Bayesian Inference for Multiparameter Models

Aleksandar Dogandžić

2017-03-23

Contents

Reading: §11 in the textbook and [Gelman et al. 2014, §3.1 and 3.2 and pp. 108–109].

Multiparameter Models

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

For example, if $X[n]$ are independent, identically distributed (i.i.d.) $\mathcal{N}(\mu, \sigma^2)$ given μ and σ^2, *both* μ and σ^2 may be *unknown* parameters. 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) is a multiparameter model.

☞ 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.[1] An example would be the DC-signal-in-additive white Gaussian noise (AWGN) model, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

[1] θ_1 and θ_2 could be vectors, but we describe the scalar case for simplicity.

Our inference should be based on

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

the marginal posterior probability density function (pdf) (or probability mass function (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):

Put the quantities we know on the right and the quantities we do not know on the left.

$$f_{\Theta_1,\Theta_2 | \boldsymbol{X}}(\theta_1, \theta_2 \,|\, \boldsymbol{x}) \propto f_{\boldsymbol{X} | \Theta_1,\Theta_2}(\boldsymbol{x} \,|\, \theta_1, \theta_2)\, f_{\Theta_1,\Theta_2}(\theta_1, \theta_2)$$

and then, in the continuous (pdf) case, *integrate out* the nuisance parameter:

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

also discussed in §10.7 of the textbook

or, equivalently,

$$f_{\Theta_1 | \boldsymbol{X}}(\theta_1 \,|\, \boldsymbol{x}) = \int f_{\Theta_1 | \Theta_2, \boldsymbol{X}}(\theta_1 \,|\, \theta_2, \boldsymbol{x})\, f_{\Theta_2 | \boldsymbol{X}}(\theta_2 \,|\, \boldsymbol{x})\, \mathrm{d}\theta_2.$$

☞ 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 of the nuisance parameter. Hence, the uncertainty due to the unknown θ_2 is taken into account.

✻ EXAMPLE: Predictive Distribution. Suppose that we wish to predict an observation $X_\star$ coming from the model

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

If $X_\star$ and $\boldsymbol{x}$ are independent given θ_1 and θ_2, then, based on (7) from handout `bpred`, we obtain:

$$\begin{aligned} f_{X_\star | \boldsymbol{X}}(x_\star \,|\, \boldsymbol{x}) &= \int\!\!\int f_{X_\star | \Theta_1,\Theta_2}(x_\star \,|\, \theta_1, \theta_2)\, f_{\Theta_1,\Theta_2 | \boldsymbol{X}}(\theta_1, \theta_2 \,|\, \boldsymbol{x})\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2 \\ &= \int\!\!\int f_{X_\star | \Theta_1,\Theta_2}(x_\star \,|\, \theta_1, \theta_2)\, f_{\Theta_1 | \Theta_2, \boldsymbol{X}}(\theta_1 \,|\, \theta_2, \boldsymbol{x})\, f_{\Theta_2 | \boldsymbol{X}}(\theta_2 \,|\, \boldsymbol{x})\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2. \end{aligned}$$

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 wish to find the mode of

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

an expectation-maximization (EM) algorithm may suffice. The EM algorithm will be discussed in detail later.

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

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

In the following, we consider two classical cases where analytical Bayesian computations are possible.

DC Level Estimation in AWGN with Unknown Variance

THE measurements $(X[n])_{n=0}^{N-1}$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ given $\boldsymbol{\theta}$, where

see [Gelman et al. 2014, §3.2]

$$\boldsymbol{\theta} = (\mu, \sigma^2)$$

are the unknown parameters.

Assume that μ and σ^2 are independent and use the standard noninformative Jeffreys' priors for each:

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

The likelihood function $f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \mid \boldsymbol{\theta})$ is

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

where

$$s^2(\boldsymbol{x}) = s^2 = \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2, \qquad \bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{2b}$$

are the sufficient statistics for μ and σ^2. The joint posterior pdf of the parameters is the product of the prior pdf (1) and likelihood (2a):

$$\begin{aligned} f_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta} \mid \boldsymbol{x}) &= f(\mu, \sigma^2 \mid \boldsymbol{x}) \\ &\propto \frac{1}{\sigma^2} (\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} [N s^2 + N(\bar{x} - \mu)^2]\right\} \\ &\quad \cdot \mathbb{1}_{[0,+\infty)}(\sigma^2). \end{aligned} \tag{3}$$

Provided that $N \geq 2$, the posterior pdf (3) is proper.

Conditional posterior pdf of μ given σ^2

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

$$\begin{aligned} f(\mu \mid \sigma^2, \boldsymbol{x}) &\propto f(\mu, \sigma^2 \mid \boldsymbol{x}) \\ &\propto \frac{1}{\sigma^2} \mathbb{1}_{[0,+\infty)}(\sigma^2)(\sigma^2)^{-N/2} \exp\Big\{-\frac{1}{2\sigma^2}[Ns^2 + N(\overline{x}-\mu)^2]\Big\} \\ &\propto \exp\Big[-\frac{N}{2\sigma^2}(\overline{x}-\mu)^2\Big] \\ &\text{is the kernel of} \quad \mathcal{N}\big(\mu \mid \overline{x}, \sigma^2/N\big) \qquad (4) \end{aligned}$$

see (3)

keep only terms that have μ

which agrees with (8) in handout `introBayes`, the case of estimating the DC level in AWGN with known variance.

We now write $f(\mu \mid \sigma^2, \boldsymbol{x})$ in (4) *with the normalizing constant*:

$$f(\mu \mid \sigma^2, \boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2/N}} \exp\Big[-\frac{(\mu-\overline{x})^2}{2\sigma^2/N}\Big]. \qquad (5)$$

see the table of distributions

Conditional posterior pdf of σ^2 given μ

A reminder:

$$\text{Inv-}\chi^2(\sigma^2 \mid \nu_0, \sigma_0^2) \propto (\sigma^2)^{-(\nu_0/2+1)} \exp\Big(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\Big) \mathbb{1}_{[0,+\infty)}(\sigma^2) \qquad (6)$$

and the full scaled inverted χ^2 pdf is

$$\begin{aligned} \text{Inv-}\chi^2(\sigma^2 \mid \nu_0, \sigma_0^2) &= \frac{(\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} (\sigma_0^2)^{\nu_0/2} (\sigma^2)^{-(\nu_0/2+1)} \\ &\quad \cdot \exp\Big(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\Big) \mathbb{1}_{[0,+\infty)}(\sigma^2) \qquad (7) \end{aligned}$$

see the table of distributions.

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

$$\begin{aligned} f(\sigma^2 \mid \mu, \boldsymbol{x}) &\propto f(\mu, \sigma^2 \mid \boldsymbol{x}) \\ &\propto \frac{1}{\sigma^2} \mathbb{1}_{[0,+\infty)}(\sigma^2)(\sigma^2)^{-N/2} \exp\Big\{-\frac{1}{2\sigma^2}[Ns^2 + N(\overline{x}-\mu)^2]\Big\} \\ &\propto (\sigma^2)^{-(N/2+1)} \exp\Big\{-\frac{1}{2\sigma^2}[Ns^2 + N(\overline{x}-\mu)^2]\Big\} \mathbb{1}_{[0,+\infty)}(\sigma^2) \\ &\text{is the kernel of} \quad \text{Inv-}\chi^2\big(\sigma^2 \mid N, s^2 + (\overline{x}-\mu)^2\big). \qquad (8) \end{aligned}$$

see (3)

keep track of σ^2

Therefore,

$$\begin{aligned} f(\sigma^2 \mid \mu, \boldsymbol{x}) &= \frac{(\frac{N}{2})^{N/2}}{\Gamma(N/2)} [s^2 + (\overline{x}-\mu)^2]^{N/2} \\ &\quad \cdot (\sigma^2)^{-(N/2+1)} \exp\Big(-\frac{Ns^2 + N(\overline{x}-\mu)^2}{2\sigma^2}\Big) \mathbb{1}_{[0,+\infty)}(\sigma^2) \end{aligned}$$

see (7).

Marginal posterior pdf of σ^2

.

THE marginal posterior pdf of σ^2 is a scaled inverted χ^2:

$$f(\sigma^2 \,|\, \boldsymbol{x}) = \text{Inv-}\chi^2\Big(N-1, \frac{1}{N-1}\sum_{n=0}^{N-1}(x[n]-\overline{x})^2\Big) \tag{9}$$

[Gelman et al. 2014, §3.2] performs the integration whereas here we present an algebraic derivation without explicit integral computation based on [Gelman et al. 2014, pp. 108-109]

i.e., $\{\sigma^2 \,|\, \boldsymbol{x} = \boldsymbol{x}\}$ can be simulated as follows:

$$\{\sigma^2 \,|\, \boldsymbol{x} = \boldsymbol{x}\} \sim \frac{\sum_{n=0}^{N-1}(x[n]-\overline{x})^2}{Z}$$

where Z is a χ^2_{N-1} random variable, see (2b) and (12) from handout `introBayes`.

To derive this result, we apply an algebraic approach for integrating μ out. The key to this approach is the fact that the conditional posterior pdf $f(\mu \,|\, \sigma^2, \boldsymbol{x})$ 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 (9):

$$
\begin{aligned}
f(\sigma^2 \,|\, \boldsymbol{x}) &= \frac{\overbrace{f(\mu, \sigma^2 \,|\, \boldsymbol{x})}^{\text{see (3)}}}{\underbrace{f(\mu \,|\, \sigma^2, \boldsymbol{x})}_{\mathcal{N}(\mu \,|\, \bar{x}, \sigma^2/N),\ \text{see (5)}}} && \text{not a function of } \mu \\
&\propto \frac{(\sigma^2)^{-N/2-1} \exp\{-\frac{1}{2\sigma^2}[N s^2 + N(\bar{x} - \mu)^2]\}\, \mathbb{1}_{[0,+\infty)}(\sigma^2)}{(\sigma^2)^{-1/2} \exp[-\frac{(\mu - \bar{x})^2}{2\sigma^2/N}]} && \text{keep track of both } \mu \text{ and } \sigma^2 \\
&\propto \frac{(\sigma^2)^{-N/2-1} \exp[-N s^2/(2\sigma^2)]\, \mathbb{1}_{[0,+\infty)}(\sigma^2)}{(\sigma^2)^{-1/2}} && \text{plug in } \mu = \bar{x} \\
&\propto (\sigma^2)^{-[(N-1)/2-1]} \exp\Big(-\frac{N s^2}{2\sigma^2}\Big)\, \mathbb{1}_{[0,+\infty)}(\sigma^2) \\
&\propto (\sigma^2)^{-[(N-1)/2-1]} \exp\Big\{-\frac{\sum_{n=0}^{N-1} (x[n] - \bar{x})^2}{2\sigma^2}\Big\}\, \mathbb{1}_{[0,+\infty)}(\sigma^2) && \text{rearrange to look like (6)} \\
&\text{is the kernel of } \text{Inv-}\chi^2\Big(\sigma^2 \,\Big|\, N-1, \frac{1}{N-1} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2\Big)
\end{aligned}
$$

❊ CHECK: μ must cancel out—important for verifying the correctness of our computations.

☞ WE have computed

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

algebraically *without* performing the actual integration.

Marginal posterior pdf of μ

APPLY the algebraic approach to integrate σ^2 out:

[Gelman et al. 2014, §3.2] performs the integration whereas here we present an algebraic derivation without explicit integral computation based on [Gelman et al. 2014, pp. 108-109].

$$
\begin{aligned}
f(\mu \mid \boldsymbol{x}) &= \frac{\overbrace{f(\mu, \sigma^2 \mid \boldsymbol{x})}^{\text{see (3)}}}{\underbrace{f(\sigma^2 \mid \mu, \boldsymbol{x})}_{\text{Inv-}\chi^2(\sigma^2 \mid N, s^2 + (\bar{x}-\mu)^2),\ \text{see (9)}}} \\
&\propto \frac{(\sigma^2)^{-N/2-1} \exp\{-[Ns^2 + N(\bar{x}-\mu)^2]/(2\sigma^2)\}\, \mathbb{1}_{[0,+\infty)}(\sigma^2)}{[s^2 + (\bar{x}-\mu)^2]^{N/2} (\sigma^2)^{-N/2-1} \exp\{-[Ns^2 + N(\bar{x}-\mu)^2]/(2\sigma^2)\}\, \mathbb{1}_{[0,+\infty)}(\sigma^2)} \\
&\propto [s^2 + (\bar{x}-\mu)^2]^{-N/2} \\
&\propto \Big[1 + \frac{1}{N-1} \frac{(\mu-\bar{x})^2}{s^2/(N-1)}\Big]^{-(N-1+1)/2}
\end{aligned}
$$

not a function of σ^2

kernel of the Student-t distribution

keep track of both μ and σ^2

☞ CHECK: σ^2 must cancel out.
From the table of distributions, we find:

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

which is a Student-t pdf with $N-1$ degrees of freedom and parameters

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

☞ WE have computed

$$
f(\mu \mid \boldsymbol{x}) = \int_{-\infty}^{+\infty} f_{\mu,\sigma^2 \mid \boldsymbol{X}}(\mu, \sigma^2 \mid \boldsymbol{x})\, \mathrm{d}\sigma^2
$$

algebraically, *without* performing the actual integration.

✻ A credible set for the DC-level-in-AWGN example. The typical 95 % highest posterior density (HPD) credible interval for μ based on

$$
f_{\mu \mid \boldsymbol{X}}(\mu \mid \boldsymbol{x}) = t_{N-1}\Big(\mu \,\Big|\, \bar{x}, \frac{s^2}{N-1}\Big)
$$

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

☞ EXAMPLE 11.4 in the textbook is equivalent to the conjugate prior case for (μ, σ^2) in [Gelman et al. 2014, §3.3], which is not practically relevant.

Computing MAP Estimates

MAXIMUM *a posteriori* (MAP) estimation corresponds to the minimization problem:

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

where

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

Newton-Raphson iteration

ITERATE

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \big(H^{(t)}\big)^{-1}\boldsymbol{g}^{(t)}$$

where t denotes the iteration index and

$$\begin{aligned} \boldsymbol{g}^{(t)} &= \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ H_i &= \frac{\partial^2 V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \end{aligned}$$

gradient

A damped Newton-Raphson variation:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mu^{(t)}\big(H^{(t)}\big)^{-1}\boldsymbol{g}^{(t)} \tag{10}$$

see handout `ml`.

✻ COMMENTS:

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

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

$$\boldsymbol{g}^{(+\infty)} = \frac{\partial V(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(+\infty)}=\widehat{\boldsymbol{\theta}}_{\text{MAP}}} = \mathbf{0}. \tag{11}$$

The EM algorithm for marginal MAP estimation is discussed in handout `emBayes`.

MAP Estimation for Multiple Parameters

CONSIDER again the case of two parameter vectors, denoted by $\boldsymbol{\theta}$ and $\boldsymbol{u}$; then, the set of all unknown parameters is $(\boldsymbol{\theta}, \boldsymbol{u})$. The joint posterior pdf for $\boldsymbol{\theta}$ and $\boldsymbol{u}$ is

$$f_{\boldsymbol{\Theta},\boldsymbol{U}\,|\,\boldsymbol{X}}(\boldsymbol{\theta},\boldsymbol{u}\,|\,\boldsymbol{x}) = f_{\boldsymbol{U}\,|\,\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{x})\, f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x}).$$

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

Approach I

MAXIMIZE the marginal posterior pdfs (pmfs)

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$$

and

$$\widehat{\boldsymbol{u}} = \arg\max_{\boldsymbol{u}} f_{\boldsymbol{U}\,|\,\boldsymbol{X}}(\boldsymbol{u}\,|\,\boldsymbol{x})$$

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 $f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, but not $f_{\boldsymbol{U}\,|\,\boldsymbol{X}}(\boldsymbol{u}\,|\,\boldsymbol{x})$.

Approach II

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

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta}\,|\,\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{x})$$

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

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

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

Finally, what if we cannot easily obtain either of the two marginal posterior pdfs (pmfs)?

Approach III

FIND the joint MAP estimate $(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{u}})$,

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

This estimation is sometimes performed as follows: iterate between

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

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta}\,|\,\boldsymbol{U},\boldsymbol{X}}(\boldsymbol{\theta}\,|\,\boldsymbol{u}^{(t)},\boldsymbol{x})$$

   and

2. finding the conditional MAP estimate of $\boldsymbol{u}$ by maximizing $f_{\boldsymbol{U}\mid\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{u} \mid \boldsymbol{\theta}^{(t+1)}, \boldsymbol{x})$:

$$\boldsymbol{u}^{(t+1)} = \arg\max_{\boldsymbol{u}} f_{\boldsymbol{U}\mid\boldsymbol{\Theta},\boldsymbol{X}}(\boldsymbol{u} \mid \boldsymbol{\theta}^{(t+1)}, \boldsymbol{x})$$

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 $\boldsymbol{u}$ in this example) and can employ many more; the components can be scalars or vectors.

✻ COMMUNICATIONS example. In communications, $\boldsymbol{\theta}$ and $\boldsymbol{u}$ can be (blocks of) symbols. Then, approach III corresponds to the Viterbi algorithm and Approach I to the Bahl-Cocke-Jelinek-Raviv (BCJR) algorithm [Bahl et al. 1974].

Acronyms

AWGN additive white Gaussian noise. 1, 4, 7

BCJR Bahl-Cocke-Jelinek-Raviv. 10

EM expectation-maximization. 2, 8

HPD highest posterior density. 7

i.i.d. independent, identically distributed. 1, 3

ICM iterated conditional modes. 10

MAP maximum *a posteriori*. 8–10

pdf probability density function. 2–5, 7–9

pmf probability mass function. 2, 5, 9

References

Bahl, L., J. Cocke, F. Jelinek, and J. Raviv (1974). "Optimal decoding of linear codes for minimizing symbol error rate". In: *IEEE Trans. Inf. Theory* 20.2, pp. 284–287 (cit. on p. 10).

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: Taylor & Francis (cit. on pp. 1, 3, 5, 7).