Bayesian computations – priors

Felipe Uribe

Computational Engineering
School of Engineering Sciences
Lappeenranta-Lahti University of Technology (LUT)

Special Course on Inverse Problems
Lappeenranta, FI — January-February, 2024

Recap: general comments

- Whenever we solve inverse problems, it is important to make an assessment of the uncertainties of the inversion estimates to determine their reliability (frequentist or Bayesian). There is no unique framework to address the question of uncertainty quantification.
- We have also realized that uncertain parameters are oftentimes spatially variable and random fields are used for their representation. When random fields are discretized, the inverse problem becomes high dimensional.
- Moreover, we know that Monte Carlo methods are required to compute samples describing the underlying probability distributions, from which we get statistical descriptors.
- MAP and ML estimators are not enough, we need to perform sampling to do UQ properly.

1a. Prior modeling: general

The data cannot make possible an event that is impossible under the prior.

Prior modeling: introduction

- So far, we are familiar with the notion of Gaussian (random field) priors.
- The majority of research in Bayesian formulations has been on:
  - Finding ways to sample from the posterior — (approximated) inferences.
  - Deriving priors: (i) more flexible/structure encoding, (ii) hyperparameters.

    Example: prior elicitation, which is the process of extracting the subjective knowledge of domain experts in a structured manner. We attempt to construct the *most suitable* prior distribution.
  - Beyond standard inferences (data assimilation, model selection, experiment design, surrogate modeling, etc.).
- In general, prior information/knowledge can be formulated in various ways. There is never a single option, and this could perhaps explain to some extent opposition to Bayesian thinking: *from the beginning there is no "absolute truth"*.

Prior modeling: types

Non-informative:

- They are also called *reference* priors, i.e., default priors when prior information is missing.
- When there is no physical theory that defines a stochastic model for the uncertain parameters.
- Non-informative actually implies something closer to being objective: *choose prior distributions without adding artificial information*.
- e.g., Jeffreys' and reference priors, maximum entropy priors, etc.

Informative:

- Express or convey some specific information about the unknown parameters.
- Information based on historical data, insight, or personal/expert beliefs. **Warning**: avoid using the same data twice (aka "double dipping"); leads to falsely overconfident results.
- e.g., prior elicitation (structural priors), computational convenient priors (conjugate priors).

Non-informative: maximum entropy (I)

- Much criticism of Bayesian inference concerns the fact that the result of the analysis depends on the choice of prior, and that the assignment of this prior seems rather subjective. *Is there some objective way of assigning a prior when we know little about its possible distribution?*
- The principle of maximum entropy (MaxEnt) states that the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy. The principle of maximum entropy can be seen as an application of Occam's razor[1].
- If nothing is known about a distribution (except that it belongs to a certain class), then the distribution with the largest entropy should be chosen as the least-informative default.

[1] The problem-solving principle saying that "the simplest explanation is usually the best one".

Non-informative: maximum entropy (II)

- The motivation is twofold: (i) maximizing entropy minimizes the amount of prior information assigned to the distribution (a distribution with high entropy carries the fewest constraints); (ii) many physical systems tend to move towards maximal entropy configurations over time.

- **Oftentimes we are not totally ignorant of the prior**. For example, maybe we know the mean value of the distribution, its variance, or the average value of some function of the parameter.

  These all are examples of constraints. The MaxEnt prior will then be the probability distribution that maximizes the entropy under such constrains.

- Thomas Jaynes argued that the MaxEnt distribution is "*uniquely determined as the one which is maximally noncommittal with regard to missing information, in that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters*".

Non-informative: maximum entropy (III)

- Consider a discrete random variable X, with outcomes $x_1, \ldots, x_n$, which occur with probability $\mathbb{P}[x_1], \ldots, \mathbb{P}[x_n]$. The **entropy**[2] of X is the average level of information inherent to its possible outcomes:

$$S = -\sum_{i=1}^{n} \mathbb{P}[x_i] \log(\mathbb{P}[x_i]).$$

- For continuous distributions, the previous (Shannon) entropy cannot be used, as it is only defined for discrete probability spaces. In this case, we have the following formula[3], which is closely related to the **differential entropy**:

$$S = -\int_{\mathbb{R}} \pi(x) \log\left(\pi(x)\right) \, \mathrm{d}x. \tag{1}$$

[2] by Charles Shannon in 1948, this generalized the concept of entropy (from thermodynamics) to a probability distribution.

[3] by Edwin Jaynes in 1963.

Non-informative: maximum entropy (IV)

- The maximum entropy principle is a means of deriving probability distributions given certain constraints and the assumption of maximizing entropy. That is, the task is to find a set of probabilities $\mathbb{P}[x_1], \ldots, \mathbb{P}[x_n]$ (discrete), or the density function (continuous) that maximizes the entropy S.

- One way for solving this maximization problem is via **Lagrange multipliers**. General steps:
  - Introduce a new variable λ, called Lagrange multiplier, and define a new target function.
  - Set the derivative of the function equal in order to zero to find the critical points.
  - Consider each resulting solution within the limits of the constraints and derive the resulting distribution.

Non-informative: maximum entropy (example 1)

- Suppose a RV X for which we have no information on its distribution (besides the fact that it should have a density). What type of PDF gives maximum entropy when the RV is bounded by a finite interval, say $a \leq X \leq b$?

- The Lagrangian equation gives:

$$\mathcal{L}[\pi] = -\int_a^b \pi(x)\log(\pi(x))\,\mathrm{d}x + \lambda\left(\int_a^b \pi(x)\,\mathrm{d}x - 1\right). \tag{2}$$

First, differentiating $\mathcal{L}$ with respect to $\pi(x)$ (this is in the sense of calculus of variations)[4]

$$\frac{\delta\mathcal{L}}{\delta\pi} = 0 \qquad \longrightarrow \qquad -1 - \log(\pi(x)) + \lambda = 0 \qquad \longrightarrow \qquad \pi(x) = \exp(\lambda - 1). \tag{3}$$

[4] The functional derivative of $\mathcal{L}[\pi]$ is defined through $\int \frac{\delta\mathcal{L}}{\delta\pi}(x)\phi(x)\,\mathrm{d}x = \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{L}[\pi + \varepsilon\phi]\right]_{\varepsilon=0}$, where ϕ is an arbitrary function.

Non-informative: maximum entropy (example 1)

Second, the result of $\pi(x)$ has to satisfy the stated constraint:

$$\int_a^b \exp(1-\lambda)\,\mathrm{d}x = 1 \qquad \longrightarrow \qquad \lambda = 1 - \log\left(\frac{1}{b-a}\right). \tag{4}$$

- Taking both solutions together, we get the following PDF:

$$\pi(x) = \exp\left(1 - \left[1 - \log\left(\frac{1}{b-a}\right)\right]\right) = \frac{1}{b-a}, \tag{5}$$

which is the uniform PDF on the interval $[a, b]$.

Hence, the MaxEnt distribution associated with X is uniform between a and b.

Non-informative: maximum entropy (example 2)

- Suppose now that X has a preassigned mean μ and standard deviation σ. Which function $\pi(x)$ gives the maximum of the entropy?
- The Lagrangian equation gives:

$$\mathcal{L} = -\int_{\mathbb{R}} \pi(x)\log(\pi(x))\,\mathrm{d}x + \lambda_0\left(\int_{\mathbb{R}} \pi(x)\,\mathrm{d}x - 1\right) + \lambda_1\left(\int_{\mathbb{R}} (x-\mu)^2\pi(x)\,\mathrm{d}x - \sigma^2\right).$$

note that only one constrained is needed, as μ is already included in the definition of σ^2.

Next, $\mathcal{L}$ is differentiated with respect to $\pi(x)$:

$$\frac{\delta\mathcal{L}}{\delta\pi(x)} = 0 \tag{6a}$$

$$-1 - \log(\pi(x)) + \lambda_0 + \lambda_1(x-\mu)^2 = 0 \quad \longrightarrow \quad \pi(x) = \exp(\lambda_0 + \lambda_1(x-\mu)^2 - 1). \tag{6b}$$

Non-informative: maximum entropy (example 2)

Now, the result of $\pi(x)$ has to satisfy the stated constraints:

$$\int_{\mathbb{R}} \exp(\lambda_0 + \lambda_1 (x-\mu)^2 - 1)\,\mathrm{d}x = 1 \quad \text{and} \quad \int_{\mathbb{R}} (x-\mu)^2 \exp(\lambda_0 + \lambda_1 (x-\mu)^2 - 1)\,\mathrm{d}x = \sigma^2$$

$$\exp(\lambda_0 - 1)\sqrt{-\frac{\pi}{\lambda_1}} = 1 \quad \text{and} \quad \exp(\lambda_0 - 1) = \sqrt{\frac{1}{2\pi\sigma^2}}. \tag{7}$$

- Putting everything together, we get the following PDF:

$$\pi(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \tag{8}$$

which is the Gaussian PDF.

Non-informative: maximum entropy (final comments)

- If we want to infer a probability distribution given certain constraints, out of all distributions compatible with them, one should pick the distribution having the largest value of S.
- A MaxEnt distribution is completely determined by features that appear explicitly in the constraints.
- An list of MaxEnt distributions (Link Wikipedia).
- A disadvantage of the MaxEnt priors is that they mostly arise from the exponential family, which is not broad enough to include some useful priors (e.g., in cases when we want to encode some structural information).

Non-informative: Jeffreys (Fisher information)

- If the only information available about $\boldsymbol{x}$ is the likelihood function, then it makes sense to use π_{like} to define the prior. **Assumption:** regularity conditions on the likelihood...

- The score function is defined as the partial derivative with respect to $\boldsymbol{x}$ of the natural logarithm of the likelihood function:

$$\mathcal{S}(\boldsymbol{x};\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{x}} \log \pi_{\text{like}}(\boldsymbol{y} \,|\, \boldsymbol{x}), \tag{9}$$

  which can be interpreted as the relative change of the likelihood at $\boldsymbol{x}$.

- The Fisher information can be understood (i) as the covariance matrix of the score function:

$$\mathcal{I}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}}\left[\mathcal{S}(\boldsymbol{x};\boldsymbol{y})\mathcal{S}(\boldsymbol{x};\boldsymbol{y})^{\mathsf{T}}\right] = \mathbb{E}_{\boldsymbol{y}}\left[\left(\frac{\partial}{\partial \boldsymbol{x}} \log \pi_{\text{like}}(\boldsymbol{y} \,|\, \boldsymbol{x})\right)^2\right], \tag{10}$$

  which can be interpreted as the *variability* of the relative change of the likelihood at $\boldsymbol{x}$.

Non-informative: Jeffreys (Fisher information)

- or (ii) as an expectation of the Hessian of the negative log-likelihood:

$$\mathcal{I}(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial^2}{\partial \boldsymbol{x}^2} \log \pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})\right] = -\mathbb{E}_{\boldsymbol{y}}[\mathbf{H}(\boldsymbol{x}; \boldsymbol{y})]\,, \tag{11}$$

  which can be seen as the curvature of the log-likelihood. Near the ML estimate, low Fisher information indicates that the ML is shallow and there are many nearby values with a similar log-likelihood. Conversely, high Fisher information indicates that the ML is sharp.

Non-informative: Jeffreys (Fisher information)

- or (ii) as an expectation of the Hessian of the negative log-likelihood:

$$\mathcal{I}(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}}\left[\frac{\partial^2}{\partial \boldsymbol{x}^2}\log \pi_{\text{like}}(\boldsymbol{y}\,|\,\boldsymbol{x})\right] = -\mathbb{E}_{\boldsymbol{y}}[\mathbf{H}(\boldsymbol{x};\boldsymbol{y})]\,, \tag{12}$$

  which can be seen as the curvature of the log-likelihood. Near the ML estimate, low Fisher information indicates that the ML is shallow and there are many nearby values with a similar log-likelihood. Conversely, high Fisher information indicates that the ML is sharp.

- We want a rule that assigns a prior π_{pr} to a given likelihood function π_{like} that does not change with different parametrizations.

- **Jeffreys Prior**: it is invariant under monotone transformations of the parameter. That is, the relative probability assigned to a volume of a probability space using a Jeffreys prior will be the same regardless of the parameterization used to define such prior.

Non-informative: Jeffreys

- Let $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ be two possible parametrizations of a statistical model, with $\boldsymbol{x}$ a continuously differentiable function $\widetilde{\boldsymbol{x}}$. We call the prior $\pi_{\mathrm{pr}}(\boldsymbol{x})$ *invariant under re-parametrization* if

$$\pi_{\mathrm{pr}}(\widetilde{\boldsymbol{x}}) = \pi_{\mathrm{pr}}(\boldsymbol{x}) \det(\mathbf{J}), \tag{13}$$

  where $\mathbf{J}$ is the Jacobian matrix with entries $\partial \boldsymbol{x}_i / \partial \widetilde{\boldsymbol{x}}_j$.

- The Fisher information matrix transforms under re-parametrizations as

$$\mathcal{I}_{\widetilde{\boldsymbol{x}}}(\widetilde{\boldsymbol{x}}) = \mathbf{J}^{\mathsf{T}} \mathcal{I}_{\boldsymbol{x}}(\boldsymbol{x})\, \mathbf{J},$$

  we have that $\det(\mathcal{I}_{\widetilde{\boldsymbol{x}}}(\widetilde{\boldsymbol{x}})) = \det(\mathcal{I}_{\boldsymbol{x}}(\boldsymbol{x})) \det(\mathbf{J})^2$.

- Hence, the prior that give the desired "invariance" is:

$$\pi_{\mathrm{pr}}(\boldsymbol{x}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{x}))}, \tag{14}$$

  this is the so-called *multivariate Jeffreys' prior*.

Non-informative: Jeffreys (example 1)

Jeffreys' prior for a Gaussian linear model (with system matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$):

- The log-likelihood of is given by (with identity covariance, $\mathbf{\Sigma}_{\text{obs}} = \mathbf{I}_m$):

$$\log \pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x}) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} (\boldsymbol{y} - \mathbf{G}\boldsymbol{x})^{\mathsf{T}} (\boldsymbol{y} - \mathbf{G}\boldsymbol{x}) . \tag{15}$$

- The score function is:

$$\mathcal{S}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log \pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathbf{G}^{\mathsf{T}}(\boldsymbol{y} - \mathbf{G}\boldsymbol{x}). \tag{16}$$

- And the Fisher information is:

$$\mathcal{I}(\boldsymbol{x}) = -\mathbb{E}_{\boldsymbol{y}}\left[-\mathbf{G}^{\mathsf{T}}\mathbf{G}\right] = \mathbf{G}^{\mathsf{T}}\mathbf{G}. \tag{17}$$

- So the prior is constant and improper, that is, $\pi_{\text{pr}}(\boldsymbol{x}) \propto \sqrt{\det(\mathbf{G}^{\mathsf{T}}\mathbf{G})}$.

Non-informative: Jeffreys (example 2)

- For the Gaussian distribution of the real value x:

$$\pi(x \mid \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{18}$$

  with μ fixed.

- The Jeffreys (hyper)prior for the standard deviation $\sigma > 0$ is

$$\begin{aligned}\pi_{\mathrm{pr}}(\sigma) \propto \sqrt{\mathcal{I}(\sigma)} &= \sqrt{\mathbb{E}_x\left[\left(\frac{\mathrm{d}}{\mathrm{d}\sigma}\log(\pi(x \mid \sigma))\right)^2\right]} \\ &= \sqrt{\mathbb{E}_x\left[\left(\frac{(x-\mu)^2 - \sigma^2}{\sigma^3}\right)^2\right]} = \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma}.\end{aligned} \tag{19}$$

Non-informative: Jeffreys (final comments)

- A valid reason to use this non-informative prior instead of others, is that the probability is independent from the set of parameters that is chosen to describe parameter space.
- Oftentimes the Jeffreys prior cannot be normalized, and thus, it is an improper prior.
- **Reference priors**: by maximizing the Kullback–Leibler divergence (KLD) from prior to posterior, we allow the data to have the maximum effect on the posterior estimates. In practice, it is the expectation of the KLD what is maximized (i.e., the mutual information).

  In 1D, reference priors and Jeffreys priors are equivalent.

Informative: conjugacy (i)

- We say that the prior $\pi_{\text{pr}}(\boldsymbol{x})$ belongs to a class of distributions $\mathcal{P}_1$ and the likelihood $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})$ to a family $\mathcal{P}_2$.

- Then the class $\mathcal{P}_1$ is conjugate for $\mathcal{P}_2$, if the posterior $\pi_{\text{pos}}(\boldsymbol{x} \mid \boldsymbol{y})$ belongs to same class as the prior (i.e. $\mathcal{P}_1$).

Prior ($\mathcal{P}_1$)	Likelihood ($\mathcal{P}_2$)
$x \sim \text{Beta}(a, b)$	$\text{Binom}(n, x)$
$x \sim \text{Gamma}(a, b)$	$\text{Pois}(x)$
$x \sim \mathcal{N}(\mu_{\text{pr}}, \sigma^2_{\text{pr}})$	$\mathcal{N}(x, \sigma^2_{\text{obs}})$ (known σ^2_{obs})
$(1/\sigma^2_{\text{obs}}) \sim \text{Gamma}(a, b)$	$\mathcal{N}(\mu, \sigma^2_{\text{obs}})$ (known μ)
$\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(\mathbf{S}, n)$	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (known $\boldsymbol{\mu}$)
... and many more; check this Link	

Informative: conjugacy (ii)

Conjugate priors offer:

- Analytical tractability — finding the posterior reduces to updating parameters of the prior distribution.
- Assist the validation of numerical techniques for Bayesian inference.
- Mixtures of conjugate priors are also conjugate.
- The main difficulty is that non-conjugate priors appear more often in practice.

Informative: conjugacy (example 1)

Example: inference about the fairness of a coin (from [4]).

- If we denote the bias-weighting by x, then $x = 0$ and $x = 1$ can represent a coin which produces a tail or a head on every flip, respectively.
- There is a continuum of possibilities for the value of x between these limits, with $x = 1/2$ indicating a fair coin.
- Our inference about the fairness of this coin is summarized by the conditional probability

$$\mathbb{P}[x \mid \text{data}] \propto \mathbb{P}[\text{data} \mid x]\ \mathbb{P}[x]\,.$$

  - Prior (beta): $x \sim \text{Beta}(x; a, b) \propto x^{a-1}(1-x)^{b-1}$.
  - Data (binomial): $\mathbb{P}[\text{data} \mid x] \sim \text{Binom}(n_H; n, x) \propto x^{n_H}(1-x)^{n-n_H}$ (n_H heads in n tosses).
  - Posterior: $x \sim \text{Beta}(a + n_H, b + n_T)$.

Informative: conjugacy (example 1)

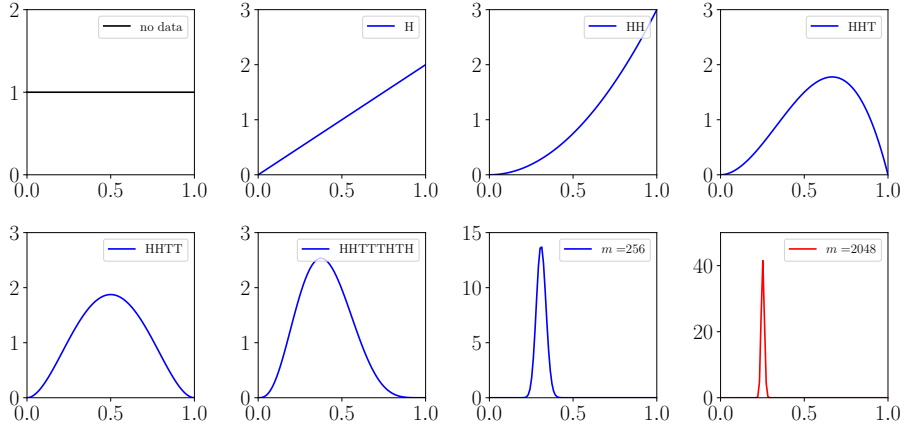

Figure: Evolution of the posterior PDF for the bias-weighting of a coin.

Informative: conjugacy (example 1)

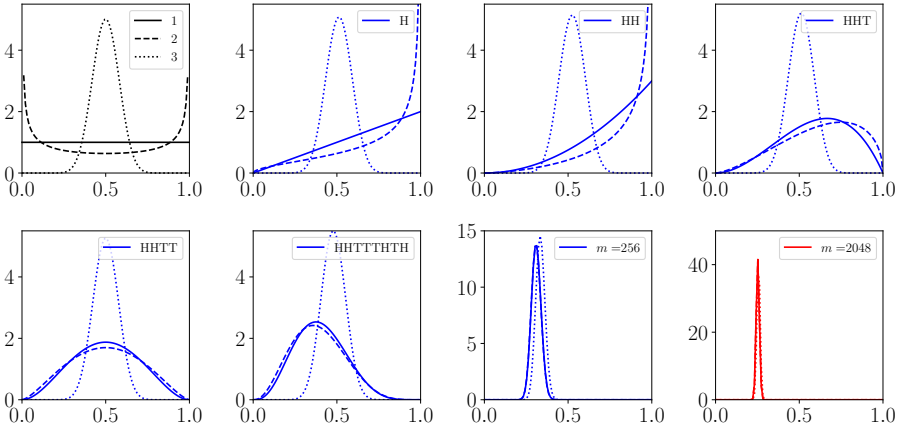

Figure: The effect of different priors on the posterior PDF for the bias-weighting of a coin.

Informative: conjugacy (example 2)

Conjugate prior for a Gaussian linear model (with system matrix $\mathbf{G} \in \mathbb{R}^{d \times d}$):

- If the prior density is Gaussian $\pi_{\text{pr}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\text{pr}}, \boldsymbol{\Sigma}_{\text{pr}})$.
- And the likelihood is also Gaussian $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \mathbf{G}\boldsymbol{x}, \boldsymbol{\Sigma}_{\text{obs}})$.
- Then the posterior is also Gaussian $\pi_{\text{pos}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{\text{pos}}, \boldsymbol{\Sigma}_{\text{pos}})$, with parameters:
  (i) Version 1:

$$\boldsymbol{\Sigma}_{\text{pos}} = \boldsymbol{\Sigma}_{\text{pr}} - \mathbf{C}\mathbf{G}\boldsymbol{\Sigma}_{\text{pr}} \qquad \boldsymbol{\mu}_{\text{pos}}(\boldsymbol{y}) = \boldsymbol{\mu}_{\text{pr}} + \mathbf{C}(\boldsymbol{y} - \mathbf{G}\boldsymbol{\mu}_{\text{pr}}), \tag{20}$$

  where $\mathbf{C} = \boldsymbol{\Sigma}_{\text{pr}}\mathbf{G}^{\mathsf{T}} \left(\mathbf{G}\boldsymbol{\Sigma}_{\text{pr}}\mathbf{G}^{\mathsf{T}} + \boldsymbol{\Sigma}_{\text{obs}}\right)^{-1}$.

  (ii) Version 2:

$$\boldsymbol{\Sigma}_{\text{pos}} = \left(\boldsymbol{\Sigma}_{\text{pr}}^{-1} + \mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}_{\text{obs}}^{-1}\mathbf{G}\right)^{-1} \qquad \boldsymbol{\mu}_{\text{pos}}(\boldsymbol{y}) = \boldsymbol{\Sigma}_{\text{pos}} \left(\mathbf{G}^{\mathsf{T}}\boldsymbol{\Sigma}_{\text{obs}}^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}_{\text{pr}}^{-1}\boldsymbol{\mu}_{\text{pr}}\right). \tag{21}$$

Informative: conjugacy (example 2)

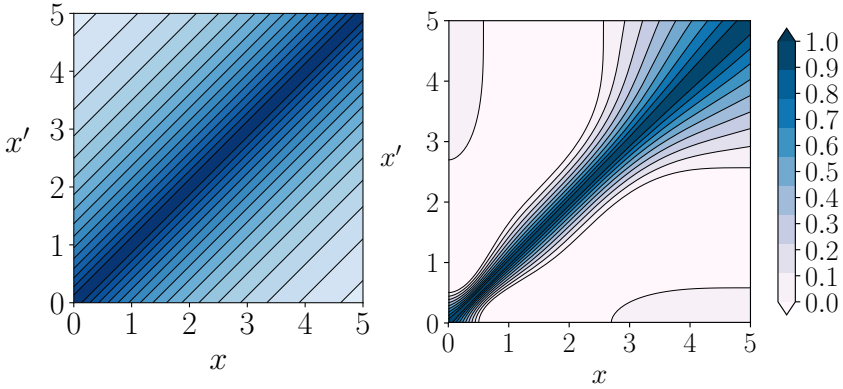

Figure: Prior (left) to posterior (right) covariance update (Matérn covariance in the prior).

Informative: conjugacy (example 3)

The case of Gaussian and gamma distributions:

- If the (hyper)prior density for a precision (inverse variance) parameter δ is gamma $\pi_{\text{hpr}}(\delta) = \text{Gamma}(\delta; \alpha, \beta) \propto \delta^{\alpha-1} \exp(-\beta\delta)$, with shape $\alpha > 0$ and rate $\beta > 0$ parameters.
- And the prior density is Gaussian $\pi_{\text{pr}}(\boldsymbol{x} \mid \delta) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, {}^1\!/\!_\delta\, \mathbf{I}_d) \propto \delta^{d/2} \exp\left(-({}^\delta\!/\!_2)\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\right)$.
- Then the conditional distribution $\pi(\delta \mid \boldsymbol{x})$ is gamma:

$$\pi(\delta \mid \boldsymbol{x}) = \pi_{\text{pr}}(\boldsymbol{x} \mid \delta)\pi_{\text{hpr}}(\delta) = \left[\delta^{d/2} \exp\left(-\frac{\delta}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\right)\right] \left[\delta^{\alpha-1} \exp(-\beta\delta)\right] \tag{22a}$$

$$= \delta^{d/2+\alpha-1} \exp\left(-\left[\frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x} + \beta\right]\delta\right), \tag{22b}$$

which is a gamma distribution with shape $\overline{\alpha} = {}^d\!/\!_2 + \alpha$ and rate $\overline{\beta} = ({}^1\!/\!_2)\boldsymbol{x}^\mathsf{T}\boldsymbol{x} + \beta$.

1b. Prior modeling: hierarchical

Hierarchical models: introduction (i)

- The prior information is rarely rich enough to define a specific prior distribution.
- There is a need for robustness.
- It often helps to decompose prior knowledge into several levels particularly when the available data is hierarchical.
- It is often desirable to reason at various levels: How does prior knowledge guide inferences at lower levels?
- One of the key flexibilities of the Bayesian construction!

Hierarchical models: introduction (ii)

- BIPs involve the construction of more complex posterior distributions when the prior and likelihood depend on unknowns parameters. For instance, if we do not know the scale of the noise, or the mean and variance of the prior.

- We can also model these as random variables such that we can write the hierarchical likelihood and prior as:

$$\pi_{\text{like}}\left(\boldsymbol{y}, \lambda \mid \boldsymbol{x}\right) = \pi_{\text{like}}\left(\boldsymbol{y} \mid \boldsymbol{x}, \lambda\right) \pi_{\text{hpr}}(\lambda) \tag{23a}$$

$$\pi_{\text{pr}}\left(\boldsymbol{x}, \delta\right) = \pi_{\text{pr}}\left(\boldsymbol{x} \mid \delta\right) \pi_{\text{hpr}}(\delta), \tag{23b}$$

  where $\pi_{\text{hpr}}(\lambda), \pi_{\text{hpr}}(\delta)$ are the so-called **hyperprior** probability densities, for the hyperparameters λ and δ.

- We also take into account the uncertainty about the hyperparameters in the Bayesian inverse problem. This approach is oftentimes called *hierarchical regularization*.

Hierarchical models: formulation

- In this case, the Bayesian inverse problem is reformulated as a *hierarchical Bayesian inverse problem*:

$$\begin{aligned}\pi(\boldsymbol{x}, \lambda, \delta) := \pi_{\text{pos}}\left(\boldsymbol{x}, \lambda, \delta \mid \boldsymbol{y}\right) &\propto \pi_{\text{like}}\left(\boldsymbol{y}, \lambda \mid \boldsymbol{x}\right) \pi_{\text{pr}}\left(\boldsymbol{x}, \delta\right) \\ &= \pi_{\text{like}}\left(\boldsymbol{y} \mid \boldsymbol{x}, \lambda\right) \pi_{\text{pr}}(\boldsymbol{x} \mid \delta) \pi_{\text{hpr}}(\lambda) \pi_{\text{hpr}}(\delta). \end{aligned} \tag{24}$$

- We can use conjugate priors in the hyperparameters as discussed previously. That is, use inverse-gamma/gamma hyperpriors for the variance/precision parameters. However, half-Cauchy priors have been advocated for variance parameters [1].

- We often use the Gibbs sampler to solve hierarchical problem. Challenges with potentially "nasty" geometry of the parameter space.

- By increasing the hierarchy, we make the prior more flexible (less influential).

Priors: final comments

- *"The support of the prior should be wide enough to allow the data to speak"*.
- Most formally the prior serves to encode information related to the problem being analyzed, but in practice it often becomes a means of stabilizing inferences in complex, high-dimensional problems.
- If information is available on the values of the parameters or their structural dependence, this should be incorporated in the prior.
- Quite frequently analysts perform sensitivity analysis by perturbing prior hyperparameters and observing the differences in the inferred results.
- Even though conjugate priors are "aesthetically" appealing and convenient, they are generally not applicable in interesting problems.
- Improper priors are allowed but one should be very careful when considering model validation.

2. Bayesian computations: beyond standard inference

Bayesian computations: beyond standard inference

- **Experimental design**: design the experiment such that it maximizes an expected utility of the experiment outcome, e.g., information gain, reduced variance, etc.

- **Data assimilation — filtering, smoothing, prediction**: problems where the data comes sequentially (e.g., weather forecasting). Filtering (present), smoothing (past), prediction (future). Common methods: Kalman filters, sequential importance sampling.

- **Model selection**: the model itself needs to be selected from a predefined collection of models. Each model in the set can have different parameters or can be based on particular mathematical assumptions.
  (i) How good is our model?
  (ii) How does it compare with other models?
  (iii) What model did most likely generate the data we observed?

Model selection

- The joint posterior density over both, model and parameters, is computed using Bayes' theorem

$$\pi_{\text{pos}}(k, \boldsymbol{x}_k \mid \boldsymbol{y}) = \frac{1}{\overline{Z}_{\boldsymbol{y}}} \overline{\pi}_{\text{pr}}(k) \pi_{\text{pr}}(\boldsymbol{x}_k \mid k)\, \pi_{\text{like}}(\boldsymbol{y} \mid k, \boldsymbol{x}_k), \tag{25}$$

  wherein the evidence is given by the theorem of total probability [3]:

$$\overline{Z}_{\boldsymbol{y}} = \sum_{k' \in \mathcal{K}} \overline{\pi}_{\text{pr}}(k') Z_{\boldsymbol{y}}(k'). \tag{26}$$

- The posterior density of the models is obtained by integrating out the parameters in eq. (25)

$$\overline{\pi}_{\text{pos}}(k \mid \boldsymbol{y}) = \frac{\overline{\pi}_{\text{pr}}(k) \int_{\boldsymbol{\Theta}_k} \pi_{\text{pr}}(\boldsymbol{x}_k \mid k)\, \pi_{\text{like}}(\boldsymbol{y} \mid k, \boldsymbol{x}_k)\, \mathrm{d}\boldsymbol{x}_k}{\sum_{k' \in \mathcal{K}} \overline{\pi}_{\text{pr}}(k') \int_{\boldsymbol{\Theta}_{k'}} \pi_{\text{pr}}(\boldsymbol{x}_{k'} \mid k')\, \pi_{\text{like}}(\boldsymbol{y} \mid k', \boldsymbol{x}_{k'})\, \mathrm{d}\boldsymbol{x}_{k'}} = \overline{\pi}_{\text{pr}}(k) \frac{Z_{\boldsymbol{y}}(k)}{\overline{Z}_{\boldsymbol{y}}}. \tag{27}$$

Model selection

In model selection, the posterior eq. (27) can be used to perform the following tasks:

- *Model choice or selection*, which requires the computation of the MAP estimator, $k_{\text{MAP}} = \arg\max_{k \in \mathcal{K}} \overline{\pi}_{\text{pos}}(k \mid \boldsymbol{y})$. Model choice is used as indicator of *model complexity*, i.e., the model that provides the best alignment with the observed data should be preferred over unnecessarily complicated ones (Occam's razor).

- *Model mixing or averaging*, which requires the consideration of the whole collection of parameters weighted by $\overline{\pi}_{\text{pos}}(k \mid \boldsymbol{y})$. The model mixing solution can be seen as a model posterior predictive distribution at the model level.

Beyond inferences: model selection example

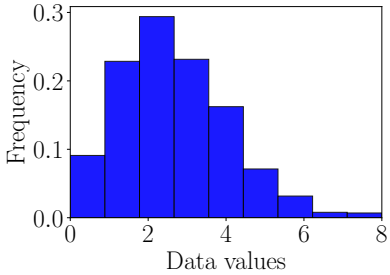

Figure: Someone gives you this data set $\widetilde{\boldsymbol{y}}$ consisting of $m = 1140$ values and ask you about the underlying data generation process, i.e., the likelihood (this Example is after [2]).

Beyond inferences: model selection example

- The objective is to find the likelihood model that yields a better representation of a data set. Two models are considered:

$$\text{(Model 1: Poisson)} \qquad \pi_{\text{like}}(\widetilde{\boldsymbol{y}} \mid \lambda) = \prod_{i=1}^{m} \frac{\lambda^{\tilde{y}_i}}{\widetilde{y}_i!} \exp(-\lambda). \tag{28a}$$

$$\text{(Model 2: Neg. Bin.)} \qquad \pi_{\text{like}}(\widetilde{\boldsymbol{y}} \mid \lambda, \kappa) = \prod_{i=1}^{m} \frac{\lambda^{\tilde{y}_i}}{\widetilde{y}_i!} \frac{\Gamma(1/\kappa + \widetilde{y}_i)}{\Gamma(1/\kappa)(1/\kappa + \lambda)^{\tilde{y}_i}} (1 + \kappa\lambda)^{-1/\kappa}. \tag{28b}$$

- We define the priors as follows (model 1: $x_1 = \lambda$ and model 2: $[x_1, x_2] = [\lambda, \kappa]$):

$$\text{(Gamma)} \qquad \pi_{\text{pr}}(x_1 \mid k = 1) = \mathsf{gamma}(x_1; \alpha_\lambda, \beta_\lambda)$$

$$\text{(Gamma)} \qquad \pi_{\text{pr}}(x_1, x_2 \mid k = 2) = \mathsf{gamma}(x_1; \alpha_\lambda, \beta_\lambda) \times \mathsf{gamma}(x_2; \alpha_\kappa, \beta_\kappa)$$

$$\text{(Discrete uniform)} \qquad \overline{\pi}_{\text{pr}}(k) = \mathsf{Unif}(a, b)$$

Beyond inferences: model selection example

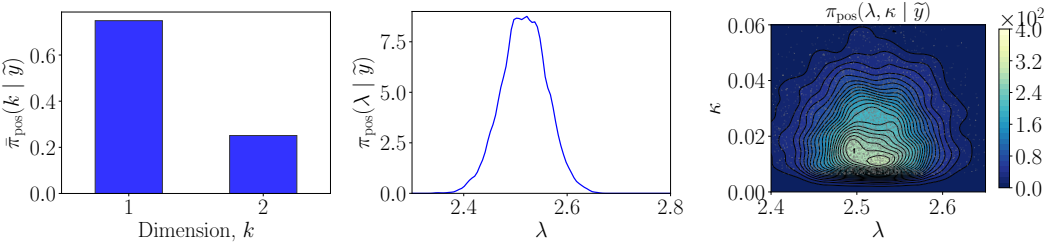

Figure: The histogram of the data set is shown in the 1st row. The posterior of the models and parameters in dimension 1 and 2 are shown in the 2nd row.

References

[1] A. Gelman. "Prior distributions for variance parameters in hierarchical models". In: *Bayesian Analysis* 1.3 (2006), 515—533.

[2] D. I. Hastie et al. "Model choice using reversible jump Markov chain Monte Carlo". In: *Statistica Neerlandica* 66.3 (2012), pp. 309–338.

[3] C. P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. 2nd ed. Springer, 2007.

[4] D. S. Sivia et al. *Data analysis: a Bayesian tutorial*. 2nd ed. Oxford University Press, 2006.