

Bayesian Inference II

Outline:

- Multiparameter models.

Reading: Chapter 10 in Kay-I.

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.

For example, classical DC-signal-in-AWGN-noise model: $X[n]$ are conditionally independent, identically distributed (i.i.d.) $\mathcal{N}(\mu, \sigma^2)$ given μ and σ^2 , where *both* μ and σ^2 are *unknown* parameters. 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 the value of σ^2 , but 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 would be the DC-signal-in-AWGN-noise model, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

We should base our inference on

$$f_{\Theta_1 | \mathbf{X}}(\theta_1 | \mathbf{x})$$

the marginal posterior pdf (pmf) of θ_1 , which accounts for the uncertainty due to the fact that θ_2 is unknown. First, we start with the joint posterior pdf (pmf):

$$f_{\Theta_1, \Theta_2 | \mathbf{x}}(\theta_1, \theta_2 | \mathbf{x}) \propto f_{\mathbf{x} | \Theta_1, \Theta_2}(\mathbf{x} | \theta_1, \theta_2) \pi(\theta_1, \theta_2)$$

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

$$f_{\Theta_1 | \mathbf{x}}(\theta_1 | \mathbf{x}) = \int f_{\Theta_1, \Theta_2 | \mathbf{x}}(\theta_1, \theta_2 | \mathbf{x}) d\theta_2$$

or, equivalently,

$$f_{\Theta_1 | \mathbf{x}}(\theta_1 | \mathbf{x}) = \int f_{\Theta_1 | \Theta_2, \mathbf{x}}(\theta_1 | \theta_2, \mathbf{x}) f_{\Theta_2 | \mathbf{x}}(\theta_2 | \mathbf{x}) d\theta_2$$

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.

Predictive Distribution: Suppose that we wish to predict an observation X_* coming from the model

$$f_{\mathbf{x} | \Theta_1, \Theta_2}(\mathbf{x} | \theta_1, \theta_2).$$

If X_* and \mathbf{x} are conditionally independent given θ_1 and θ_2 , then, based on (10) from handout # 4, we obtain:

$$\begin{aligned} & f_{X_* | \mathbf{x}}(x_* | \mathbf{x}) \\ &= \int \int f_{X_* | \Theta_1, \Theta_2}(x_* | \theta_1, \theta_2) \cdot f_{\Theta_1, \Theta_2 | \mathbf{x}}(\theta_1, \theta_2 | \mathbf{x}) d\theta_1 d\theta_2 \\ &= \int \int f_{X_* | \Theta_1, \Theta_2}(x_* | \theta_1, \theta_2) \cdot f_{\Theta_1 | \Theta_2, \mathbf{x}}(\theta_1 | \theta_2, \mathbf{x}) \\ &\quad \cdot f_{\Theta_2 | \mathbf{x}}(\theta_2 | \mathbf{x}) d\theta_1 d\theta_2. \end{aligned}$$

The above integrals are often difficult to evaluate analytically and may be highly multidimensional if θ_2 and θ_1 are vectors. Typically, we need Monte Carlo methods to handle practical cases. If we just need to find the mode of

$$f_{\Theta_1 | \mathbf{x}}(\theta_1 | \mathbf{x})$$

an EM algorithm may suffice. The EM algorithm will be discussed in detail later in this handout.

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 a couple of (rare) classical practical cases where analytical Bayesian computations are possible.

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

See also Section 3.2 in

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, 2nd 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:

$$\begin{aligned}\pi(\mu, \sigma^2) &= \pi(\mu) \pi(\sigma^2) \propto \mathbf{1} \frac{1}{\sigma^2} i_{[0, +\infty)}(\sigma^2) \\ &= \frac{1}{\sigma^2} i_{[0, +\infty)}(\sigma^2).\end{aligned}\tag{1}$$

DC level in additive white Gaussian noise: $X[n]$, $n = 0, 1, \dots, N - 1$ are conditionally i.i.d. $\mathcal{N}(\mu, \sigma^2)$ given $\boldsymbol{\theta}$, where

$$\boldsymbol{\theta} = [\mu, \sigma^2]^T.$$

Suitably arranging the terms in the expression for the likelihood

function $f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ leads to

$$\begin{aligned}
 f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \\
 &\cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}[(x[n]-\bar{x})+(\bar{x}-\mu)]^2\right\} i_{[0,+\infty)}(\sigma^2) \\
 &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{N s^2(\mathbf{x}) + N(\bar{x}-\mu)^2}{2\sigma^2}\right\} i_{[0,+\infty)}(\sigma^2) \quad (2)
 \end{aligned}$$

where

$$s^2(\mathbf{x}) = s^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2, \quad \bar{x} = \frac{1}{N}\sum_{n=0}^{N-1}x[n] \quad (3)$$

are the sufficient statistics for μ and σ^2 . The product of the above prior pdf (1) and likelihood (2) is proportional to the posterior pdf:

$$\begin{aligned}
 f_{\boldsymbol{\theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x}) &= f_{\mu,\Sigma^2|\mathbf{X}}(\mu,\sigma^2|\mathbf{x}) \\
 &\propto \frac{1}{\sigma^2}(\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}[N s^2 + N(\bar{x}-\mu)^2]\right\} i_{[0,+\infty)}(\sigma^2). \quad (4)
 \end{aligned}$$

Provided that $N \geq 2$, the posterior pdf (4) is proper.

Conditional posterior pdf of μ given $\Sigma^2 = \sigma^2$

The conditional posterior pdf of μ given $\Sigma^2 = \sigma^2$ is proportional to the joint posterior density with σ^2 held constant:

$$\begin{aligned} f_{\mu | \Sigma^2, \mathbf{x}}(\mu | \sigma^2, \mathbf{x}) &\propto f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) \\ &\stackrel{\text{see (4)}}{\propto} \frac{1}{\sigma^2} \cdot (\sigma^2)^{-N/2} \cdot \exp \left\{ -\frac{1}{2\sigma^2} [N s^2 + N(\bar{x} - \mu)^2] \right\} \\ &\underbrace{\propto}_{\text{keep track of } \mu} \exp \left\{ -\frac{N}{2\sigma^2} (\bar{x} - \mu)^2 \right\} \\ &\text{(look up the table of distributions)} \\ &\text{is the kernel of } \mathcal{N}\left(\mu \mid \bar{x}, \frac{\sigma^2}{N}\right) \end{aligned} \quad (5)$$

which agrees with (16) in handout # 4, the case of estimating the DC level in AWGN noise having known variance.

We now write $f_{\mu | \Sigma^2, \mathbf{x}}(\mu | \sigma^2, \mathbf{x})$ *with the normalizing constant*:

$$f_{\mu | \Sigma^2, \mathbf{x}}(\mu | \sigma^2, \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2/N}} \cdot \exp \left[-\frac{(\mu - \bar{x})^2}{2\sigma^2/N} \right] \quad (6)$$

see the table of distributions.

A reminder:

$$\text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right) i_{[0,+\infty)}(\sigma^2) \quad (7)$$

and the full scaled inverted χ^2 pdf is

$$\begin{aligned} \text{Inv-}\chi^2(\sigma^2 | \nu_0, \sigma_0^2) &= \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} (\sigma_0^2)^{\nu_0/2} \\ &\cdot (\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right) i_{[0,+\infty)}(\sigma^2) \end{aligned} \quad (8)$$

see the table of distributions.

Conditional posterior pdf of $\Sigma^2 = \sigma^2$ given μ

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

$$f_{\Sigma^2 | \mu, \mathbf{x}}(\sigma^2 | \mu, \mathbf{x}) \propto f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x})$$

$$\stackrel{\text{see (4)}}{\propto} \frac{1}{\sigma^2} \cdot (\sigma^2)^{-N/2} \cdot \exp\left\{-\frac{1}{2\sigma^2} [N s^2 + N(\bar{x} - \mu)^2]\right\}$$

$\underbrace{\propto}_{\text{keep track of } \sigma^2}$

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

(look up the table
of distributions)

$$\text{is the kernel of } \text{Inv-}\chi^2(\sigma^2 | N, s^2 + (\bar{x} - \mu)^2). \quad (9)$$

Therefore,

$$f_{\Sigma^2 | \mu, \mathbf{x}}(\sigma^2 | \mu, \mathbf{x}) = \frac{(N/2)^{N/2}}{\Gamma(N/2)} [s^2 + (\bar{x} - \mu)^2]^{N/2} \cdot (\sigma^2)^{-(N/2+1)} \exp\left(-\frac{N s^2 + N (\bar{x} - \mu)^2}{2 \sigma^2}\right) i_{[0, +\infty)}(\sigma^2). \quad (10)$$

Marginal Posterior Pdf of Σ^2

The marginal posterior pdf of Σ^2 is a scaled inverted χ^2 :

$$f_{\Sigma^2 | \mathbf{X}}(\sigma^2 | \mathbf{x}) = \text{Inv-}\chi^2(N - 1, s^2) \quad (11)$$

i.e. $\{\Sigma^2 | \mathbf{X} = \mathbf{x}\}$ can be simulated as follows:

$$\underbrace{\{\Sigma^2 | \mathbf{X} = \mathbf{x}\}}_{\text{coming from } f_{\Sigma^2 | \mathbf{X}}(\sigma^2 | \mathbf{x})} \sim \frac{\sum_{n=0}^{N-1} (x[n] - \bar{x})^2}{Z}$$

where Z is a χ_{N-1}^2 random variable, see (3) and p. 33 in handout # 4.

To derive this result, we apply a Bayesian trick for integrating μ out without actually performing the integration. The key to this trick is that the conditional posterior pdf $f_{\mu | \Sigma^2, \mathbf{X}}(\mu | \sigma^2, \mathbf{x})$ is known exactly (*including the normalizing constant*), e.g. it belongs to a family of pdfs (pmfs) that appear in our table of distributions.

We derive (11):

$$\underbrace{f_{\Sigma^2 | \mathbf{X}}(\sigma^2 | \mathbf{x})}_{\text{not a function of } \mu} = \frac{\overbrace{f_{\mu, \Sigma^2 | \mathbf{X}}(\mu, \sigma^2 | \mathbf{x})}^{\text{see (4)}}}{\underbrace{f_{\mu | \Sigma^2, \mathbf{X}}(\mu | \sigma^2, \mathbf{x})}_{\mathcal{N}(\mu | \bar{x}, \frac{\sigma^2}{N}), \text{ see (6)}}}$$

\propto
keep track of both σ^2 and
 μ , **be careful with (6)**

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

plug in $\underline{\mu} = \bar{x}$ $\frac{(\sigma^2)^{-N/2-1} \cdot \exp \left(-\frac{1}{2\sigma^2} N s^2 \right)}{(\sigma^2)^{-1/2}}$

rearrange to look like (7) $= (\sigma^2)^{-[(N-1)/2-1]} \cdot \exp \left(-\frac{N s^2}{2\sigma^2} \right)$

the kernel of $= \text{Inv-}\chi^2(\sigma^2 | N - 1, \frac{N}{N-1} s^2)$

$$= \text{Inv-}\chi^2 \left(\sigma^2 | N - 1, \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \right)$$

Additional checking: μ must cancel out.

We computed

$$f_{\Sigma^2 | \mathbf{x}}(\sigma^2 | \mathbf{x}) = \int_{-\infty}^{+\infty} f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) d\mu$$

algebraically, without performing the actual integration.

Marginal Posterior Pdf of μ

Apply the Bayesian trick and integrate Σ^2 out:

$$\underbrace{f_{\mu | \mathbf{x}}(\mu | \mathbf{x})}_{\text{not a function of } \sigma^2} = \frac{\overbrace{f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x})}^{\text{see (4)}}}{\underbrace{f_{\Sigma^2 | \mu, \mathbf{x}}(\mu | \sigma^2, \mathbf{x})}_{\text{Inv-}\chi^2(\sigma^2 | N, s^2 + (\bar{x} - \mu)^2), \text{ see (10)}}}$$

\propto
 keep track of both μ
 σ^2 , be careful with (10)

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

$$\propto [s^2 + (\bar{x} - \mu)^2]^{-N/2}$$

$$\propto \left[1 + \frac{1}{N-1} \cdot \frac{(\mu - \bar{x})^2}{s^2/(N-1)} \right]^{-(N-1+1)/2}$$

which is kernel of the scaled t distribution. Additional checking: σ^2 must cancel out. From the table of distributions, we find:

$$f_{\mu | \mathbf{x}}(\mu | \mathbf{x}) = t_{N-1} \left(\mu \mid \bar{x}, \frac{s^2}{N-1} \right)$$

which is the pdf of the t *distribution* with $N - 1$ degrees of

freedom and parameters

$$\bar{x} \quad \text{mean, see (3)} \quad \text{and} \quad \frac{s^2}{N-1} = \frac{\sum_{n=0}^{N-1} (x[n] - \bar{x})^2}{N(N-1)} \quad \text{scale.}$$

We have managed to compute

$$f_{\mu | \mathbf{x}}(\mu | \mathbf{x}) = \int_{-\infty}^{+\infty} f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) d\sigma^2$$

algebraically, without performing the actual integration.

How To Summarize the Obtained Posterior Distributions?

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

We can make “interval inferences” based on the posterior distributions and construct *credible sets*, also known as *Bayesian confidence intervals*.

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

$$\Pr_{\Theta | \mathbf{x}}\{\theta \in A | \mathbf{x}\} = c.$$

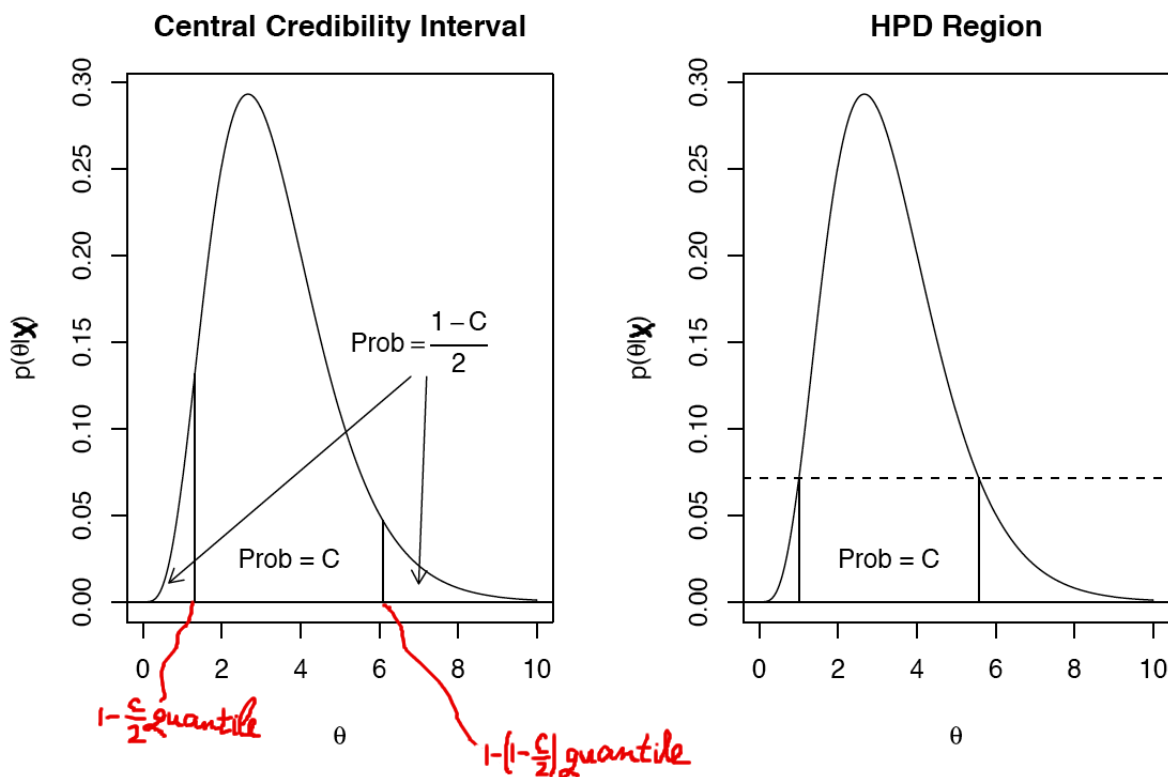
The most common approach to credible intervals (scalar credible sets) are *central credible intervals*. A central credible interval $[\tau_l, \tau_u]$ satisfies

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

where τ_l and τ_u are the $\frac{1-c}{2}$ and $1 - \frac{1-c}{2}$ quantiles of the posterior pdf, see also the figure in Example 1 on the next

page. An alternative is the $100c\%$ *highest posterior density* (HPD) region, which is defined as the smallest region of the parameter space with probability c .

Example 1:



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, since it only involves finding quantiles of the posterior distribution.

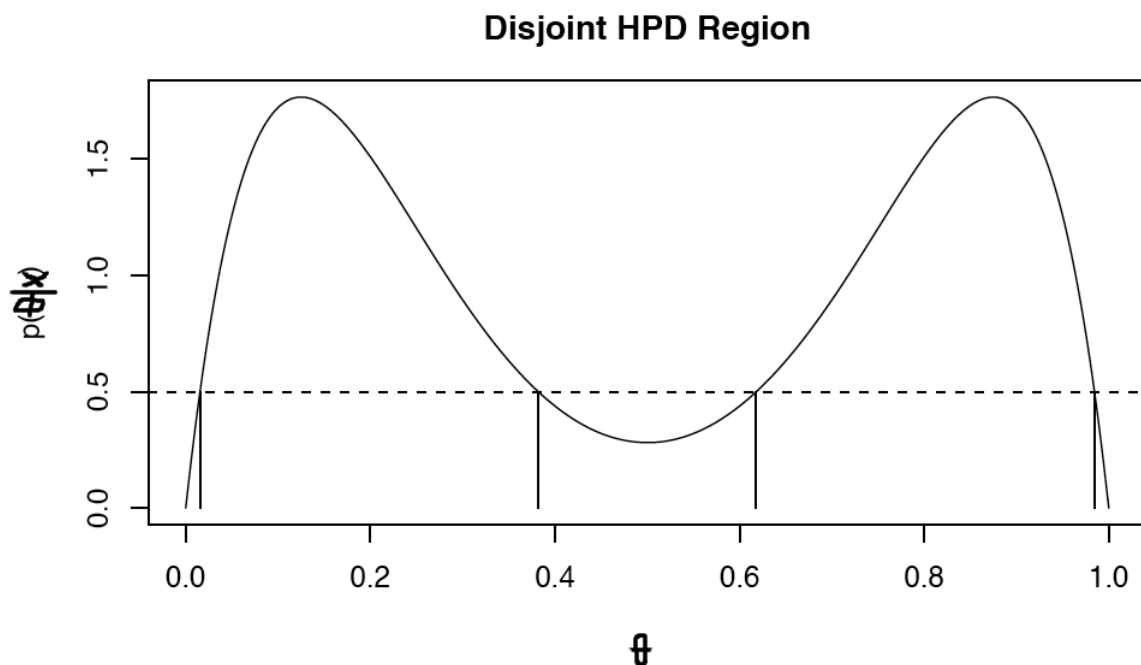
Example 2: Go back to the DC-level in AWGN example on pp. 6–15. A 95% (say) HPD credible region for μ based on

$$f_{\mu | \mathbf{x}}(\mu | \mathbf{x}) = t_{N-1} \left(\mu \mid \bar{x}, \frac{s^2}{N-1} \right)$$

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 not look “flat” under another parametrization. We have seen such an example earlier in handout # 4; here is another example that focuses on HPD credible regions.

Suppose that $0 < \theta < 1$ is a parameter of interest, but that we are equally interested in

$$\gamma = \frac{1}{1 - \theta}.$$

If the posterior density $f_{\Theta | \mathbf{X}}(\theta | \mathbf{x})$ is, say,

$$f_{\Theta | \mathbf{X}}(\theta | \mathbf{x}) = \begin{cases} 2\theta, & 0 < \theta < 1 \\ 0, & \text{otherwise} \end{cases} = 2\theta i_{(0,1)}(\theta)$$

the corresponding cdf is

$$F_{\Theta | \mathbf{X}}(\theta | \mathbf{x}) = \begin{cases} 0, & \theta < 0 \\ \theta^2, & 0 < \theta < 1 \\ 1, & \theta > 1 \end{cases}$$

implying that a 95% HPD credible set for θ is $(\sqrt{0.05}, 1)$. Now, despite the fact that

$$\gamma = \frac{1}{1 - \theta}$$

is a monotone function of θ , the interval

$$\left(\frac{1}{1 - \sqrt{0.05}}, +\infty \right) \quad (12)$$

is not an HPD credible set for γ . Clearly, (12) is a 95% credible set for γ , but it is not HPD. To see this, we find the cdf $F_{\Gamma | \mathbf{X}}(\gamma | \mathbf{x})$: for $t \geq 1$,

$$\begin{aligned} F_{\Gamma | \mathbf{X}}(t | \mathbf{x}) &= \Pr\{\Gamma \leq t | \mathbf{X} = \mathbf{x}\} \\ &= \Pr\left\{\frac{1}{1 - \Theta} \leq t \mid \mathbf{X} = \mathbf{x}\right\} \\ &= \Pr\left\{\Theta \leq 1 - \frac{1}{t} \mid \mathbf{X} = \mathbf{x}\right\} = \left(1 - \frac{1}{t}\right)^2. \end{aligned}$$

Therefore, $\{\Gamma | \mathbf{X} = \mathbf{x}\}$ has the pdf

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

with $f_{\Gamma | \mathbf{X}}(1 | \mathbf{x}) = 0$ and, consequently, HPD intervals for $\{\Gamma | \mathbf{X} = \mathbf{x}\}$ must be two-sided, which is in contrast with (12).

Computing *Maximum A Posteriori (MAP)* Estimates

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$$

where

$$V(\boldsymbol{\theta}) = -\ln f_{\boldsymbol{\Theta} | \mathbf{X}}(\boldsymbol{\theta} | \mathbf{x}).$$

Newton-Raphson Iteration:

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

where

$$\mathbf{g}_i = \left. \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}$$
$$H_i = \left. \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(i)}}.$$

We could also introduce a damped Newton-Raphson iteration:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mu_i H_i^{-1} \mathbf{g}_i \quad (13)$$

see handout # 3.

Comments:

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

Upon convergence (i.e. as $i \nearrow \infty$) and if we reached the right (global) optimum, we have

$$\mathbf{g}_\infty = \left. \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\infty)}=\hat{\boldsymbol{\theta}}_{\text{MAP}}} = \mathbf{0}. \quad (14)$$

Posterior Approximation Around the MAP Estimate

Expanding the posterior distribution $f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})$ in Taylor series around the MAP estimate

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \hat{\boldsymbol{\theta}}_{\text{MAP}}(\mathbf{x})$$

and keeping the first three terms yields:

$$\begin{aligned} \ln f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x}) &\approx \ln f_{\Theta | \mathbf{x}}(\hat{\boldsymbol{\theta}}_{\text{MAP}} | \mathbf{x}) \\ &+ \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}})^T \frac{\partial^2 \ln f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{MAP}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{MAP}}). \end{aligned} \quad (15)$$

The second term in the Taylor-series expansion vanishes because the log posterior pdf (pmf) has zero derivative at the MAP estimate, see (14).

If the number of measurements is large, the posterior distribution $f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})$ will be *unimodal*. Furthermore, if $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ is in the interior of the parameter space sp_{Θ} (preferably *far from the boundary* of sp_{Θ}), then we can use the following approximation for the posterior pdf:

$$f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x}) \approx \mathcal{N}\left(\hat{\boldsymbol{\theta}}_{\text{MAP}}, -\left[\frac{\partial^2 \ln f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{MAP}}}\right]^{-1}\right).$$

This approximation follows by looking at (15) as a function of $\boldsymbol{\theta}$.

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

If $f_{\boldsymbol{\theta} | \boldsymbol{x}}(\boldsymbol{\theta} | \boldsymbol{x})$ has multiple modes (and we can find them all), it can be approximated by a Gaussian mixture or, more generally, a t -distribution mixture.

More on Asymptotic Normality and Consistency for Bayesian Models

Consider conditionally i.i.d. observations $X[n]$, $n = 0, 1, \dots, N - 1$ given θ , following

$$\{X[n] \mid \Theta = \theta\} \sim \underbrace{f_{\text{TRUE}}(x[n])}_{\text{true distribution of the data}}. \quad (16)$$

As before, we also have

$f_{X \mid \Theta}(x \mid \theta)$, data model, likelihood;

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

Define

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[N-1] \end{bmatrix}.$$

Note that we may not be modeling the data correctly. Here, modeling the data correctly means that $f_{\text{TRUE}}(x) =$

$f_{X|\Theta}(x|\theta_0)$ 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* $D(\mathbf{p} \parallel \mathbf{q})$ from one pmf (\mathbf{p}) to another (\mathbf{q}):

$$D(\mathbf{p} \parallel \mathbf{q}) = \sum_k p_k \ln \frac{p_k}{q_k}$$

and apply it to measure distance between $p_{\text{TRUE}}(x)$ and $p_{X|\Theta}(x|\theta)$ when these distributions are discrete (pmfs):

$$\begin{aligned} D(\mathbf{p}_{\text{TRUE}} \parallel \mathbf{p}_{\Theta}) &= \sum_x p_{\text{TRUE}}(x) \ln \frac{p_{\text{TRUE}}(x)}{p_{X|\Theta}(x|\theta)} \\ &\stackrel{\text{if data modeled correctly}}{=} \sum_x p_{X|\Theta}(x|\theta_0) \ln \frac{p_{X|\Theta}(x|\theta_0)}{p_{X|\Theta}(x|\theta)} \\ &= \mathbb{E}_{X|\Theta} \left[\ln \frac{p_{X|\Theta}(X|\theta_0)}{p_{X|\Theta}(X|\theta)} \mid \theta_0 \right]. \end{aligned}$$

which is nonnegative and minimized at θ_0 , yielding

$$\mathbb{E}_{X|\Theta} \left[\ln \frac{p_{X|\Theta}(X|\theta_0)}{p_{X|\Theta}(X|\theta_0)} \mid \theta_0 \right] = 0$$

see also handout *emlecture*. **In the following discussion, we assume that the data model is correct and that θ_0 is**

the unique minimizer of

$$\mathbb{E}_{X|\Theta} \left[\ln \frac{f_{X|\Theta}(X|\theta_0)}{f_{X|\Theta}(X|\theta)} \mid \theta_0 \right].$$

Theorem 1. [Convergence in discrete parameter space.]
If the parameter space Θ is finite and

$$\underbrace{\Pr_{\Theta}\{\theta = \theta_0\}}_{\pi(\theta_0)=p_{\Theta}(\theta_0)} > 0$$

then

$$\underbrace{\Pr_{\Theta|\mathbf{X}}\{\theta = \theta_0 \mid \mathbf{X} = \mathbf{x}\}}_{p_{\Theta|\mathbf{X}}(\theta_0 \mid \mathbf{x})} \rightarrow 1 \quad \text{as } N \nearrow +\infty.$$

Proof. Terminology: we refer to the posterior ratio

$$\frac{p_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x})}{p_{\Theta|\mathbf{X}}(\theta_0 \mid \mathbf{x})}$$

as *posterior-odds ratio* (θ vs. θ_0 , in this case). Now, consider

the log posterior odds:

$$\ln \left(\frac{p_{\Theta | \mathbf{X}}(\theta | \mathbf{x})}{p_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x})} \right) = \ln \left(\frac{\pi(\theta)}{\pi(\theta_0)} \right) + \sum_{n=0}^{N-1} \ln \left(\frac{f_{X | \Theta}(x[n] | \theta)}{f_{X | \Theta}(x[n] | \theta_0)} \right). \quad (17)$$

The second term in this expression is the sum of N conditionally i.i.d. random variables given θ_0 . Recall that $X[0], X[1], \dots, X[n-1]$ are coming from $f_{\text{TRUE}}(x) = f_{X | \Theta}(x | \theta_0)$. Then

$$\mathbb{E}_{X | \Theta} \left[\ln \left(\frac{f_{X | \Theta}(X[n] | \theta)}{f_{X | \Theta}(X[n] | \theta_0)} \right) \middle| \theta_0 \right] \leq 0$$

where, by our assumption that θ_0 is the unique minimizer of $\mathbb{E}_{X | \Theta} \left[\ln \frac{p_{X | \Theta}(X | \theta_0)}{p_{X | \Theta}(X | \theta)} \middle| \theta_0 \right]$, the equality holds only for $\theta = \theta_0$. If $\theta \neq \theta_0$, the second term in (17) is the sum of N i.i.d. random variables with negative mean, which diverges to $-\infty$ as $N \nearrow \infty$. As long as

$\Pr_{\Theta} \{ \Theta = \theta_0 \} = \pi(\theta_0) > 0$ making the first term in (17) finite

the log posterior odds in (17) $\searrow -\infty$ as $N \nearrow +\infty$. Thus, if $\theta \neq \theta_0$, the posterior odds go to zero:

$$\frac{p_{\Theta | \mathbf{X}}(\theta | \mathbf{x})}{p_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x})} \rightarrow 0$$

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

$$p_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x}) \rightarrow 1.$$

□

Theorem 2. [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*

$$\Pr_{\Theta | \mathbf{X}}\{\theta \in A | \mathbf{X} = \mathbf{x}\} \rightarrow 1 \quad \text{as } N \nearrow \infty.$$

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 2 can be relaxed.

- Similarly, Theorem 1 can often be extended to allow for an infinite discrete parameter space.

Theorem 3. [Asymptotic Normality of $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$] 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 (16),

$$\sqrt{N} [\hat{\theta}_{\text{MAP}}(\mathbf{X}) - \theta_0] \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\theta_0)^{-1}) \quad \text{as } N \nearrow +\infty$$

where $\mathcal{I}_1(\theta_0)$ is the Fisher information for $\theta = \theta_0$ and a single measurement (say $X[0]$):

$$\begin{aligned} \mathcal{I}_1(\theta_0) &= \mathbb{E}_{X[0]|\Theta} \left[\left(\frac{d \ln f_{X|\Theta}(X[0]|\theta)}{d\theta} \right)^2 \middle| \theta_0 \right] \\ &= -\mathbb{E}_{X[0]|\Theta} \left[\frac{d^2 \ln f_{X|\Theta}(X[0]|\theta)}{d\theta^2} \middle| \theta_0 \right]. \end{aligned}$$

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

Here are some useful observations to help justify Theorem 3. Consider the scalar version of the Taylor-series expansion (15):

$$\begin{aligned} \ln f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) &\approx \ln f_{\Theta|\mathbf{X}}(\hat{\theta}_{\text{MAP}}|\mathbf{x}) \\ &\quad + \frac{1}{2} (\theta - \hat{\theta}_{\text{MAP}})^2 \frac{d^2}{d\theta^2} [\ln f_{\Theta|\mathbf{X}}(\hat{\theta}_{\text{MAP}}|\mathbf{x})]. \end{aligned}$$

Now, study the behavior of the negative Hessian of the log posterior pdf at θ_0 :

$$\begin{aligned} -\frac{d^2 \ln f_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x})}{d\theta^2} &= -\frac{d^2 \ln \pi(\theta_0)}{d\theta^2} - \frac{d^2 \ln f_{\mathbf{X} | \Theta}(\mathbf{x} | \theta_0)}{d\theta^2} \\ &= -\frac{d^2 \ln \pi(\theta_0)}{d\theta^2} - \sum_{n=0}^{N-1} \frac{d^2 \ln f_{X | \Theta}(x[n] | \theta_0)}{d\theta^2} \end{aligned}$$

and, therefore,

$$\mathbb{E}_{\mathbf{X} | \Theta} \left[-\frac{d^2 \ln f_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x})}{d\theta^2} \middle| \theta_0 \right] = -\frac{d^2 \ln \pi(\theta_0)}{d\theta^2} + N \mathcal{I}_1(\theta_0)$$

implying that, as N grows,

$$-\frac{d^2 \ln f_{\Theta | \mathbf{X}}(\theta_0 | \mathbf{x})}{d\theta^2} \approx N \mathcal{I}_1(\theta_0).$$

To summarize: for a large number of i.i.d. measurements (i.e. asymptotically), MAP and classical ML estimates give equivalent answers.

MAP Estimator Computation for Multiple (Subsets of) Parameters

Consider again the case of two parameter vectors, denoted by $\boldsymbol{\theta}$ and \boldsymbol{u} ; then, the vector of all unknown parameters is

$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{u} \end{bmatrix}.$$

The joint posterior pdf for $\boldsymbol{\theta}$ and \boldsymbol{u} is

$$\underbrace{f_{\boldsymbol{\theta}, \boldsymbol{u} | \boldsymbol{x}}(\boldsymbol{\theta}, \boldsymbol{u} | \boldsymbol{x})}_{f_{\boldsymbol{\rho} | \boldsymbol{x}}(\boldsymbol{\rho} | \boldsymbol{x})} = f_{\boldsymbol{u} | \boldsymbol{\theta}, \boldsymbol{x}}(\boldsymbol{u} | \boldsymbol{\theta}, \boldsymbol{x}) \cdot f_{\boldsymbol{\theta} | \boldsymbol{x}}(\boldsymbol{\theta} | \boldsymbol{x}).$$

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

Here is our *first attempt to estimate $\boldsymbol{\theta}$ and \boldsymbol{u}* : maximize the marginal posterior pdfs (pmfs)

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} f_{\boldsymbol{\theta} | \boldsymbol{x}}(\boldsymbol{\theta} | \boldsymbol{x})$$

and

$$\hat{\boldsymbol{u}} = \arg \max_{\boldsymbol{u}} f_{\boldsymbol{u} | \boldsymbol{x}}(\boldsymbol{u} | \boldsymbol{x})$$

which take into account the uncertainties about the other parameter. We can perform the two optimizations separately.

But, what if we cannot easily obtain these two marginal posterior pdfs (pmfs)? Suppose now that we can obtain $f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})$, but not $f_{U | \mathbf{x}}(\mathbf{u} | \mathbf{x})$.

Second attempt to estimate $\boldsymbol{\theta}$ and \mathbf{u} :

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

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} f_{\Theta | \mathbf{x}}(\boldsymbol{\theta} | \mathbf{x})$$

which, as desired, takes into account the uncertainty about \mathbf{u} by integrating \mathbf{u} out from the joint posterior pdf.

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

$$\hat{\mathbf{u}} = \arg \max_{\mathbf{u}} f_{U | \Theta, \mathbf{x}}(\mathbf{u} | \hat{\boldsymbol{\theta}}, \mathbf{x}).$$

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

$$(\hat{\boldsymbol{\theta}}, \hat{\mathbf{u}}) = \arg \max_{\boldsymbol{\theta}, \mathbf{u}} f_{\Theta, U | \mathbf{x}}(\boldsymbol{\theta}, \mathbf{u} | \mathbf{x}).$$

This estimation is sometimes done as follows: iterate between

1. finding the conditional MAP estimate of $\boldsymbol{\theta}$ by maximizing $f_{\boldsymbol{\theta} | U, \mathbf{x}}(\boldsymbol{\theta} | \mathbf{u}_p, \mathbf{x})$:

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

and

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

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

known as the *iterated conditional modes (ICM) algorithm*. This is simply an application of the *stepwise-ascent approach to optimization*. A general ICM algorithm is not restricted to two components ($\boldsymbol{\theta}$ and \mathbf{u} in this example) and can employ many more; the components can be scalars or vectors.

EM Algorithm for Computing Marginal MAP Estimates

We wish to find the marginal MAP estimate of θ by maximizing

$$f_{\theta | \mathbf{x}}(\theta | \mathbf{x}) = \int f_{\theta, U | \mathbf{x}}(\theta, \mathbf{u} | \mathbf{x}) d\mathbf{u}.$$

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

$$f_{\theta, U | \mathbf{x}}(\theta, \mathbf{u} | \mathbf{x}) = f_{U | \theta, \mathbf{x}}(\mathbf{u} | \theta, \mathbf{x}) \cdot f_{\theta | \mathbf{x}}(\theta | \mathbf{x})$$

is easy, we can treat \mathbf{u} as the *missing data* and apply the EM algorithm: start with

$$\ln f_{\theta | \mathbf{x}}(\theta | \mathbf{x}) = \ln f_{\theta, U | \mathbf{x}}(\theta, \mathbf{u} | \mathbf{x}) - \ln f_{U | \theta, \mathbf{x}}(\mathbf{u} | \theta, \mathbf{x})$$

and take the expectation of this expression with respect to $f_{U | \theta, \mathbf{x}}(\mathbf{u} | \theta_p, \mathbf{x})$:

$$\begin{aligned} \ln f_{\theta | \mathbf{x}}(\theta | \mathbf{x}) &= \underbrace{\mathbb{E}_{U | \theta, \mathbf{x}}[\ln f_{\theta, U | \mathbf{x}}(\theta_p, U | \mathbf{x}) | \theta_p, \mathbf{x}]}_{Q(\theta | \theta_p)} \\ &\quad - \underbrace{\mathbb{E}_{U | \theta, \mathbf{x}}[\ln f_{U | \theta, \mathbf{x}}(U | \theta_p, \mathbf{x}) | \theta_p, \mathbf{x}]}_{H(\theta | \theta_p)}. \end{aligned}$$

Recall that our goal is to maximize $\ln f_{\theta | \mathbf{x}}(\theta | \mathbf{x})$ with respect to θ . 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 (10) in **emlecture**. Hence, finding a $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta} | \boldsymbol{\theta}_p)$ will increase $\ln f_{\boldsymbol{\Theta} | \mathbf{X}}(\boldsymbol{\theta} | \mathbf{x})$:

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

EM Algorithm:

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

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}_p) &= \mathbb{E}_{U | \boldsymbol{\Theta}, \mathbf{X}}[\ln f_{\boldsymbol{\Theta}, U | \mathbf{X}}(\boldsymbol{\theta}_p, \mathbf{U} | \mathbf{x}) | \boldsymbol{\theta}_p, \mathbf{x}] \\ &= \int \ln f_{U, \boldsymbol{\Theta} | \mathbf{X}}(\mathbf{u}, \boldsymbol{\theta} | \mathbf{x}) f_{U | \boldsymbol{\Theta}, \mathbf{X}}(\mathbf{u} | \boldsymbol{\theta}_p, \mathbf{x}) d\mathbf{u}. \end{aligned}$$

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

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

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

Bayesian EM Example (from Ch. 12.3 in Gelman, Carlin, Stern, and Rubin)

Consider the classical DC-level estimation problem where the measurements $X[n]$, $n = 0, 1, \dots, N - 1$ are conditionally i.i.d. given $\boldsymbol{\theta} = [\mu, \sigma^2]^T$, following

$$\{X[n] \mid \Theta = \boldsymbol{\theta}\} \sim \mathcal{N}(\mu, \sigma^2).$$

Note that both μ and σ^2 are *unknown*.

Choose *semi-conjugate* priors with μ and σ^2 independent *a priori*:

$$\underbrace{\pi(\mu, \sigma^2)} = \pi(\mu) \cdot \pi(\sigma^2) \quad (18)$$

not a conjugate prior pdf for μ and Σ^2

and

$$\pi(\mu) = \mathcal{N}(\mu \mid \mu_0, \tau_0^2) \quad (19)$$

$$\pi(\sigma^2) \propto \underbrace{1/\sigma^2}_{\text{Jeffreys' prior}} \quad \text{see (28) in handout \# 4.} \quad (20)$$

Comments:

(What is a semi-conjugate prior?) If $\Sigma^2 = \sigma^2$ 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 (i.e. it is not a conjugate prior for $\Theta = [\mu, \Sigma^2]^T$), hence the prefix “semi.”

(The conjugate prior is obscure in this example): Conjugate prior exists for μ and Σ^2 (Θ) 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 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

$$\mu_{\text{MAP}} = \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{N}{\sigma^2} \bar{x}}{\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}} \quad (21)$$

see (15) in handout # 4 and Example 10.2 in Kay-I.

Since μ and σ^2 are independent and this prior is not conjugate (but is intuitively appealing), this problem does not have a closed-form solution for the MAP estimate of μ .

Find the marginal posterior mode (MAP estimate) of μ . In the EM framework that we introduced, μ corresponds to the parameter of interest and σ^2 corresponds to the missing data (missing datum, in this case).

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

$$\begin{aligned}
 f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) &\propto \underbrace{\exp \left[-\frac{1}{2\tau_0^2} (\mu - \mu_0)^2 \right]}_{\text{prior pdf}} \cdot (\sigma^2)^{-1} \\
 &\cdot \underbrace{(\sigma^2)^{-N/2} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]}_{\text{likelihood}}. \quad (22)
 \end{aligned}$$

This joint posterior pdf is key to all steps of our EM algorithm derivation. First, find the *conditional posterior pdf of the “missing” σ^2 given μ and \mathbf{x} , evaluated at $\mu = \mu^{(p)}$* :

$$\begin{aligned}
 f_{\Sigma^2 | \mu, \mathbf{x}}(\sigma^2 | \mu^{(p)}, \mathbf{x}) \\
 \propto (\sigma^2)^{-N/2-1} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \right]
 \end{aligned}$$

(look up the table
of distributions)

is the kernel of $\text{Inv-}\chi^2 \left(\Sigma^2 | N, \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \right)$. (23)

Now, write down the log of (22) as

$$\ln f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) = \underbrace{\text{const}}_{\text{not a function of } \mu} - \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2.$$

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

Bayesian EM Example: E Step

Conditional on $\mu^{(p)}$ and \mathbf{x} , find the expectation of $\ln f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x})$ by averaging over Σ^2 [i.e. the expectation is with respect to (23)]:

$$\begin{aligned} Q(\mu | \mu^{(p)}) &= \mathbb{E}_{\Sigma^2 | \mu, \mathbf{x}} \left[\ln f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \Sigma^2 | \mathbf{x}) \mid \mu^{(p)}, \mathbf{x} \right] \\ &= \underbrace{\text{const}}_{\text{not a function of } \mu} \underbrace{-\frac{1}{2\tau_0^2} (\mu - \mu_0)^2}_{\text{no } \Sigma^2, \text{ expectation disappears}} \\ &\quad - \frac{1}{2} \cdot \mathbb{E}_{\Sigma^2 | \mu, \mathbf{x}} \left(\frac{1}{\Sigma^2} \mid \mu^{(p)}, \mathbf{x} \right) \cdot \sum_{n=0}^{N-1} (x[n] - \mu)^2. \end{aligned}$$

Comments:

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

$$\mathbb{E}_{\Sigma^2 | \mu, \mathbf{x}} \left(\frac{1}{\Sigma^2} \mid \mu^{(p)}, \mathbf{x} \right).$$

But, (23) implies that $\{\Sigma^2 | \mu^{(p)}, \mathbf{x}\}$ is distributed as

$$\frac{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \cdot N}{Z}$$

where Z is a Chi-square χ_N^2 random variable, see p. 33 of handout # 4. Since the mean of a χ_N^2 random variable is N (see the distribution table), we have:

$$E_{\Sigma^2 | \mu, \mathbf{x}} \left(\frac{1}{\Sigma^2} \mid \mu^{(p)}, \mathbf{x} \right) = \left[\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \right]^{-1}$$

which is intuitively appealing: this expression is simply an inverse of the sample estimate of σ^2 for known μ , with μ replaced by its latest estimate $\mu^{(p)}$. Finally,

$$\begin{aligned} Q(\mu | \mu^{(p)}) &= \underbrace{\text{const}}_{\text{not a function of } \mu} - \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 \\ &\quad - \frac{1}{2} \cdot \left[\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \right]^{-1} \cdot \sum_{m=0}^{N-1} (x[m] - \mu)^2 \\ &= \underbrace{\text{const}}_{\text{not a function of } \mu} - \frac{1}{2\tau_0^2} (\mu^2 - 2\mu\mu_0) \\ &\quad - \frac{1}{2} \cdot \left[\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2 \right]^{-1} \cdot \left\{ N\mu^2 - 2\mu \left(\sum_{m=0}^{N-1} x[m] \right) \right\} \\ &= \text{const} - \frac{1}{2} \left[\frac{1}{\tau_0^2} + \frac{N}{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2} \right] \mu^2 \\ &\quad + \left(\frac{\mu_0}{\tau_0^2} + \frac{N\bar{x}}{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2} \right) \cdot \mu. \end{aligned}$$

Bayesian EM Example: M Step

Find μ that maximizes $Q(\mu | \mu^{(p)})$ and choose it to be $\mu^{(p+1)}$:

$$\mu^{(p+1)} = \frac{\frac{1}{\tau_0^2} \cdot \mu_0 + \frac{N}{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2} \cdot \bar{x}}{\frac{1}{\tau_0^2} + \frac{N}{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2}} \quad (24)$$

which is simple and intuitive. Compare (24) with (21) and note that

$$\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu^{(p)})^2$$

estimates σ^2 based on $\mu^{(p)}$. Our iteration (24) converges to the marginal posterior mode of $f_{\mu | \mathbf{x}}(\mu | \mathbf{x})$.

Why Not Integrate Σ^2 Out from (22)?

We can, but still no closed-form MAP Estimate.

A related question: why isn't the prior in (18)–(20) nice, i.e. conjugate?

Recall (22):

$$f_{\mu, \Sigma^2 | \mathbf{x}}(\mu, \sigma^2 | \mathbf{x}) \propto \underbrace{\exp \left[-\frac{1}{2\tau_0^2} (\mu - \mu_0)^2 \right] \cdot (\sigma^2)^{-1}}_{\text{prior pdf}} \cdot \underbrace{(\sigma^2)^{-N/2} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]}_{\text{likelihood}} \quad (25)$$

and note that

$$f_{\Sigma^2 | \mu, \mathbf{x}}(\sigma^2 | \mu, \mathbf{x}) \propto (\sigma^2)^{-N/2-1} \cdot \exp \left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]$$

(look up the table of distributions)

is the kernel of $\text{Inv-}\chi^2 \left(\Sigma^2 | N, \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right)$

and, therefore [see (8)]:

$$\begin{aligned}
 f_{\Sigma^2 | \mu, \mathbf{X}}(\sigma^2 | \mu, \mathbf{x}) &= \text{Inv-}\chi^2(\sigma^2 | N, \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2) \\
 &= \frac{(N/2)^{N/2}}{\Gamma(N/2)} \left[\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]^{N/2} \\
 &\cdot (\sigma^2)^{-(N/2+1)} \exp \left(- \frac{N \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2}{2 \sigma^2} \right) i_{[0, +\infty)}(\sigma^2). \quad (26)
 \end{aligned}$$

Now, utilize (25) and (26) and apply our Bayesian trick:

$$\underbrace{f_{\mu | \mathbf{X}}(\mu | \mathbf{x})}_{\text{not a function of } \sigma^2} = \frac{f_{\mu, \Sigma^2 | \mathbf{X}}(\mu, \Sigma^2 | \mathbf{x})}{f_{\Sigma^2 | \mu, \mathbf{X}}(\mu | \sigma^2, \mathbf{x})}$$

∞
 keep track of
 μ , be careful with (26)

$$\frac{e^{-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2} (\sigma^2)^{-N/2-1} \cdot \exp \left\{ - \frac{1}{2\sigma^2} [N s^2 + N (\bar{x} - \mu)^2] \right\}}{\left[\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]^{N/2} (\sigma^2)^{-N/2-1} \cdot \exp \left\{ - \frac{1}{2\sigma^2} \left[\sum_{n=0}^{N-1} (x[n] - \mu)^2 \right] \right\}}$$

$\overset{\sigma^2 \text{ cancels out}}{\infty} \exp \left[- \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 \right] \cdot \left[\sum_{n=0}^{N-1} (x[n] - \mu)^2 \right]^{-N/2} \quad ???$

This pdf is not in the distribution table and we cannot find a closed form for its maximization with respect to μ . This is why we need an iteration, such as EM algorithm.

Yet, the above expression is useful: we can derive a (damped) Newton-Raphson iteration for its maximization. A Newton-Raphson iteration is an alternative to the EM algorithm that we presented for this problem and can be computationally more efficient than the EM algorithm.