldsymbol{x}, \{\theta_j : j \neq k\}), \quad k = 1, 2, \ldots, p.$$

Hence, in some cases, the problems (*), (**), and (***) may all reduce to generating samples from a joint distribution when all we know are the full conditional pdfs/pmfs. This is not always true, but is true often enough to make Gibbs sampling popular.

Simulation From a Joint Distribution Using Conditionals

To illustrate the basic idea consider a simple bivariate case: two random variables x_1 and x_2 with specified conditional pdfs/pmfs: $p_{x_1|x_2}(x_1|x_2)$ and $p_{x_2|x_1}(x_2|x_1)$. Given these conditionals, we wish to draw samples from the joint pdf/pmf $p(x_1, x_2)$.

If there exists a joint distribution $p(x_1, x_2)$, then the conditionals $p_{x_1|x_2}(x_1|x_2)$ and $p_{x_2|x_1}(x_2|x_1)$ are said to be *compatible*. Essentially, we need the existence of marginals $p_{x_1}(x_1)$ and $p_{x_2}(x_2)$ so that

$$p_{x_1|x_2}(x_1|x_2) = \frac{p(x_1, x_2)}{p_{x_2}(x_2)} \quad \text{and} \quad p_{x_2|x_1}(x_2|x_1) = \frac{p(x_1, x_2)}{p_{x_1}(x_1)}.$$

Assume that $p(x_1, x_2), p_{x_1}(x_1), p_{x_2}(x_2)$, $p_{x_1|x_2}(x_1|x_2), p_{x_2|x_1}(x_2|x_1)$ all exist, but all we know are the conditionals. If $p_{x_1|x_2}(x_1|x_2)$ and $p_{x_2|x_1}(x_2|x_1)$ are compatible and easy to simulate from, can we use them to generate samples from the joint pdf/pmf $p(x_1, x_2)$?

First, note that

$$
\begin{aligned}
p_{x_1}(x_1) &= \int \overbrace{p_{x_1|x_2}(x_1|x_2)\,p_{x_2}(x_2)}^{p(x_1,x_2)}\,dx_2 && (29)\\
p_{x_2}(x_2) &= \int p_{x_2|x_1}(x_2|x_1)\,p_{x_1}(x_1)\,dx_1. && (30)
\end{aligned}
$$

Now, substitute (30) into (29) to get

$$
\begin{aligned}
p_{x_1}(x_1) &= \int p_{x_1|x_2}(x_1|x_2) \int p_{x_2|x_1}(x_2|t)\,p_{x_1}(t)\,dt\,dx_2\\
&= \int \underbrace{\int p_{x_1|x_2}(x_1|x_2)\,p_{x_2|x_1}(x_2|t) dx_2}_{q(x_1,t)}\,p_{x_1}(t)\,dt\\
&= \int q(x_1,t)\,p_{x_1}(t)\,dt. && (31)
\end{aligned}
$$

Note:

- $q(\cdot,\cdot)$ is not necessarily a joint density or mass function.

To solve for p_{x_1}, consider (in general)

$$
\mathcal{T}(h,x) = \int q(x,t)\,h(t)\,dt
$$

an integral transformation of a function $h(\cdot)$ and a value x, where this transformation is defined by the function $q(x,t)$ which is determined entirely by the conditional specifications $p_{x_1|x_2}(\cdot|\cdot)$ and $p_{x_2|x_1}(\cdot|\cdot)$:

$$q(x,t) = \int p_{x_1|x_2}(x|x_2)\, p_{x_2|x_1}(x_2|t)\, dx_2.$$

Then, the marginal density $p_{x_1}(\cdot)$ is defined by a function that results in $\mathcal{T}$ being a fixed-point transformation [see (31)]:

$$p_{x_1}(x) = \mathcal{T}(p_{x_1}, x).$$

Now, consider a given value $x_1 \Longrightarrow$ we wish to evaluate the *unknown* $p_{x_1}(x_1)$. Theoretically, we could start with an initial guess, say $h^{(0)}(\cdot)$, for $p_{x_1}(\cdot)$ and compute

$$h^{(1)}(x_1) = \mathcal{T}(h^{(0)}, x_1) = \int q(x_1,t)\, h^{(0)}(t)\, dt$$

and repeat (i.e. apply successive substitution)

$$\begin{array}{rcl} h^{(2)}(x_1) & = & \mathcal{T}(h^{(1)}, x_1) \\ & \vdots & \\ h^{(m+1)}(x_1) & = & \mathcal{T}(h^{(m)}, x_1). \end{array}$$

When $h^{(m+1)}(x_1) = h^{(m)}(x_1)$ we have reached a fixed-point solution of the transformation and would take

$$p_{x_1}(x_1) = h^{(m+1)}(x_1) = h^{(m)}(x_1).$$

A drawback of successive substitution (to find fixed-point solutions to transformation problems) is that it does not guarantee a unique solution (the same is true for the EM algorithm). (Notice similarity with the EM algorithm, which is also a fixed-point method!)

Here, we simply assume a unique solution.

Finally, recall the discussion on p. 89 and note that

- Given $p_{x_1|x_2}(x_1|x_2)$ and $p_{x_2}(x_2)$, if we
  - draw one value x_2' from $p_{x_2}(x_2)$ and then
  - use that result to make one draw from $p_{x_1|x_2}(x_1|x_2')$, say x_1',

  then we have one draw (x_1', x_2') from the joint $p(x_1, x_2)$ and thus also from the marginals $p_{x_1}(x_1)$ and $p_{x_2}(x_2)$.

We are now ready to derive the Gibbs sampler for the bivariate case.

THE GIBBS SAMPLER

For a nice tutorial, see

G. Casella and E.I. George, "Explaining the Gibbs sampler," *American Statistician,* vol. 46, pp. 167–174, Aug. 1992

which is posted on WebCT.

The Bivariate Case: Gibbs Sampler as a Substitution Algorithm

Our goal is to draw samples from the joint distribution $p(x_1, x_2)$ *given only* its conditionals $p_{x_1|x_2}(x_1|x_2)$ and $p_{x_2|x_1}(x_2|x_1)$.

Assume

(i) $p_{x_1|x_2}(x_1 \,|\, x_2)$ and $p_{x_2|x_1}(x_2 \,|\, x_1)$ are compatible and

(ii) the fixed-point solution to $\mathcal{T}(h, x)$ is unique

where

$$\mathcal{T}(h, x) = \int q(x, t)\, h(t)\, dt$$

and

$$q(x, t) = \int p_{x_1|x_2}(x|x_2)\, p_{x_2|x_1}(x_2|t)\, dx_2.$$

[Here, we have two h functions since we wish to approximate the marginal distributions for both the marginals of x_1 and

x_2. Without loss of generality, we focus on the estimation of $p_{x_1}(x_1)$.] Then, an iterative algorithm (Gibbs sampler) to generate observations from the joint pdf/pmf $p(x_1, x_2)$ is developed as follows. Consider again the equations

$$\begin{aligned} p_{x_1}(x_1) &= \int p_{x_1|x_2}(x_1|x_2)\, p_{x_2}(x_2)\, dx_2 \\ p_{x_2}(x_2) &= \int p_{x_2|x_1}(x_2|x_1) p_{x_1}(x_1) dx_1. \end{aligned}$$

Suppose that we start with a value $x_2^{(0)}$ considered as a draw from some approximation to $p_{x_2}(\cdot)$, denoted by $\widehat{p}_{x_2}^{(0)}(\cdot)$. Given $x_2^{(0)}$, we generate a value $x_1^{(1)}$ from $p_{x_1|x_2}(x_1|x_2^{(0)})$. Thus, we have a draw from

$$\widehat{p}_{x_1}^{(1)}(x_1) = \int p_{x_1|x_2}(x_1|x_2)\, \widehat{p}_{x_2}^{(0)}(x_2)\, dx_2$$

i.e. we have

$$\begin{aligned} x_2^{(0)} &\quad \text{from} \quad \widehat{p}_{x_2}^{(0)}(\cdot) \\ x_1^{(1)} &\quad \text{from} \quad \widehat{p}_{x_1}^{(1)}(\cdot). \end{aligned}$$

Then, complete one “cycle” (in terms of the superscript index) by generating a value $x_2^{(1)}$ from $p_{x_2|x_1}(x_2|x_1^{(1)})$ which will be a value from

$$\widehat{p}_{x_2}^{(1)}(x_2) = \int p_2(x_2|x_1)\,\widehat{p}_{x_1}^{(1)}(x_1)\,dx_1.$$

At the end of one “cycle,” we have

$$\begin{aligned} x_1^{(1)} & \quad \text{from} \quad \widehat{p}_{x_1}^{(1)}(\cdot) \\ x_2^{(1)} & \quad \text{from} \quad \widehat{p}_{x_2}^{(1)}(\cdot). \end{aligned}$$

We start over with $x_2^{(1)}$ replacing $x_2^{(0)}$. Repeating the process k times, we end up with

$$\begin{aligned} x_1^{(k)} & \quad \text{from} \quad \widehat{p}_{x_1}^{(k)}(\cdot) \\ x_2^{(k)} & \quad \text{from} \quad \widehat{p}_{x_2}^{(k)}(\cdot). \end{aligned}$$

Now,

$$\begin{aligned} \widehat{p}_{x_1}^{(k)}(x_1) & \quad \to \quad p_{x_1}(x_1) \\ \widehat{p}_{x_2}^{(k)}(x_2) & \quad \to \quad p_{x_2}(x_2) \quad \text{as } k \to \infty. \end{aligned}$$

Why? Because at each step we have used the transformations

$$\begin{aligned}
\widehat{p}_{x_1}^{(k+1)}(x_1) &= \int p_{x_1|x_2}(x_1|x_2) \underbrace{\int p_{x_2|x_1}(x_2|t)\widehat{p}_{x_1}^{(k)}(t)\,dt}_{\widehat{p}_{x_2}^{(k)}(x_2)} dx_2 \\
&= \int\int p_{x_1|x_2}(x_1|x_2)\, p_{x_2|x_1}(x_2|t)\, \widehat{p}_{x_1}^{(k)}(t)\,dt\,dx_2 \\
&= \int q(x_1, t)\, \widehat{p}_{x_1}^{(k)}(t)\,dt = \mathcal{T}(\widehat{p}_{x_1}^{(k)}, x_1).
\end{aligned}$$

Similarly,

$$\widehat{p}_{x_2}^{(k+1)}(x_2) = \mathcal{T}(\widehat{p}_{x_2}^{(k)}, x_2)$$

(for which we need a different h function). At each step, we have used the transformations whose fixed-point solutions give $p_{x_1}(x_1)$ and $p_{x_2}(x_2)$ and we have used these in a successive-substitution manner.

Thus, if we apply the above substitution a large number of times K, we will end up with *one* pair of values (x_1', x_2') generated from $p(x_1, x_2)$. Consequently, x_1' is a draw from $p_{x_1}(x_1)$ and x_1' a draw from $p_{x_2}(x_2)$.

If we repeat the entire procedure N times (with different starting values), we end up with N i.i.d. pairs $(x_{1,i}, x_{2,i}), i = 1, \ldots, N$.

Note: Although we have independence between pairs, we will not (and should not, in general) have independence within a pair (i.e. if we have independence within a pair, we do not need Gibbs sampling).

Thus, the joint empirical distribution of $(x_{1,i}, x_{2,i})$ (i.e. *histogram*) should converge (as $N \to \infty$) to a joint distribution having marginals $p_{x_1}(\cdot)$ and $p_{x_2}(\cdot)$ and conditionals $p_{x_1|x_2}(\cdot|\cdot)$ and $p_{x_2|x_1}(\cdot|\cdot)$.

Summary of the Gibbs Sampling Algorithm: To obtain one observation [i.e. one pair (x_1, x_2) from the joint distribution $p(x_1, x_2)$]:

(i) Start with initial value $x_2^{(0)}$;

(ii) Generate $x_1^{(1)}$ from $p_{x_1|x_2}(x_1|x_2^{(0)})$;

(iii) Generate $x_2^{(1)}$ from $p_{x_2|x_1}(x_2|x_1^{(1)})$;

(iv) Return to step (ii) with $x_2^{(1)}$ in place of $x_2^{(0)}$;

(v) Repeat a large number of times.

Comments:

(i) To generate N i.i.d. samples, apply the algorithm N times with different starting values:

$$x_2^{(0,1)}, x_2^{(0,2)}, \ldots x_2^{(0,N)}.$$

(ii) To generate N dependent (but, under mild conditions, ergodic) samples, use the algorithm with one starting value $x_2^{(0)}$ and keep all values after K cycles:

$$\{(x_1^{(K+1)}, x_2^{(K+1)}), \ldots, (x_1^{(K+N)}, x_2^{(K+N)})\}.$$

Under ergodic theorems, this provides good results for a histogram (empirical distribution) of $p_{x_1}(\cdot)$, $p_{x_2}(\cdot)$, and $p(x_1, x_2)$ and Monte Carlo estimation of various quantities.

(iii) The sequence in comment (ii) forms a Markov Chain; hence, Gibbs sampling is one type of Markov Chain Monte Carlo (MCMC) methods.

Gibbs Sampling in Higher Dimensions

Gibbs sampling in p dimensions ($p > 2$) is a direct extension of the bivariate algorithm. Suppose that we wish to sample from a joint distribution $p(x_1, \ldots, x_p)$ but have available only the conditionals

$$p_{x_i|\{x_k : k \neq i\}}(x_i|\{x_k : k \neq i\}), \quad i = 1, \ldots, p.$$

Note that these are full conditional pdfs/pmfs. A Gibbs sampler for generating one sample from $p(x_1, \ldots, x_p)$ is

(i) Obtain starting values $x_1^{(0)}, \ldots, x_{p-1}^{(0)}$;

(ii) Draw $x_p^{(1)}$ from $p_{x_p|x_1,x_2,\ldots,x_{p-1}}(x_p|x_1^{(0)}, \ldots, x_{p-1}^{(0)})$;

(iii) Draw $x_{p-1}^{(1)}$ from

$$p_{x_{p-1}|x_1,x_2,\ldots,x_p}(x_{p-1}|x_1^{(0)}, \ldots, x_{p-2}^{(0)}, x_p^{(1)});$$

(iv) Draw $x_{p-2}^{(1)}$ from

$$p_{x_{p-2}|x_1,x_2,\ldots,x_p}(x_{p-2}|x_1^{(0)}, \ldots, x_{p-3}^{(0)}, x_{p-1}^{(1)}, x_p^{(1)});$$

(v) Continue sequentially until one full cycle results in

$$\boldsymbol{x}^{(1)} = [x_1^{(1)}, \ldots, x_p^{(1)}]^T;$$

(vi) Return to step (ii) with $\boldsymbol{x}^{(0)}$ replaced by $\boldsymbol{x}^{(1)}$;

(vii) Repeat steps (ii)–(vi) a large number of times K.

We then have one observation $\boldsymbol{x}^{(k)} = [x_1^{(k)}, \ldots, x_p^{(k)}]^T$ from $p(x_1, \ldots, x_p)$.

Comments:

- Again, we may wish to either repeat this entire process using different starting values (i.i.d. samples) or depend (rely) on ergodicity and keep all values after a given point.

- If we rely on ergodicity, a major issue is how large k should be so that $\boldsymbol{x}^{(k)}$ may be reliably considered as a value from the joint distribution.

  $*$ In MCMC terminology, the value of K after which $\boldsymbol{x}^{(k)}$, $k \geq K$ are "reliable" defines the "burn-in" period.

- *Compromise between "pure" i.i.d. and ergodic samples*: Compute the sample autocorrelation of $\boldsymbol{x}^{(k)}$. If this autocorrelation at lag l and larger is negligible, then we can generate *almost i.i.d.* samples by keeping only every lth draw in the chain after convergence.

- Clearly, we can also cycle $x_1^{(t+1)} \longrightarrow x_2^{(t+1)} \longrightarrow \cdots \longrightarrow x_{p-1}^{(t+1)} \longrightarrow x_p^{(t+1)}$, say, rather than $x_p^{(t+1)} \longrightarrow x_{p-1}^{(t+1)} \longrightarrow \cdots \longrightarrow x_2^{(t+1)} \longrightarrow x_1^{(t+1)}$, or pick up any other cycling order. This is a trivial observation.

A Toy Example: Drawing Samples from a Bivariate Gaussian Pdf

Suppose that we wish to sample from

$$
\begin{aligned}
p_{X,Y}(x,y) = \mathcal{N}\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \Big) \\
\propto \quad \exp\Big\{ -\frac{1}{2(1-\rho^2)} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Big\} \\
\propto \quad \exp\Big[-\frac{1}{2(1-\rho^2)} \cdot (x^2 + y^2 - 2\,\rho\,x\,y) \Big]
\end{aligned}
$$

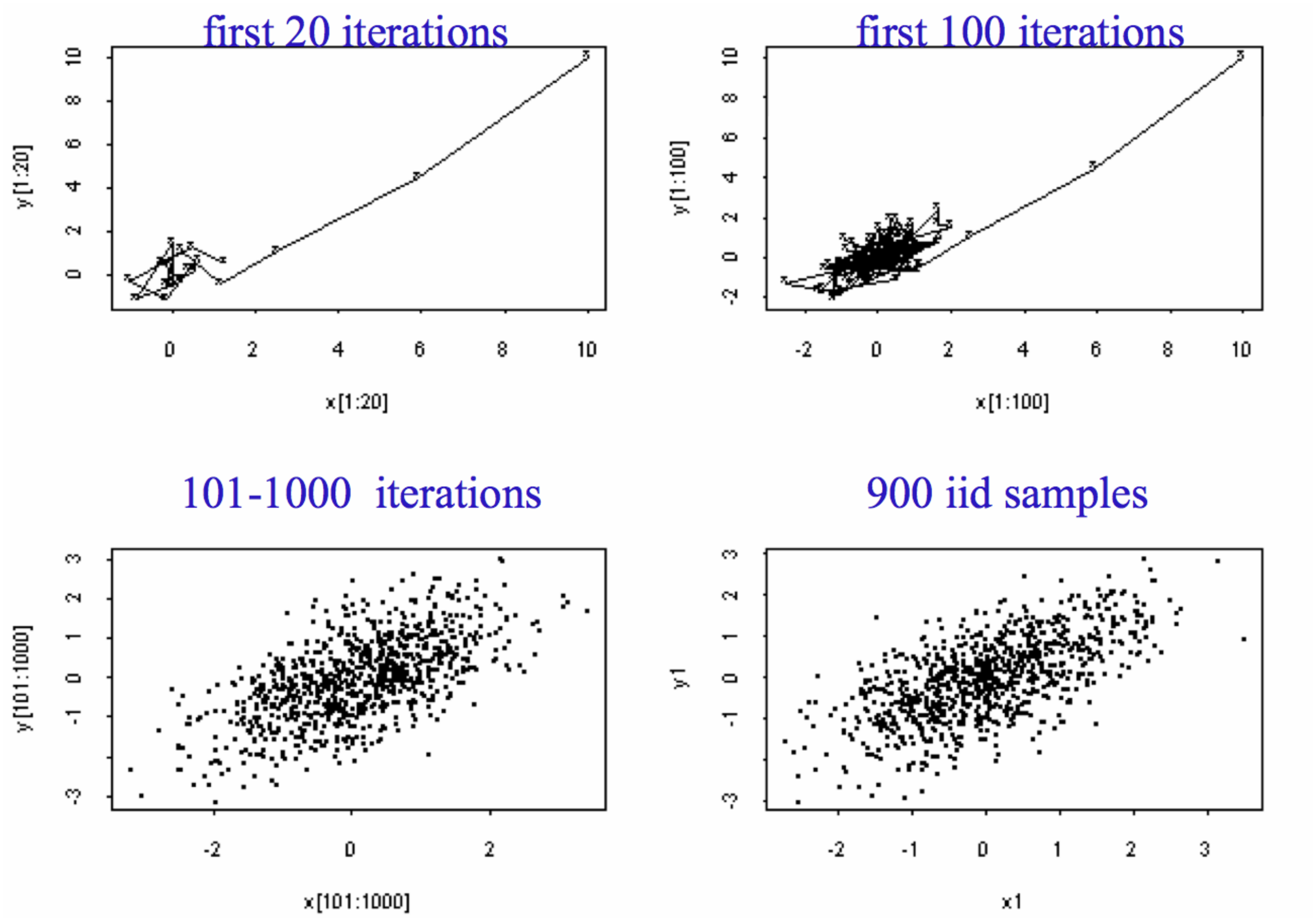
using the Gibbs sampler, see also p. 27 in handout # 1 for a general pdf expression for a bivariate Gaussian pdf. We need conditional pdfs $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$, which easily follow:

$$
\begin{aligned}
p_{X|Y}(x\,|\,y) \quad & \propto \quad \exp\Big[-\frac{1}{2(1-\rho^2)} \cdot (x^2 - 2\,\rho\,y\,x) \Big] \\
& = \quad \mathcal{N}(\rho\,y, 1-\rho^2)
\end{aligned}
$$

and, by symmetry,

$$
p_{Y|X}(y\,|\,x) \quad = \quad \mathcal{N}(\rho\,x, 1-\rho^2).
$$

Let us start from, say, $(x_0, y_0) = (10, 10)$ and run a few iterations for $\rho = 0.6$.



Why might Gibbs work (as it apparently does in this example)?

A fixed-point argument (certainly not a proof):

$$\int p_X(x)\, p_{Y\,|\,X}(y|x)\, dx \quad = \quad p_Y(y)$$

$$\int\int p_X(x)\, p_{Y\,|\,X}(y\,|\,x)\, p_{X\,|\,Y}(x'\,|\,y)\, dx\, dy \quad = \quad p_X(x').$$

Remember: Gibbs does not always work; later, we will give an example where Gibbs *does not* work.

Bayesian Example: Gibbs Sampler for the Gaussian Semi-Conjugate Model

Simple DC-level-in-AWGN measurement model:

$$x_i \,|\, \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$$

with parameters μ and σ^2. Clearly, given μ and σ^2, x_i, $i = 1, 2, \ldots, N$ are i.i.d.

Consider the semi-conjugate prior:

$$\begin{aligned}\pi(\mu) &= \mathcal{N}(\mu_0, \tau_0^2)\\ \pi(\sigma^2) &= \text{Inv-}\chi^2(\nu_0, \sigma_0^2).\end{aligned}$$

The joint posterior distribution is

$$\begin{aligned}p(\mu, \sigma^2 | \boldsymbol{x}) \quad \propto \quad & \exp\Big[-\frac{1}{2\tau_0^2}\,(\mu - \mu_0)^2\Big]\\ & \cdot(\sigma^2)^{-(\nu_0/2+1)} \cdot \exp\Big(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\Big)\\ & \cdot(\sigma^2)^{-N/2} \cdot \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\Big].\end{aligned}$$

To implement the Gibbs sampler for simulating from this joint posterior pdf, we need the full conditional pdfs: $p(\mu \,|\, \sigma^2, \boldsymbol{x})$ and $p(\sigma^2 \,|\, \mu, \boldsymbol{x})$.

The full conditional posterior pdf for μ is

$$
\begin{aligned}
& p(\mu \,|\, \sigma^2, \boldsymbol{x}) \overset{\text{sufficiency}}{=} p(\mu \,|\, \sigma^2, \overline{x}) \propto p(\mu, \overline{x} \,|\, \sigma^2) \\
& \propto \underbrace{p(\overline{x} \,|\, \mu, \sigma^2)}_{\mathcal{N}(\mu, \frac{\sigma^2}{N})} \cdot \; \underbrace{\pi(\mu)}_{\mathcal{N}(\mu_0, \tau_0^2)} \\
& \propto \exp\Big\{ -\frac{1}{2} \Big[\frac{N}{\sigma^2} (\overline{x} - \mu)^2 + \frac{1}{\tau_0^2} (\mu - \mu_0)^2 \Big] \Big\} \\
& \propto \exp\Big\{ -\frac{1}{2} \Big[\frac{N\tau_0^2 + \sigma^2}{\sigma^2 \tau_0^2} \mu^2 - 2 \frac{N\tau_0^2 \overline{x} + \sigma^2 \mu_0}{\sigma^2 \tau_0^2} \mu \Big] \Big\} \\
& \propto \exp\Big\{ -\frac{1}{2} \frac{N\tau_0^2 + \sigma^2}{\sigma^2 \tau_0^2} \Big[\mu^2 - 2 \frac{N\tau_0^2 \overline{x} + \sigma^2 \mu_0}{N\tau_0^2 + \sigma^2} \mu \Big] \Big\} \\
& \propto \mathcal{N}(\mu_N, \tau_N^2)
\end{aligned}
$$

where

$$
\begin{aligned}
\mu_N &= \frac{\frac{N}{\sigma^2} \overline{x} + \frac{1}{\tau_0^2} \mu_0}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}} \\
\tau_N^2 &= \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\tau_0^2}}.
\end{aligned}
$$

The full conditional posterior pdf for σ^2 is

$$\begin{aligned} p(\sigma^2|\mu, \boldsymbol{x}) &\propto (\sigma^2)^{-(\frac{N+\nu_0}{2}+1)} \\ &\quad \cdot \exp\Big\{ -\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{N}(x_i-\mu)^2 + \nu_0\sigma_0^2\Big]\Big\} \\ &\propto \text{Inv-}\chi^2(\nu_N, \sigma_N^2) \end{aligned}$$

where

$$\begin{aligned} \nu_N &= \nu_0 + N \\ \sigma_N^2(\mu) &= \frac{1}{\nu_N}\cdot\Big[\sum_{i=1}^{N}(x_i-\mu)^2 + \nu_0\sigma_0^2\Big]. \end{aligned}$$

Now, we are ready to describe the Gibbs sampler for simulating from $\mu, \sigma^2|\boldsymbol{x}$ under the above model:

(a) start with a guess for σ^2, $(\sigma^2)^{(0)}$;

(b) draw $\mu^{(1)}$ from a Gaussian distribution $\mathcal{N}\big(\mu_N((\sigma^2)^{(0)}), \tau_N((\sigma^2)^{(0)})\big)$, where

$$\begin{aligned} \mu_N((\sigma^2)^{(0)}) &= \frac{\frac{N}{(\sigma^2)^{(0)}}\overline{x} + \frac{1}{\tau_0^2}\mu_0}{\frac{N}{(\sigma^2)^{(0)}} + \frac{1}{\tau_0^2}} \\ \tau_N^2((\sigma^2)^{(0)}) &= \frac{1}{\frac{N}{(\sigma^2)^{(0)}} + \frac{1}{\tau_0^2}}; \end{aligned}$$

(c) next, draw $(\sigma^2)^{(1)}$ from Inv-$\chi^2(\nu_N, \sigma_N^2(\mu^{(1)}))$, where

$$\sigma_N^2(\mu^{(1)}) = \frac{1}{\nu_N} \cdot \Big[\sum_{i=1}^{N} (x_i - \mu^{(1)})^2 + \nu_0 \sigma_0^2 \Big];$$

(d) iterate between (b) and (c) upon convergence.

Grouping and Collapsing

Suppose that it is possible to sample from the conditional distribution

$$p_{x_{p-1},x_p\,|\,x_2,x_3,\ldots,x_{p-2}}(x_{p-1},x_p\,|\,x_2,x_3,\ldots,x_{p-2})$$

directly. Then we can use grouped Gibbs, described below.

Suppose that it is possible to integrate out x_p and that we can sample from the conditional distributions

$$p_{x_i\,|\,x_1,x_2,\ldots,x_{p-2}}(x_i\,|\,x_1,x_2,\ldots,x_{p-2}), \quad i=1,2,\ldots,p-1$$

directly. Then we can use collapsed Gibbs, described below.

Consider now the following three schemes:

Standard Gibbs: $x_1^{(t+1)} \to x_2^{(t+1)} \to \cdots \to x_{p-1}^{(t+1)} \to x_p^{(t+1)}$.

Grouped Gibbs: $x_1^{(t+1)} \to x_2^{(t+1)} \to \cdots \to x_{p-2}^{(t+1)} \to (x_{p-1}^{(t+1)}, x_p^{(t+1)})$ where the last draw is from

$$p_{x_{p-1},x_p\,|\,x_2,x_3,\ldots,x_{p-2}}(x_{p-1},x_p\,|\,x_2^{(t+1)},x_3^{(t+1)},\ldots,x_{p-2}^{(t+1)})$$

which we assumed we know how to do. Here is one situation where this can happen. Suppose that it is possible to integrate out x_p and that we can sample from the conditional distribution

$$p_{x_{p-1}\,|\,x_1,x_2,\ldots,x_{p-2}}(x_{p-1}\,|\,x_1,x_2,\ldots,x_{p-2})$$

directly. Then, grouped Gibbs can be implemented as:

$x_1^{(t+1)} \rightarrow x_2^{(t+1)} \rightarrow \cdots \rightarrow x_{p-3}^{(t+1)} \rightarrow x_{p-2}^{(t+1)}$ using full conditionals; then sample $x_{p-1}^{(t+1)}$ from $p_{x_{p-1}\,|\,x_1,x_2,\ldots,x_{p-2}}(x_{p-1}\,|\,x_1^{(t+1)},x_2^{(t+1)},\ldots,x_{p-2}^{(t+1)})$, followed by drawing $x_p^{(t+1)}$ from the full conditional $p_{x_p\,|\,x_1,x_2,\ldots,x_{p-1}}(x_p\,|\,x_1^{(t+1)},x_2^{(t+1)},\ldots,x_{p-1}^{(t+1)})$.

Collapsed Gibbs: $x_1^{(t+1)} \rightarrow x_2^{(t+1)} \rightarrow \cdots \rightarrow x_{p-3}^{(t+1)} \rightarrow x_{p-1}^{(t+1)}$ using

$$p_{x_i\,|\,x_1,x_2,\ldots,x_{p-2}}(x_i\,|\,x_1,x_2,\ldots,x_{p-2}), \quad i=1,2,\ldots,p-1.$$

It turns out that there is an ordering $\succeq$ (better than or as good as) for these schemes in terms of their "convergence speed":

Collapsed Gibbs $\succeq$ Grouped Gibbs $\succeq$ Standard Gibbs

see Ch. 6 in

J.S. Liu, *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag, 2001.

Rao-Blackwellization

Suppose that we ran a two (block) component Gibbs sampler and that we have (post burn-in) samples

$$(x_1^{(t)}, x_2^{(t)}), \quad t = 1, 2, \ldots, T$$

from $p(x_1, x_2)$.

Say we are interested in estimating

$$\begin{aligned} G &= E_{p(x_1,x_2)}[g(X_1)] = \int\int g(x_1)\, p(x_1, x_2)\, dx_1\, dx_2 \\ &= \int g(x_1)\, p(x_1)\, dx_1 \end{aligned}$$

and suppose that $E[g(X_1) \,|\, X_2 = x_2]$ can be computed analytically. Now, consider the following two competing estimators of G:

$$\underbrace{\widehat{G} = \frac{1}{T} \cdot \sum_{t=1}^{T} g(x_1^{(t)})}_{\text{simple MC average estimator}} \quad , \quad \widetilde{G} = \underbrace{\frac{1}{T} \cdot \sum_{t=1}^{T} \mathrm{E}\,[g(X_1 \,|\, x_2^{(t)})]}_{\text{Rao-Blackwellized estimator}} \ .$$

It can be shown that, as expected from conditioning,

$$\operatorname{var}(\widetilde{G}) \leq \operatorname{var}(\widehat{G}).$$

Say we wish to estimate the marginal posterior pdf $p(x_1)$ assuming that we can compute $p(x_1 \,|\, x_2)$ analytically. Then, apply the following mixture pdf estimator:

$$\widehat{p}(x_1) = \frac{1}{T} \cdot \sum_{t=1}^{T} p(x_1 \,|\, x_2^{(t)}).$$