Part III: INVERSE UNCERTAINTY QUANTIFICATION

This is related with 2 tasks:

1. Parameter estimation → Bayesian framework (we look for the distribution of parameters)
2. Data assimilation (time dependent / sequential parameter estimation)

↓

this means that parameters are no longer fixed (and unknown) numbers, but parameters are statistical objects that can be described in terms of distributions. Data will come and will allow us to update our believes on parameters.

parameters are random variables ←

This is the essence of Bayesian framework: we need to assimilate the knowledge from the data (the measurements) to update the prior believes we have on the parameters (to improve the knowledge) to decrease the uncertainty.

Chapter 9

Statistical Inverse Problems and Parameter Estimation

9.1 Basics on Parameter Estimation

We first review some basic notions related with parameter estimation, ranging from the frequentist approach to the Bayesian framework. Before addressing this latter, we recall ordinary least squares and maximum likelihood estimators, applying these notions to the case of systems governed by differential problems.

9.1.1 Estimators, Estimates, and Sampling Distributions

We start by summarizing some basic ideas related with the estimation of unknown parameters through samples, observations or measurements. Let us consider a fixed but unknown parameter $\boldsymbol{\theta} \in \mathcal{P} \subset \mathbb{R}^p$. A point estimate is a vector in $\mathbb{R}^p$ that represents $\boldsymbol{\theta}$. An interval estimate provides an interval that quantifies the plausible location of components of $\boldsymbol{\theta}$. The mean or the median of a sampling distribution are examples of point estimates, whereas confidence intervals are interval estimates.

For the sake of simplicity, let us consider in this section scalar parameters (in dimension $p = 1$). An *estimator* is a rule or procedure that specifies how to construct estimates for θ based on random samples $X_1, \ldots, X_n$. An estimator is a random variable with an associated distribution (sampling distribution), whereas an *estimate* is a realization of the estimator, so it is a function of the realized values $x_1, \ldots, x_n$. An estimator is *unbiased* if its mean is equal to the value of the parameter being estimated, otherwise it is said to be biased. Ordinary least squares and maximum likelihood estimators are two among the most common estimators. Finally, a *statistic* is a measurable function of one or more random variables that does not depend on unknown parameters.

For instance, given $X_1, \ldots, X_n$ random variables associated with a sample of size n, the sample mean and variance

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

are estimators of the population mean μ and variance σ^2. If we additionally assume that $X_i \sim N(\mu, \sigma^2)$, then

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right), \qquad S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1).$$

On the other hand, the goal when constructing an interval estimate is to determine functions $\theta_L(x)$, $\theta_R(x)$ such that $\theta_L(x) < \theta < \theta_R(x)$ of p based on realizations $\mathbf{x} = \{x_1, \ldots, x_n\}$ of a random

sample $\mathbf{X} = \{X_1, \dots, X_n\}$. The random interval $[\theta_L(\mathbf{X}), \theta_R(\mathbf{X})]$ is an *interval estimator*. In combination with a confidence coefficient, an interval estimator is called *confidence interval*. The confidence coefficient can be interpreted as the frequency of times, in repeated sampling, that the interval will contain the target parameter θ. The $(1-\alpha)100\%$ confidence interval is the pair of statistics $(\theta_L(\mathbf{X}), \theta_R(\mathbf{X}))$ such that for all $\theta \in \mathcal{P}$,

$$P\left(\theta \in (\theta_L(\mathbf{X}), \theta_R(\mathbf{X}))\right) = 1 - \alpha.$$

For instance, given a sequence of n random variables $X_1, \dots, X_n$ with $X_i \sim N(\mu, \sigma^2)$ with known variance and unknown mean, since

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

we have that

$$P\left(\mu \in \left(\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha.$$

9.1.2 Ordinary least squares estimator

We recall the basic features of the (ordinary) least squares estimator. Consider the statistical model

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \varepsilon_i, \qquad i = 1, \dots, n \tag{9.1}$$

where Y_i are random variables whose realizations y_i are a set of n measurements from an experiment and $f(\mathbf{x}_i, \boldsymbol{\theta})$ is the parameter-dependent model response or QoI at corresponding locations (e.g., in space or time; we will be more precise later on). The random variables ε_i account for errors between the model and measurements. Finally, $\boldsymbol{\theta}_0$ denotes the true (but unknown) parameter value that we cannot measure directly but instead must infer from realizations of the random variable Y_i. Note that in this context, $\boldsymbol{\theta}_0$ is *not* a random variable.

Let us also assume that the errors ε_i are iid, unbiased ($\mathbb{E}[\varepsilon_i] = 0$) and have true but unknown variance $\mathrm{Var}[\varepsilon_i] = \sigma_0^2$). We assume that the true parameter θ_0 is in an admissible parameter space $\mathcal{P}$. The *ordinary least square (LS) estimator* is

$$\hat{\boldsymbol{\theta}}_{LS} = \arg\min_{\mathbf{q}\in\mathcal{P}} \sum_{i=1}^{n} (Y_i - f(\mathbf{x}_i, \mathbf{q}))^2$$

and the corresponding estimate is the realization in $\mathbb{R}^p$ that minimize the sum of squares errors,

$$\boldsymbol{\theta}_{LS} = \arg\min_{\mathbf{q}\in\mathcal{P}} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \mathbf{q}))^2.$$

The case of a linear model

The simplest case is the one of a linear model with $p+1$ parameters, with

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \ldots + \theta_p x_p.$$

If we have noisy measurements $\mathbf{y} = (y_1, y_2, \dots, y_n)$ obtained at n points $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$ where $i = 1, \dots, n$, the model can be written in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ are the obtained measurements, $\boldsymbol{\theta}_0 = (\theta_{00}, \dots, \theta_{0p})$ is the vector of true (but unknown) parameters and $\mathbf{X}$ is the (known, deterministic) design matrix that contains the measured values for the input variables, augmented with a column of ones to account for the intercept term:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

We assume that there are more measurements than parameters, that is, $n > p + 1$. For linear models, we can derive a direct formula for the OLS estimator; indeed, the OLS estimate that minimizes

$$SS(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2$$

is obtained as the solution to the normal equations

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y} \qquad \Rightarrow \qquad \boldsymbol{\theta}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},$$

that is, the OLS estimator is

$$\hat{\boldsymbol{\theta}}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

of which $\boldsymbol{\theta}_{LS}$ is a realization.

Under the assumption that we have independent and identically distributed measurement noise with measurement error variance σ_0^2, $Cov(\mathbf{Y}) = \sigma_0^2 I$, where I is the identity matrix, we have that

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}_{LS}\right] = \boldsymbol{\theta}_0 \qquad \text{and} \qquad Cov(\hat{\boldsymbol{\theta}}_{LS}) = \sigma_0^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

Indeed, setting $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$,

$$\mathbb{E}\left[\hat{\boldsymbol{\theta}}_{LS}\right] = \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right] = \mathbb{E}\left[\mathbf{A}\mathbf{Y}\right] = \mathbf{A}\mathbb{E}\left[\mathbf{Y}\right] = \mathbf{A}\mathbf{X}\boldsymbol{\theta}_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0$$

and (noting that $\hat{\boldsymbol{\theta}}_{LS} = \mathbf{A}\mathbf{Y} = \mathbf{A}(\mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) = \boldsymbol{\theta}_0 + \mathbf{A}\boldsymbol{\varepsilon}$)

$$Cov(\hat{\boldsymbol{\theta}}_{LS}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{LS} - \boldsymbol{\theta}_0)^\top\right] = \mathbb{E}\left[(\mathbf{A}\boldsymbol{\varepsilon})(\mathbf{A}\boldsymbol{\varepsilon})^\top\right] = \mathbf{A}\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\mathbf{A}^\top = \sigma_0^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

If moreover measurement errors are Gaussian, then $\hat{\boldsymbol{\theta}}_{LS}$ is also Gaussian, being a linear transformation of a Gaussian random variable $\mathbf{Y}$. In this case,

$$\hat{\boldsymbol{\theta}}_{LS} \sim N(\boldsymbol{\theta}_0, \sigma_0^2(\mathbf{X}^\top\mathbf{X})^{-1}).$$

It is also possible to show that the unbiased estimator for the error covariance σ_0^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\mathbf{R}}\hat{\mathbf{R}}^\top$$

where $\hat{\mathbf{R}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}_{LS}$ denotes the residual estimator.

Remark 9.1.1. *Hence, the estimator $\hat{\boldsymbol{\theta}}_{LS}$ has a distribution – sometimes referred to as* sampling distribution *– that can be used to construct confidence intervals for the estimation process, provided that errors are normally distributed. The same results also holds true in the case of sufficiently large samples, so that the central limit theorem can be invoked. For sufficiently large n, with errors that are i.i.d. with variance σ_0^2 fixed (but likely unknown), the sampling distribution for $\hat{\boldsymbol{\theta}}_{LS}$ is asymptotically normal $N(\boldsymbol{\theta}_0, \sigma_0^2(\mathbf{X}^\top\mathbf{X})^{-1})$. In both cases, confidence intervals or regions will be centered at $\boldsymbol{\theta}_{LS}$, and will be based on standard results (CI for the mean with unknown variance).*

9.1.3 Maximum likelihood estimator

Maximum likelihood estimators can also be used to achieve the goal of estimating a parameters vector $\boldsymbol{\theta}$ based on random samples $Y_1, \ldots, Y_n$. Let $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ be a parameter-dependent joint probability density function associated with a random vector $\mathbf{Y} = [Y_1, \ldots, Y_n]$ where $\boldsymbol{\theta} \in \mathcal{P}$ is an unknown parameter vector, and let $\mathbf{y} = [y_1, \ldots, y_n]$ be a realization of $\mathbf{Y}$. The likelihood function $L : \mathcal{P} \to [0, \infty)$ is defined by

$$L_{\mathbf{y}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$$

where the observed sample is fixed and θ varies over all admissible parameter values. For iid variables, it follows that the likelihood function

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i;\boldsymbol{\theta}).$$

Finally, we denote the log-likelihood function by

$$l_{\mathbf{y}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{y}) = \log L(\boldsymbol{\theta}|\mathbf{y}).$$

Estimates for $\boldsymbol{\theta}_0$ are commonly constructed by computing the value of $\boldsymbol{\theta}$ that maximizes the likelihood which is termed a maximum likelihood estimate (MLE). For iid samples, the MLE is

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\mathcal{P}} \prod_{i=1}^{n} f_{Y_i}(y_i;\boldsymbol{\theta}).$$

To illustrate, we consider (9.1) with the assumption that errors are iid, unbiased and normally distributed with true but unknown variance σ_0^2 so that $\varepsilon_i \sim N(0,\sigma_0^2)$ and hence $Y_i \sim N(f(t_i,\boldsymbol{\theta}_0),\sigma_0^2)$. In this case, $\boldsymbol{\theta}$ and σ_0^2 are both parameters, and the likelihood function is

$$L(\boldsymbol{\theta},\sigma^2|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f(t_i,\boldsymbol{\theta}))^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(t_i,\boldsymbol{\theta}))^2\right)$$

The MLE for θ_0 and σ_0^2 is

$$[\boldsymbol{\theta},\sigma^2]_{MLE} = \arg\max_{\mathbf{q}\in\mathcal{P},\sigma^2>0} L(\mathbf{q},\sigma^2|\mathbf{y}).$$

Due to the monotonicity of the logarithm function, maximizing $L(\mathbf{q},\sigma^2|\mathbf{y})$ is equivalent to maximizing the log-likelihood

$$l(\boldsymbol{\theta},\sigma^2|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(t_i,\boldsymbol{\theta}))^2.$$

With the assumption of iid, unbiased normally distributed errors, the maximum likelihood solution $\boldsymbol{\theta}_{MLE}$ to the problem

$$\sum_{i=1}^{n}(y_i - f(t_i,\boldsymbol{\theta}))\nabla f(t_i,\boldsymbol{\theta}) = \mathbf{0}$$

is the same as the least squares estimate $\boldsymbol{\theta}_{LS}$. In frequentist inference, the MLE $\boldsymbol{\theta}_{MLE}$ is the parameter value that makes the observed output most likely; it should not be interpreted as the most likely parameter value resulting from the data since this would require it to be a random variable. But this is what Bayesian statistics, and not frequentist inference, assume...

9.1.4 More on parameter estimation from a frequentist perspective: the case of nonlinear models

For nonlinear models, no such direct methods are available, and one has to resort to numerical methods and different approximations. A possible strategy is to linearize the the nonlinear model and use the linear theory.

Assume that the observed model response or QoI is given by a generic nonlinear function (in the parameters), where χ are independent variables (e.g., time t or spatial coordinates $\mathbf{x}$) or other known inputs, and $\boldsymbol{\theta} \in \mathcal{P}$ is a vector of parameters. The function f generically denotes the map from the independent variables and parameters to the response. We assume that f is fixed and known, in the sense that there exists a unique modeled response; in particular, we can think to the case where $f(\chi,\boldsymbol{\theta}) = Q(u(\chi,\boldsymbol{\theta}))$, being $u = u(\chi,\boldsymbol{\theta})$ the solution of a PDE or ODE system, so that we must rely on numerical approximations for f.

Remark 9.1.2. *Even if the underlying mathematical model is linear in the solution (e.g., a linear elliptic PDE), the map between the parameter vector and the response is usually nonlinear.*

We assume that we have observations (χ_i, y_i), $i = 1, \dots, n$ where the measured quantity of interest y_i is corrupted by measurement errors ε_i so that

$$y_i = f(\chi_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots, n.$$

The mathematical inverse problem associated with parameter estimation is: given these noisy measurements, determine $\boldsymbol{\theta}$ in a stable manner. The associated *statistical inverse problem* (or *inverse UQ problem*) is to additionally quantify uncertainties associated with $\boldsymbol{\theta}$ due to measurement errors. The assumptions required to approximate $\boldsymbol{\theta}$ and quantify its uncertainty define frequentist and Bayesian techniques for parameter estimation.

For sensitivity analysis and uncertainty propagation (or forward UQ) the specific roles of the independent variables are typically of secondary importance and we are interested in how the model solution depends on the input parameters $\boldsymbol{\theta}$. Hence, we represent

$$y_i = f_i(\boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots, n.$$

where $f_i(\boldsymbol{\theta}) \in \mathbb{R}^z$ denotes the observed model response and $y_i \in \mathbb{R}^{n_z}$ denote measured data. The model response can be expressed as $n \times n_z$ vectors, like

- $f(\boldsymbol{\theta}) = [f(t_1, \boldsymbol{\theta}), \dots, f(t_n, \boldsymbol{\theta})]^\top$ for an evolution process; ← if we observe a process at different time steps (e.g. ODE)
- $f(\boldsymbol{\theta}) = [f(x_1, \boldsymbol{\theta}), \dots, f(x_n, \boldsymbol{\theta})]^\top$ for a stationary process;
- $f(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), \dots, f_n(\boldsymbol{\theta})]^\top$ for an algebraic model.

The dependence of the observed model response on the independent variables is thus suppressed in the notation $f(\boldsymbol{\theta})$. For evolution models, we will have $n_z \geq 1$ experimental measurements and model responses at each time t_j, $j = 1, \dots, n$. For stationary processes and algebraic models, we consider scalar measurements and model evaluations, so that $n_z = 1$.

Let us consider the statistical model

$$\mathbf{Y} = f(\boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = [Y_1, \dots, Y_n]^\top$ is a random vector whose realization $\mathbf{y} = [y_1, \dots, y_n]^\top$ is made by measurements from an experiment. Modeling and measurement errors are represented by the random vector $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]^\top$, and are assumed[1] to be unbiased and i.i.d. .

The goal when calibrating models is to determine parameter estimates $\boldsymbol{\theta}$ so that the model response $f(\boldsymbol{\theta})$ fits the data in some optimal sense. This can be achieved by constructing an estimator $\hat{\boldsymbol{\theta}}$ that estimates $\boldsymbol{\theta}_0$ in a statistically reasonable manner. To this goal, OLS estimators

$$\hat{\boldsymbol{\theta}}_{LS} = \arg\min_{\boldsymbol{\theta} \in \mathcal{P}} \sum_{i=1}^{n} (Y_i - f_i(\boldsymbol{\theta}))^2$$

and estimates

$$\boldsymbol{\theta}_{LS} = \arg\min_{\boldsymbol{\theta} \in \mathcal{P}} \sum_{i=1}^{n} (y_i - f_i(\boldsymbol{\theta}))^2$$

can be built, as well as ML estimators – both LS and ML estimators achieve this goal and are equivalent for certain assumptions regarding the distribution of errors ε_i.

[1] Sometimes this might not be true; multiplicative errors might be more appropriate, under the form $Y_i = f_i(\boldsymbol{\theta}_0)(1 + \varepsilon_i)$, $i = 1, \dots, n$, since $\mathrm{Var}[Y_i]$ might depend on the magnitude of $f_i(\boldsymbol{\theta}_0)$.

The former is a nonlinear least square problem. Note that since the estimator $\hat{\boldsymbol{\theta}}$ is a random variable (or vector), it has a mean, covariance, and distribution (the sampling distribution). With appropriate assumptions on the distribution of ε_i, $\mathrm{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}_0$ and the covariance will quantify the variability of the errors; confidence limits for the sampling distribution can be used to quantify the accuracy of the estimation process.

However, the sampling distribution does not provide a distribution for the model parameters since $\boldsymbol{\theta}_0$ is not a random variable in frequentist inference. For certain problems, the sampling distribution coincides with the parameter distribution constructed in a Bayesian framework.

9.1.5 Scalar observations

Let us assume that the model exhibits nonlinear parameter dependencies, summarized by the function $f(\cdot)$. We take $\boldsymbol{\theta} \in \mathbb{R}^p$ and let $\boldsymbol{\theta}_0$ designate the true but unknown parameter that generates the response $\mathbf{y} \in \mathbb{R}^n$.

Moreover, assume that there are more measurements than parameters, so that $n > p$, and let $\mathcal{P}$ denote the admissible parameter space. To construct parameter and error variance estimators, we require ε_i to be iid with zero mean and fixed but unknown variance σ_0^2. With this assumption,

$$\mathrm{E}[Y_i] = f_i(\boldsymbol{\theta}_0), \qquad \mathrm{Var}(Y_i) = \sigma_0^2.$$

The OLS estimate for the scalar case is obtained as

$$\boldsymbol{\theta}_{LS} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \sum_{i=1}^{n} (y_i - f_i(\boldsymbol{\theta}))^2$$

and since $f(\cdot)$ is nonlinear, estimates must be obtained by minimizing the least squares functional numerically, by using e.g. gradient-based methods (like sequential quadratic programming, or Levenberg-Marquardt method). Linearization about $\boldsymbol{\theta}_0$ yields the approximate covariance relation

$$Cov(\hat{\boldsymbol{\theta}}_{LS}) \approx \sigma_0^2 (\mathbf{J}(\boldsymbol{\theta}_0)^\top \mathbf{J}(\boldsymbol{\theta}_0))^{-1} \approx \hat{\sigma}^2 (\mathbf{J}(\boldsymbol{\theta}_{LS})^\top \mathbf{J}(\boldsymbol{\theta}_{LS}))^{-1}$$

where $\mathbf{J}(\boldsymbol{\theta}) \in \mathbb{R}^{n\times p}$ is the Jacobian (or sensitivity) matrix,

$$(\mathbf{J}(\boldsymbol{\theta}))_{ik} = \frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_k}.$$

This matrix can be built by finite difference approximations, solving the sensitivity equations, introducing a suitable adjoint problem, or automatic differentiation. Note that, in the case the PDE (or ODE) solution requires a high number of degrees of freedom, this task features big computational costs.

Since the error variance σ_0^2 is unknown, we can consider the unbiased variance estimator and estimate

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \hat{\mathbf{R}}^\top \hat{\mathbf{R}}, \qquad \sigma^2 = \frac{1}{n-p-1} \mathbf{R}^\top \mathbf{R},$$

where $\hat{\mathbf{R}} = \mathbf{Y} - f(\hat{\boldsymbol{\theta}}_{LS})$ and $\mathbf{R} = \mathbf{y} - f(\boldsymbol{\theta}_{LS})$ are the residual estimator and estimate, respectively. This yields the estimate

$$V = \sigma^2 [\mathbf{J}(\boldsymbol{\theta})^\top \mathbf{J}(\boldsymbol{\theta})]^{-1}$$

for the covariance matrix. To specify a sampling distribution for $\hat{\boldsymbol{\theta}}_{LS}$, we again require either the assumption that errors are i.i.d. and $\varepsilon \sim N(0, \sigma_0^2)$, or that n is sufficiently large that we can invoke the central limit theorem. This directly or asymptotically establishes that

$$\hat{\boldsymbol{\theta}}_{LS} \sim N(\boldsymbol{\theta}_0, \sigma_0^2 [\mathbf{J}(\boldsymbol{\theta}_0)^\top \mathbf{J}(\boldsymbol{\theta}_0)]^{-1})$$

where the covariance matrix is approximated by $\sigma^2 [\mathbf{J}(\boldsymbol{\theta}_0)^\top \mathbf{J}(\boldsymbol{\theta}_0)]^{-1}$.

$\vec{Y} = f(\vec{\theta}_0) + \vec{\varepsilon}$

- **LINEAR MODEL** : (f)

$\hat{\vec{\theta}}_{OLS} = \underset{\vec{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f_i(\vec{\theta}))^2$

$= (\vec{X}^T \vec{X})^{-1} \vec{X}^T \vec{Y}$

$\mathbb{E}[\hat{\vec{\theta}}_{OLS}] = \vec{\theta}_0$,

$cov(\hat{\vec{\theta}}_{OLS}) = \sigma_0^2 (\vec{X}^T \vec{X})^{-1}$

- **NONLINEAR MODEL** (f)

$\hat{\vec{\theta}}_{OLS} = \underset{\vec{\theta}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f_i(\vec{\theta}))^2$

$\mathbb{E}[\hat{\vec{\theta}}_{OLS}] = \vec{\theta}_0$

$cov(\hat{\vec{\theta}}_{OLS}) = \hat{\sigma}_0^2 (J(\vec{\theta}_0)^T J(\vec{\theta}_0))^{-1}$

the covariance is related with the linearization

Remark 9.1.3. *Indeed, if we consider a Taylor expansion of*

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - f_i(\boldsymbol{\theta}))^2$$

about $\boldsymbol{\theta} = \boldsymbol{\theta}_0$*, we have*

$$\mathcal{J}(\boldsymbol{\theta}) \approx \mathcal{J}(\boldsymbol{\theta}_0) + \nabla \mathcal{J}(\boldsymbol{\theta})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{H}(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

where the elements of the Hessian matrix H *are*

$$(\mathbf{H}(\boldsymbol{\theta}_0))_{pq} = \frac{\partial^2 \mathcal{J}(\boldsymbol{\theta}_0)}{\partial \theta_p \partial \theta_q} = 2 \sum_{i=1}^{n} \frac{\partial f_i(\boldsymbol{\theta}_0)}{\partial \theta_p} \frac{\partial f_i(\boldsymbol{\theta}_0)}{\partial \theta_q} + 2 \sum_{i=1}^{n} (f_i(\boldsymbol{\theta}_0) - y_i) \frac{\partial^2 f_i(\boldsymbol{\theta}_0)}{\partial \theta_p \partial \theta_q}.$$

Assuming that the residuals $f_i(\boldsymbol{\theta}_0) - y_i$ *are small, we can approximate*

$$(\mathbf{H}(\boldsymbol{\theta}_0))_{pq} \approx 2 \sum_{i=1}^{n} \frac{\partial f_i(\boldsymbol{\theta}_0)}{\partial \theta_p} \frac{\partial f_i(\boldsymbol{\theta}_0)}{\partial \theta_q} \qquad \Rightarrow \qquad \mathbf{H}(\boldsymbol{\theta}_0) \approx 2\mathbf{J}^\top(\boldsymbol{\theta}_0)\mathbf{J}(\boldsymbol{\theta}_0).$$

A similar Taylor expansion in the case of a linear model would give (note the equality)

$$\mathcal{J}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta}_0) + \nabla \mathcal{J}(\boldsymbol{\theta})^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

so that, in the linear approximation, the Jacobian matrix $\mathbf{J}$ *assumes the role of the design matrix* $\mathbf{X}$ *of the linear case, so that - recalling that in the linear case* $Cov(\hat{\boldsymbol{\theta}}_{LS}) = \sigma_0^2 (\mathbf{X}^\top \mathbf{X})^{-1}$*,*

$$Cov(\hat{\boldsymbol{\theta}}_{LS}) = \sigma_0^2 (\mathbf{J}(\boldsymbol{\theta}_0)^\top \mathbf{J}(\boldsymbol{\theta}_0))^{-1}.$$

Remark 9.1.4. *Numerical approximations of the Jacobian matrix are usually computed, using, for instance, the centered finite difference formula, yielding*

$$(\mathbf{J}(\boldsymbol{\theta}))_{ik} = \frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_k} \approx \frac{f_i(\boldsymbol{\theta} + h\mathbf{e}_k) - f_i(\boldsymbol{\theta} - h\mathbf{e}_k)}{2h}$$

where $h > 0$ *is a small constant added to the* k*-th component of the* $\boldsymbol{\theta}$ *vector.*

Example: a spring model

Consider the spring model

$$\begin{cases} u'' + Cu' + Ku = 0, & t > 0 \\ u(0) = 2 \\ u'(0) = -C \end{cases}$$

with displacement observation, so that

$$z = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} u \\ u' \end{pmatrix} = u.$$

The previous problem admits the solution

$$u(t) = 2e^{-Ct/2} \cos(\sqrt{K - C^2/4}\, t)$$

provided that $C^2 - 4K < 0$. We take $K = 20.5$ to be known and let $\theta = C$ be the parameter considered in the statistical analysis. Note that the dependence of $u(t, \theta)$ on θ is nonlinear.

To numerically generate synthetic data, we employ $C_0 = 1.5$ and add noise $\varepsilon \sim N(0, \sigma_0^2)$ where $\sigma_0 = 0.1$; the model and one realization of the data at $n = 501$ points are reported in Figure 9.1 (a); residuals are plotted in Figure 9.1 (b). The $n \times 1$ sensitivity matrix is

$$\mathbf{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial z}{\partial C}(t_1, \boldsymbol{\theta}) \\ \cdots \\ \frac{\partial z}{\partial C}(t_n, \boldsymbol{\theta}) \end{pmatrix}, \quad \frac{\partial z}{\partial C}(t, \boldsymbol{\theta}) = e^{-Ct/2} \left[\frac{Ct}{\sqrt{4K - C^2}} \sin(\sqrt{K - C^2/4}\ t) - t \cos(\sqrt{K - C^2/4}\ t) \right].$$

Since in this case we assume σ_0^2 to be known, we obtain the covariance value

$$\sigma_C^2 = \sigma_0^2 [\mathbf{J}(\boldsymbol{\theta})^\top \mathbf{J}(\boldsymbol{\theta})]^{-1} = 3.35 \cdot 10^{-4}$$

so that $\sigma_C = 0.0183$. Since $\varepsilon_i \sim N(0, \sigma_0^2)$, the random variable $\hat{\theta} = \hat{C}$ has the sampling distribution $\hat{C} \sim N(C_0, \sigma_C^2)$, which is reported in Figure 9.2. The OLS estimate for the provided data is $C = 1.4792$ and the 95% confidence interval is $[1.4433, 1.5150]$.

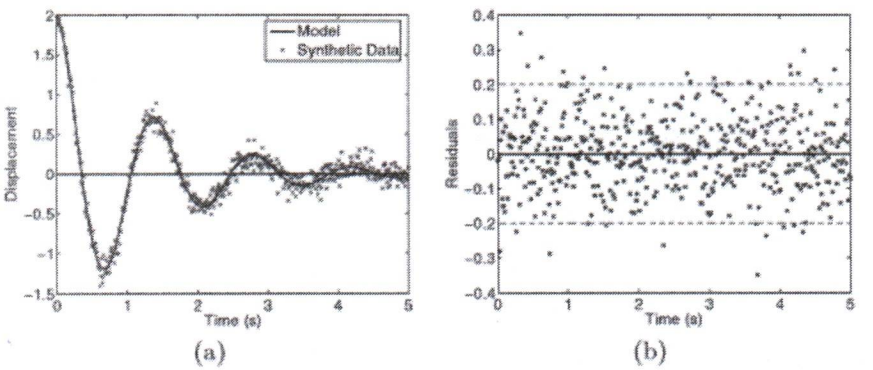


Figure 9.1: Synthetic data and modeled displacement (left) and residuals at $n = 501$ points (right). Taken from [45].

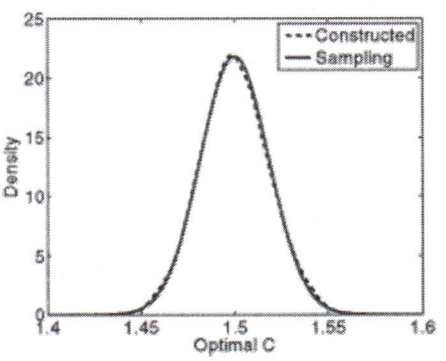


Figure 9.2: Sampling density $N(C_0, \sigma_C^2)$ for $\hat{C}$ and density constructed from 10 000 simulations. Taken from [45].

9.1.6 Multiple responses*

We can also generalize the estimation process to the case of $n_z > 1$ data measurements. The statistical model in this case is

$$\mathbf{Z}_i = \mathbf{f}_i(\boldsymbol{\theta}_0) + \boldsymbol{\varepsilon}_i, \qquad i = 1, \dots, n$$

where $\mathbf{Z}_i$ and $\boldsymbol{\varepsilon}_i$ are random n_z-vectors and each $\mathbf{f}_i : \mathbb{R}^p \to \mathbb{R}^{n_z}$. To take into account the possibility that error distributions associated with individual components of the observations could differ, we let $\sigma^2_{0_j}$ denote the fixed but unknown variance of the error associated with he j-th observation. These values are listed in the $n_z \times n_z$ diagonal measurement error covariance matrix $\mathbf{V}_0 = diag(\sigma^2_{0_1}, \dots, \sigma^2_{0_{n_z}})$. Errors, as before, are assumed to be unbiased; the matrix $\mathbf{V}_0$ is fixed but typically unknown.

The construction of parameter (and covariance) estimators is similar to the scalar case $n_z = 1$, but is more involved due to the fact that variances of the error components are typically different. In this case, the OLS estimator and estimate are given by

$$\hat{\boldsymbol{\theta}}_{LS} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \sum_{i=1}^{n} (\mathbf{Z}_i - \mathbf{f}_i(\boldsymbol{\theta}))^\top \mathbf{V}_0^{-1} (\mathbf{Z}_i - \mathbf{f}_i(\boldsymbol{\theta})),$$

$$\boldsymbol{\theta}_{LS} = \arg\min_{\boldsymbol{\theta}\in\mathcal{P}} \sum_{i=1}^{n} (\mathbf{z}_i - \mathbf{f}_i(\boldsymbol{\theta}))^\top \mathbf{V}_0^{-1} (\mathbf{z}_i - \mathbf{f}_i(\boldsymbol{\theta}))$$

where $\mathbf{V}_0^{-1}$ weights the response components by the reciprocals of the corresponding error variance associated with each component. The estimate of this matrix is provided by

$$\mathbf{V} = \text{diag}\left(\frac{1}{n-p} \sum_{i=1}^{n} (\mathbf{z}_i - \mathbf{f}_i(\boldsymbol{\theta}_{LS}))(\mathbf{z}_i - \mathbf{f}_i(\boldsymbol{\theta}_{LS}))^\top \right) \in \mathbb{R}^{n_z \times n_z}$$

Note that, unlike the scalar case, parameter and covariance estimates must be found by solving a unique, larger problem.

To specify a sampling distribution, we need to assume that the error in the j-th component of $\mathbf{Z}_i$ is $\varepsilon_{ij} \sim N(0, \sigma^2_{0_j})$, so that $\boldsymbol{\varepsilon}_i \sim N(0, V_0)$. For n sufficiently large, the central limit theorem can be invoked to state that the sampling distribution for $\hat{\boldsymbol{\theta}}_{LS}$ is asymptotically normal, so that

$$\hat{\boldsymbol{\theta}}_{LS} \sim N(\boldsymbol{\theta}_0, \mathcal{V}_0) \approx N(\boldsymbol{\theta}_{LS}, \mathcal{V})$$

where

$$\mathcal{V}_0 \approx \left(\sum_{j=1}^{n} \mathbf{J}_j^\top(\boldsymbol{\theta}_0) \mathbf{V}_0^{-1} \mathbf{J}_j(\boldsymbol{\theta}_0) \right)^{-1}$$

is the $p \times p$ covariance matrix and

$$\mathbf{J}_j(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f_j^{(1)}(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial f_j^{(1)}(\boldsymbol{\theta})}{\partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial f_j^{(n_z)}(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial f_j^{(n_z)}(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix}$$

is the $n_z \times p$ sensitivity matrix related with the j-th observation. For ease of implementation, $\mathcal{V}_0$ is approximated by

$$\mathcal{V} = \left(\sum_{j=1}^{n} \mathbf{J}_j^\top(\boldsymbol{\theta}_{LS}) \mathbf{V}^{-1} \mathbf{J}_j(\boldsymbol{\theta}_{LS}) \right)^{-1}$$

where the sensitivity matrix must be evaluated at each time step.

in the **frequentist** approach the parameter is unknown but fixed. The estimator of the parameter is a random variable (since it's based on X which is random). In the **Bayesian** framework the parameter is itself a random variable and so, the goal is to estimate its distribution (starting from a prior and updated through observations)

9.2 A Bayesian framework for inverse UQ

The goal in statistical inference is to deduce the structure of (or make conclusion about) a phenomenon based on observed data. This often involves the determination of an unknown distribution based on observed data in which case the problem of statistical inference can be stated as follows: given a set $S = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, $\mathbf{z}_j \in \mathbb{R}^{n_z}$ of observed realizations of a random variable $\mathbf{Z}$, we want to infer the underlying probability distribution that produces the data S. In particular, we assume to deal with parametric[2] statistical inference.

9.2.1 Introduction: a crash course in Bayesian Statistics

Frequentist and Bayesian inference differ in the underlying assumption made regarding the nature of probabilities, models and parameters:

- From a frequentist perspective, probabilities are defined as the frequencies with which an event occurs if the experiment is repeated a large number of times. Hence, they are objective and are not updated as data is acquired. *Parameters are considered to be unknown but fixed, so they are deterministic.* To establish, from a statistical point of view, confidence in the estimation process, we have seen in the previous section that we can construct estimators (such as LS or ML estimators) and that, based on either the assumption of normality for the errors or asymptotic theory (e.g., central limit theorem), we can then construct sampling distributions and confidence intervals for the parameter estimators. Because parameters are fixed but unknown values in this framework, this latter cannot be applied to obtain parameter densities that can be propagated through models to quantify model uncertainty.
- Probabilities are treated as possibly subjective in the Bayesian framework, and they can be updated to reflect new information. Moreover, *parameters are considered to be random variables with associated density, and the solution of the parameter estimation problem is a probability density (the posterior distribution).* The Bayesian perspective is thus natural for model uncertainty quantification since it provides densities that can be propagated through models, as we did when we dealt with forward UQ.

> Bayesian inference is based on the assumption that probabilities, and more generally speaking our state of knowledge regarding an observed phenomenon, can be updated as additional information is obtained. In the context of parametric models, parameters are treated as random variables having associated densities.

Bayes' formula

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

for probabilities provides a natural way for introducing Bayesian inference. In the context of parameters[3] $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ that are quantified based on observations $\mathbf{z} = (z_1, \dots, z_n)$ we can employ the relation

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{\pi(\mathbf{z}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{\pi_Z(\mathbf{z})}$$

where $\pi_0(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{z})$ denote, respectively, the prior and posterior densities, $\pi(\mathbf{z}|\boldsymbol{\theta})$ is a likelihood, and the marginal density $\pi_Z(\mathbf{z})$ is a normalization factor. In particular:

[2] Statistical inference can be roughly categorized as being *parametric* or *non-parametric*. In the former case, one assumes that the underlying distributions can be adequately described in terms of a parametric relation having a relatively small number of parameters (e.g., mean and variance). The inference problem is to estimate those parameters or the distribution of those parameters. This approach often features a small number of parameters, but might suffer of limited accuracy if the assumed functional relation is incorrect.

[3] For ease of notation, often we avoid capital letters to denote parameters, even if they are now random variables.

- the prior density $\pi_0(\boldsymbol{\theta})$ reflects any prior knowledge that may be known about the parameter before data is taken into account (e.g., previous similar models, related experiments, literature, ...). Non informative priors can otherwise be used, such as the uniform density or unnormalized uniform, posed on the parameter support. For instance, one might employ $\pi_0(\theta) = \mathbb{I}_{[0,+\infty)}(\theta)$ for a positive parameter; this choice is *improper* in the sense that the integral of $\pi_0(\theta)$ is unbounded. Another option is to use data-dependent priors, estimated using frequentist techniques such as maximum likelihood estimators;

- the term $\pi(\mathbf{z}|\boldsymbol{\theta})$, which is a function of $\boldsymbol{\theta}$ with $\mathbf{z}$ fixed, quantifies the likelihood $L(\boldsymbol{\theta}|\mathbf{z})$ of observing $\mathbf{z}$ given parameter realizations $\boldsymbol{\theta}$. The likelihood function is the probability of observing the data at hand under the assumption that $\mathbf{Z} \sim \pi(\mathbf{z} \mid \boldsymbol{\theta})$. Seen as function of the parameters, the likelihood function is defined by

$$\boldsymbol{\theta} \mapsto \pi(\mathbf{z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(z_i \mid \boldsymbol{\theta})$$

  (measured observations $(z_1, \ldots, z_N)$ are viewed as independent realizations of $Z \sim \pi(z \mid \boldsymbol{\theta})$);

- the joint density is given by $\pi(\mathbf{z}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})$ and is normalized to unity by the marginal density function $\pi_Z(\mathbf{z})$, that plays the role of normalizing factor;

- the posterior density $\pi(\boldsymbol{\theta}|\mathbf{z})$ quantifies the probability of obtaining parameters θ given observations $\mathbf{z}$. It is the posterior density that we will be estimating using the Bayesian parameter estimation technique in the following. Note that data directly informs the posterior only through the likelihood.

Representing $\pi_Z(\mathbf{z})$ as the integral over all possible joint densities yields the Bayes relation

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{\pi(\mathbf{z}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{\int_{\mathcal{P}} \pi(\mathbf{z}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

commonly employed for parameter estimation (model calibration) and data assimilation. Note that the normalizing integral can be analytically evaluated only in special cases, and classical quadrature techniques are effective only in low dimension, say, with $p \leq 4$.

Remark 9.2.1. *In particular, we will assume that model inputs are given by vectors of parameters, and that the underlying state system can be given by a suitable discretization of a differential problem; to provide more practical tools, we do not take into account, for instance, Bayesian inversion in infinite-dimensional spaces, as done instead in [46] and [47].*

- **Example 9.2.1.** *Let us consider a case where the posterior density can be computed explicitly. Consider the results from tossing a possibly biased coin. The random variable*

$$Z_i(\omega) = \begin{cases} 0, & \omega = T \\ 1, & \omega = H \end{cases}$$

represents the result from the i-th toss, and the parameter θ is the probability of getting heads. Let us now consider the probability of getting N_1 heads and N_0 tails in a series of $N = N_0 + N_1$ flips of the coin. Since coin flips are independent events, with only two possible outcomes, the likelihood of observing a sequence $\mathbf{z} = [z_1, \ldots, z_N]$ given the probability θ, is

$$\pi(\mathbf{z}|\theta) = \prod_{i=1}^{N} \theta^{z_i}(1-q)^{1-z_i} = \theta^{\sum_i z_i}(1-\theta)^{N-\sum_i z_i} = \theta^{N_1}(1-\theta)^{N_0}$$

that is, a (scaled) binomial density. We consider first a noninformative prior

$$\pi_0(\theta) = \begin{cases} 1, & \theta \in [0,1] \\ 0, & else \end{cases}$$

which yields the posterior density

$$\pi(\theta|\mathbf{z}) = \frac{\theta^{N_1}(1-\theta)^{N_0}}{\int_0^1 \theta^{N_1}(1-\theta)^{N_0} d\theta} = \frac{(N+1)!}{N_0!N_1!}\theta^{N_1}(1-\theta)^{N_0}.$$

In this case, the denominator is the integral of a beta function which admits an analytic solution; in general, quadrature formulas must be used to approximate the integral.

For a fair coin with $\theta_0 = 1/2$, the posterior densities associated with various realizations N_1 and N_0 are reported below. Note that the variability of $\pi(\theta|\mathbf{z})$ decreases as N increases.

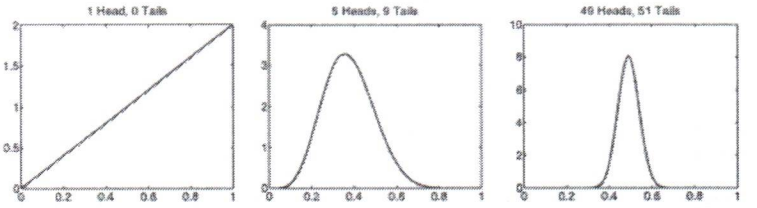


Figure 9.3: Posterior densities associated with a noninformative prior, coin toss experiment. From left to right: 1 Head, 0 Tails; 5 Heads, 9 Tails; 49 Heads, 51 Tails.

In the case, for the same fair coin with $\theta_0 = 1/2$, we had chosen a poor *prior density, for instance*

$$\pi_0(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

with $\mu = 0.3$ and $\sigma = 0.1$, even for a realization of 50 heads and 50 tails, the mode of the posterior is still smaller than $\theta_0 = 1/2$. In this case, Gaussian quadrature formulas have been used to evaluate numerically the denominator. Hence, rather than using a non valid informative prior, it is better to use a noninformative prior.

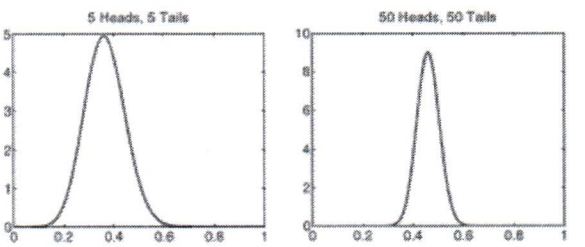


Figure 9.4: Posterior densities associated with a poor informative prior, coin toss experiment. Left: 5 Heads, 5 Tails; Right: 50 Heads, 50 Tails.

The property that the prior and the posterior distributions have the same parametric form is called *conjugacy*. When this occurs, the prior $\pi_0(\boldsymbol{\theta})$ is termed a conjugate prior for the likelihood $\pi(\mathbf{z}|\boldsymbol{\theta})$. Parameters in the prior density are often referred to as prior *hyper-parameters* to distinguish them from the model parameter $\boldsymbol{\theta}$. The corresponding parameters in the posterior density are called posterior *hyper-parameters*. The use of conjugate priors, when possible, is advantageous since closed-form expressions for the posterior are then available.

- **Example 9.2.2.** *Assume that we are interested in forecasting the value of a scalar state variable x which could be a temperature. The prior is $x \sim N(\mu_X, \sigma_x^2)$, which could come from a forecast model; that is,*

$$\pi_0(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}(x-\mu_X)^2\right).$$

Suppose we haven independent, noisy observations

$$\mathbf{z} = [z_1, \ldots . z_n]$$

each with conditional distribution $z_i|x \sim N(x, \sigma^2)$ that are conditioned on the true value of the process x. In other words, the likelihood of observing $\mathbf{z} = [z_1, \ldots . z_n]$ iid measurements under these assumptions is

$$\pi(\mathbf{z}|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z_i - x)^2\right) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - x)^2\right).$$

From Bayes' law,

$$\pi(x|\mathbf{z}) \propto \pi(\mathbf{z}|x)\pi_0(x)$$

so that, using the data and the prior distributions, we have

$$\begin{aligned}\pi(x|\mathbf{z}) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(z_i-x)^2}{\sigma^2} + \frac{(x-\mu_X)^2}{\sigma_X^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(x^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_X^2}\right) - 2\left(\sum_{i=1}^{n}\frac{z_i}{\sigma^2} + \frac{\mu_X}{\sigma_X^2}\right)x\right)\right)\end{aligned}$$

Therefore, we have obtained the product of two Gaussians, which, by completing the square, can be shown to be Gaussian itself. This produces the posterior distribution

$$x|\mathbf{z} \sim N(\mu_{x|\mathbf{z}}, \sigma_{x|\mathbf{z}}^2) \tag{9.2}$$

where

$$\mu_{x|\mathbf{z}} = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_X^2}\right)^{-1}\left(\sum_{i=1}^{n}\frac{z_i}{\sigma^2} + \frac{\mu_X}{\sigma_X^2}\right), \qquad \sigma_{x|\mathbf{z}}^2 = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_X^2}\right)^{-1}.$$

We first remark that the inverse of the posterior variance, (also called the posterior precision*), is equal to the sum of the prior precision, $1/\sigma_X^2$, and the data precision, n/σ^2. Second, the posterior mean, or conditional expectation, can also be written as a sum of two terms:*

$$\mathrm{E}[x|\mathbf{z}] = \frac{\sigma^2\sigma_X^2}{\sigma^2 + n\sigma_X^2}\left(\frac{n}{\sigma^2}\bar{z}_n + \frac{\mu_X}{\sigma_X^2}\right) = w_z\bar{z}_n + w_{\mu_X}\mu_X$$

where the sample mean,

$$\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$$

and the two weights,

$$w_z = \frac{n\sigma_X^2}{\sigma^2 + n\sigma_X^2}, \qquad w_{\mu_X} = \frac{\sigma^2}{\sigma^2 + n\sigma_X^2}$$

add up to $w_z + w_{\mu_X} = 1$. We observe immediately that the posterior mean is the weighted sum/average of the data mean ($\bar{z}_n$) and the prior mean (μ_X). Now let us examine the weights themselves. In particular:

- *If there is a large uncertainty in the prior, then $\sigma_X^2 \to \infty$ and hence $w_z \to 1$, $w_{\mu_X} \to 0$ and the likelihood dominates the prior, leading to what is known as the sampling distribution for the posterior:*

$$\pi(x|\mathbf{z}) \to N(\bar{z}_n, \sigma^2/n).$$

- *If we have a large number of observations, then $n \to \infty$ and the posterior now tends to the sample mean, whereas if we have few observations, then $n \to 0$ and the posterior*

$$\pi(x|\mathbf{z}) \to N(\mu_X, \sigma_X^2).$$

tends to the prior. In the case of equal uncertainties between data and prior, $\sigma^2 = \sigma_X^2$, and the prior mean has the weight of a single additional observation. Finally, if the uncertainties are small, either the prior is infinitely more precise than the data ($\sigma_X^2 \to 0$) or the data are perfectly precise ($\sigma^2 \to 0$).

Note that in this case we can rewrite the posterior mean and variance in a special form. Let us start with the mean:

$$\mathrm{E}[x|\mathbf{z}] = \mu_X + \frac{n\sigma_X^2}{\sigma^2 + n\sigma_X^2}(\bar{z}_n - \mu_X) = \mu_X + G(\bar{z}_n - \mu_X)$$

In other words, the prior mean μ_X is adjusted toward the sample mean $\bar{z}_n$ by a gain (or amplification factor) of $G = 1/(1 + \sigma^2/n\sigma_X^2)$, multiplied by the innovation $\bar{z}_n - \mu_X$, and we observe that the variance ratio, between data and prior, plays an essential role.
In the same way, the posterior variance can be reformulated as

$$\sigma_{x|\mathbf{z}}^2 = (1 - G)\sigma_X^2,$$

and the posterior variance is thus updated from the prior variance according to the same gain G. These two equations are particular instances of updating *occurring in* data assimilation, *since they express the interplay between prior and data, and their effect on the posterior.*

Let us illustrate this with two initial numerical examples. Suppose we have a prior distribution $x \sim N(\mu_X, \sigma_X^2)$ with mean 20 and variance 3. Suppose that our data model has the conditional law $z_i|x \sim N(x, \sigma^2)$ with variance 1. Here the data are relatively precise compared to the prior. Say we have acquired two observations, $\mathbf{y} = (19, 23)^\top$. Now we can compute the posterior distribution:

$$\mathrm{E}[x|\mathbf{z}] = 20 + \frac{6}{1+6}(21 - 20) = 20.86, \qquad \sigma_{x|\mathbf{z}}^2 = (1 - 6/7)3 = 0.43,$$

thus yielding the posterior distribution $z_i|x \sim N(20.86, 0.43)$, which represents the update of the prior according to the observations and takes into account all the uncertainties available (see the Figure 9.5, left). In other words, we have obtained a complete forecast at a given point in time.

Now consider the same prior, $x \sim N(20, 3)$, but with a relatively uncertain/imprecise observation model, $z_i|x \sim N(x, 10)$, and the same two measurements, $\mathbf{z} = (19, 23)^\top$. Redoing the above calculations, we now find

$$\mathrm{E}[x|\mathbf{z}] = 20 + \frac{6}{10+6}(21 - 20) = 20.375, \qquad \sigma_{x|\mathbf{z}}^2 = (1 - 6/16)3 = 1.875,$$

thus yielding the new posterior distribution, $z_i|x \sim N(20.375, 1.875)$, which has virtually the same mean but a much larger variance (see Figure 9.5, right, where the scales on both axes have changed).

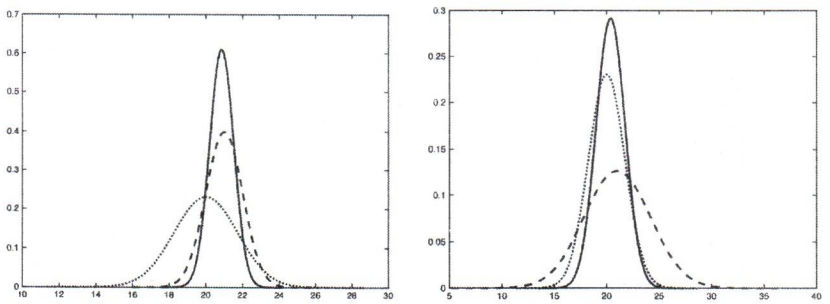


Figure 9.5: Scalar Gaussian distribution example. Left: prior $N(20,3)$ (dotted), instrument $N(x,1)$ (dashed), and posterior $N(20.86,0.43)$ (solid) distributions. Right: prior $N(20,3)$ (dotted), instrument $N(x,10)$ (dashed), and posterior $N(20.375,1.875)$ (solid) distributions.

- **Example 9.2.3.** *We present an example of a simple mechanical system and seek an estimation of its parameters from noisy measurements. A simple gravity pendulum is an idealized mathematical model of a real pendulum; this latter is a weight on the end of a massless cord suspended from a pivot, without friction. Under suitable assumptions (e.g., no frictional energy loss, inextensible massless cord, motion occurring in two dimensions, ...), the differential equation which represents the motion of a simple pendulum is*

$$\frac{d^2x}{dt^2} + \frac{g}{L}\sin x = 0$$

where g is acceleration due to gravity, L is the length of the pendulum, and x is the angular displacement. Adding a restriction to the size of the oscillation's amplitude gives a form whose solution can be easily obtained.

If it is assumed that the angle is $x \ll 1$, then substituting for $\sin x$ into the previous equation and using the small-angle approximation, $\sin x \approx x$, yields the equation for a harmonic oscillator,

$$\frac{d^2x}{dt^2} + \frac{g}{\ell}x = 0.$$

Given the initial conditions $x(0) = x_0$ and $x'(0) = 0$, the solution becomes

$$x(t) = x_0 \cos\left(\sqrt{\frac{g}{L}}\, t\right) \qquad x_0 \ll 1.$$

Hence, we consider a model for the angular displacement, x_t , of an ideal pendulum (no friction, no drag), with $x_0 = 1$,

$$x_t = \cos(\theta t) + \varepsilon_t,$$

where ε_t is a Gaussian noise with zero mean and variance σ^2, the pendulum parameter is denoted by θ, and t is time.

From these noisy measurements (suppose that the instrument is not very accurate) of x_t we want to estimate θ, which represents the physical properties of the pendulum – in fact $\theta = \sqrt{g/L}$. Using this physical model, can we estimate (or infer) the unknown physical parameters of the pendulum?

If the measurements are independent, then the likelihood of a set of T observations $x_1, \ldots, x_T$ is given by the product

$$\pi(x_1, \ldots, x_T|\theta) = \prod_{i=1}^{T} p(x_t|\theta).$$

data = position of the pendulum over time, in this particular case we know that these data are of the form:
$x_t = x_0 \cos(\theta t)$
what is θ?
The parameter we're trying to estimate
(here it is:
$y = f(\theta) + \varepsilon$
where
$f(\theta) = x_0 \cos(\theta t)$)

Notice that $f(\theta)$ is a model that links the parameter(s) to the observations. Here the situation is good because even if f is non linear, we still have the form (explicit) of f. However, f could be anything: f could also require the solution of a differential problem (suppose $f(\theta) = Q(u(\theta))$) and in that case each evaluation of the likelihood would be extremely **expensive** since for every evaluation we would have to solve a diff. problem.
(BTW, this is why surrogate models are important)

In addition, suppose that we have some prior estimation (before obtaining the measurements) of the probabilities of a set of possible values of θ. Then the posterior distribution of θ, given the measurements, can be calculated from Bayes' law, as seen above,

$$\pi(\theta|x_1 \ldots, x_T) \propto \pi_0(\theta) \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_t - \cos(\theta t))^2\right)$$

where we have omitted the denominator.

We are given the (discrete) prior distribution in Figure 9.6, center. After performing numerical simulations, we observe (see Figure 9.6, right) that the posterior for θ develops a prominent peak for a large number ($T = 100$) of measurements, centered around the real value $\theta = 0.2$ (which was used to generate the time series, x_t).

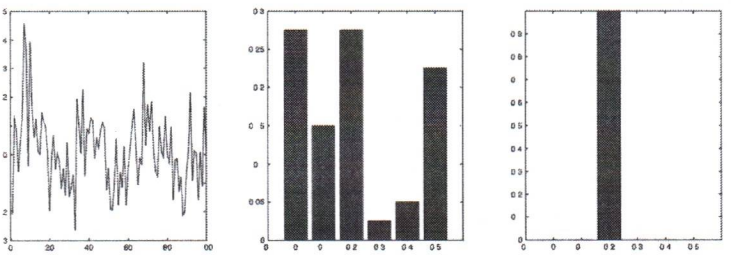


Figure 9.6: Bayesian estimation of noisy pendulum parameter, $\theta = 0.2$. Observations of 100 noisy positions (left). Prior distribution of parameter values (center). Posterior distribution for θ (right).

Example 9.2.4. (Example 1 revisited). *As a final case that will lead us naturally to the case of* filtering methods, *let us consider the vector/multivariate extension of the example 1. We will now study a vector process, $\mathbf{x}$, with n components and a prior distribution*

$$\mathbf{x} \sim N(\boldsymbol{\mu}, B)$$

where the mean vector $\boldsymbol{\mu}$ and the covariance matrix B are assumed to be known (as usual from historical data, model forecasts, etc.). The observation now takes the form of a data vector, $\mathbf{z}$, of dimension p and has the conditional distribution/model:

$$\mathbf{z}|\mathbf{x} \sim N(H\mathbf{x}, R),$$

where the observation matrix $H \in \mathbb{R}^{p\times n}$ maps the process to the measurements and the error covariance matrix R is known. As before, we would like to calculate the posterior conditional distribution of $\mathbf{x}|\mathbf{z}$, given by

$$\pi(\mathbf{x}|\mathbf{z}) \propto \pi(\mathbf{z}|\mathbf{x})\pi_0(\mathbf{x}).$$

Just as with the scalar/univariate case, the product of two Gaussians is Gaussian, and the posterior law is the multidimensional analogue of (9.2) *and can be shown to take the form*

$$\mathbf{x}|\mathbf{z} \sim N(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}}, \Sigma_{\mathbf{x}|\mathbf{z}}) \tag{9.3}$$

where

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{z}} = (H^\top R^{-1} H + B^{-1})^{-1} \left((H^\top R^{-1}\mathbf{z} + B^{-1}\boldsymbol{\mu}\right)$$

and

$$\Sigma_{\mathbf{x}|\mathbf{z}} = (H^\top R^{-1} H + B^{-1})^{-1}.$$

As above, we will now rewrite the posterior mean and variance in a special form. The posterior conditional mean becomes

$$\mathrm{E}[\mathbf{x}|\mathbf{z}] = (H^\top R^{-1} H + B^{-1})^{-1} H^\top R^{-1}\mathbf{z} + (H^\top R^{-1} H + B^{-1})^{-1} B^{-1}\boldsymbol{\mu} = \boldsymbol{\mu} + K(\mathbf{z} - H\boldsymbol{\mu})$$

where K, the gain matrix, *is*

$$K = BH^\top (R + HBH^\top)^{-1}.$$

In the same manner, the posterior conditional covariance matrix can be reformulated as

$$\mathrm{Cov}(\mathbf{x}|\mathbf{z}) = (H^\top R^{-1} H + B^{-1})^{-1} = (I - KH)B$$

with the same gain matrix K as for the posterior mean. These last two equations are fundamental for a good understanding of data assimilation, since they clearly express the interplay between prior and data, and the effect that each has on the posterior.

9.2.2 Parameter estimation in a Bayesian framework

Regardless of the specific context (model calibration or inversion), all Bayesian inverse problems share the same ingredients: a computational forward model $\mathbf{f}$, a set of input parameters $\boldsymbol{\theta} \in \mathcal{P}$ that need to be inferred, and a set of experimental data $\mathbf{z}$. In a Bayesian context, parameters now become random variables $\boldsymbol{\Theta}$ with realizations $\boldsymbol{\theta} = \boldsymbol{\Theta}(\omega)$ and associated densities that incorporate known information or information obtained as measurements are acquired.

The forward model $\boldsymbol{\theta} \mapsto \mathbf{f}(\boldsymbol{\theta})$ is a mathematical representation of the system under consideration – this might involve the solution of a differential problem, and the evaluation of the output $\mathbf{f}(\boldsymbol{\theta})$ depending on it. All models are always simplifications of the real world. Thus, to connect model predictions $\tilde{\mathbf{z}} = \mathbf{f}(\boldsymbol{\theta})$ to the observations $\mathbf{z}$, a discrepancy term shall be introduced. We consider a well-established format, in which

$$\mathbf{z} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}, \tag{9.4}$$

where $\boldsymbol{\varepsilon}$ is the term that describes the discrepancy between an experimental observation $\mathbf{z}$ and the model prediction. Equivalently,

$$z_i = f_i(\boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\mathbf{z}, \boldsymbol{\varepsilon} \in \mathbb{R}^{n_z}$, $\boldsymbol{\theta} \in \mathbb{R}^p$ are random variables[4]. Here $\mathbf{z}$ is called the measurement (data are nothing but the actual realization of the measurement).

We will assume that discrepancies are modeled as additive and mutually independent from $\boldsymbol{\theta}$. This discrepancy term represents in practice the effects of measurement error (on y_i) and model inaccuracy. In the above equation, again for the sake of simplicity, this term is supposed to have a zero mean, but it could more generally include a model bias term. Hereon we will refer to this term as *noise*, but bare in mind that this is not only referred to a measurement noise – also model inaccuracy is included.

Moreover, let $\pi_{\text{noise}}(\boldsymbol{\varepsilon})$ denote the PDF of the noise $\boldsymbol{\varepsilon}$, usually encoding experimental errors. Before performing output measurements, all the information (e.g., structure or regularity) about the distribution of the input $\boldsymbol{\theta}$ are encapsulated in the *prior* PDF $\pi_{\text{prior}}(\boldsymbol{\theta})$, to be selected according to problem-specific considerations. The prior density incorporates any knowledge that we have about parameters prior to obtaining observations $\mathbf{z}$; this could come from previous similar experiments or analysis regarding similar models.

[4]More rigorously, we should write $Z_i = f_i(\boldsymbol{\Theta}) + \varepsilon_i$, $i = 1, \ldots, n$, where Z_i, ε_i and $\boldsymbol{\Theta}$ are random variables representing measurements, measurement errors, and parameters, respectively, and $f_i(\boldsymbol{\Theta})$ denotes the parameter-dependent model response. However, we do not adopt here the usual notation distinguishing between random variables and their realizations.

In the context of model inversion or calibration, the goal is to find the optimal values of the parameters $\boldsymbol{\theta}$ that allow one to fit model predictions to the observations. In this respect, epistemic uncertainty (lack of knowledge) on the input parameters is modeled by considering input parameters as a random vector $\boldsymbol{\theta} \sim \pi_{prior}(\boldsymbol{\theta})$, with given prior distribution.

The conditional probability $\pi(\mathbf{z} \mid \boldsymbol{\theta})$ of of $\mathbf{z}$ conditioned on $\boldsymbol{\Theta} = \boldsymbol{\theta}$ is the so-called (*conditional*) *likelihood function*, and expresses the likelihood of different measurement outcomes $\mathbf{z}$ given $\boldsymbol{\Theta} = \boldsymbol{\theta}$. The likelihood function $\pi(\mathbf{z} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{z})$ incorporates information provided by the samples and constitutes the mechanism through which data informs the posterior density. Thanks to the assumption of mutual independence of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$, $\mathbf{z} \mid \boldsymbol{\Theta} = \boldsymbol{\theta}$ is distributed like $\boldsymbol{\varepsilon}$, that is, the likelihood function is

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) = \pi_{\text{noise}}(\mathbf{z} - \mathbf{f}(\boldsymbol{\theta})).$$

this means that in the likelihood we reflect the model that we set for the measurements error

Given the observed data $\mathbf{z}$, in the Bayesian framework the inverse UQ problem of *parameter estimation* is to find the conditional PDF $\pi(\boldsymbol{\theta} \mid \mathbf{z})$ of $\boldsymbol{\theta}$. This latter is the *posterior* PDF of $\boldsymbol{\theta}$ given the data $\mathbf{z}$,

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}) = \frac{\pi_{\text{prior}}(\boldsymbol{\theta})\pi(\mathbf{z} \mid \boldsymbol{\theta})}{\pi(\mathbf{z})} = \frac{\pi_{\text{prior}}(\boldsymbol{\theta})\pi(\mathbf{z} \mid \boldsymbol{\theta})}{\int_{\mathcal{P}} \pi(\mathbf{z} \mid \boldsymbol{\theta})\pi_{\text{prior}}(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{9.5}$$

thanks to Bayes' theorem. The denominator $\pi(\mathbf{z})$ plays the role of a normalization constant, and often has little importance from a computational standpoint.

we can change $\pi(\vec{z} \mid \vec{\theta}) = \pi_{noise}(\vec{z} - \vec{f}(\vec{\theta}))$ (likelihood that is related with the distribution of the noise)

Remark 9.2.2. *The specification of the likelihood function depends on the assumptions made regarding the distribution of errors. For instance, by using the statistical model* (9.4) *with the assumption that errors are i.i.d. and* $\varepsilon_i \sim N(0, \sigma^2)$*, where* σ^2 *is fixed, then the likelihood function is*

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{z}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - f_i(\boldsymbol{\theta}))^2\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} SS_{\boldsymbol{\theta}}\right)$$

where

$$SS_{\boldsymbol{\theta}} = \sum_{i=1}^{n}(z_i - f_i(\boldsymbol{\theta}))^2.$$

Solving an inverse UQ problem in the case in which input parameters do not depend on time – alternatively, in the literature such a problem is referred to as the *stationary inverse problem* – thus consists in finding a prior PDF $\pi_{\text{prior}}(\boldsymbol{\theta})$, expressing the likelihood function $\pi(\mathbf{z} \mid \boldsymbol{\theta})$ using the interplay between the observation and the unknown, and finally developing suitable numerical techniques to explore the posterior PDF. Note that observations can be acquired at different locations in the physical domain, and/or be provided as a single quantity of interest measured at different time instants if the differential problem is time-dependent. Each of these tasks is a challenging problem from a computational standpoint.

In the case where the unknown is a random variable with few components, the posterior PDF can also be visualized in the form of a non-negative function of these variables. Most applications, however, yield larger-scale inverse UQ problems, and resulting PDFs in high-dimensional spaces, for which it is much more effective to evaluate suitable *point estimators*, such as the *maximum a posteriori* estimator

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} \pi(\boldsymbol{\theta} \mid \mathbf{z})$$

The prior predictive distribution is the distribution of $\vec{z}$ "averaged" over all possible values of $\vec{\theta}$:

$$\pi'_{pred}(\vec{z}) = \int_{(H)} p(\vec{z}, \vec{\theta})\, d\vec{\theta}$$
$$= \int_{(H)} \pi(\vec{z}|\vec{\theta})\, \pi_{prior}(\vec{\theta})\, d\vec{\theta}$$

this distribution is a prior and so it does not rely on any observation

in the same way we can define the posterior predictive distribution based on an observed $\vec{z}$:

$$\pi''_{pred}(z_{new}|\vec{z}) = \int_{(H)} p(z_{new}, \vec{\theta}|\vec{z})\, d\vec{\theta}$$
$$= \int_{(H)} \frac{p(z_{new}, \vec{\theta}, \vec{z})}{p(\vec{z})}\, d\vec{\theta}$$
$$= \int_{(H)} \frac{p(z_{new}|\vec{\theta}, \vec{z})\, p(\vec{\theta}, \vec{z})}{p(\vec{z})}\, d\vec{\theta}$$
$$= \int_{(H)} \frac{p(z_{new}|\vec{\theta}, \vec{z})\, p(\vec{\theta}|\vec{z})\, \cancel{p(\vec{z})}}{\cancel{p(\vec{z})}}\, d\vec{\theta}$$

$z_{new}|\vec{\theta} \perp\!\!\!\perp \vec{z}$

$$= \int_{(H)} \pi(z_{new}|\vec{\theta})\, \pi(\vec{\theta}|\vec{z})\, d\vec{\theta}$$

the posterior predictive distribution is constructed in the same way as the prior predictive distribution, however in the prior one we weight with $\pi_{prior}(\vec{\theta})$ while in the posterior one we weight with $\pi(\vec{\theta}|\vec{z})$ (which is the updated knowledge about $\vec{\theta}$)

or the *conditional mean*

$$\boldsymbol{\theta}_{\mathrm{CM}} = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}] = \int_{\mathbb{R}^p} \boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\theta.$$

Evaluating the former requires the solution of an optimization problem, using iterative, gradient-based methods; computing the latter involves a numerical quadrature problem in high-dimensional spaces. The evaluation of variability estimators such as the *conditional covariance*

$$\mathrm{Cov}\,(\boldsymbol{\theta} \mid \mathbf{z}) = \int_{\mathbb{R}^p} (\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{CM}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{CM}})^{\mathrm{T}}\, \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\theta \in \mathbb{R}^{p \times p},$$

or *confidence regions*, also provides further indicators for uncertainty quantification.

One may also be interested in posterior marginals of a specific parameter θ_i, that can be computed by integration over the other components (sometimes called nuisance parameters):

$$\pi(\theta_i \mid \mathbf{z}) = \int_{\mathcal{P}_{\sim i}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\boldsymbol{\theta}_{\sim i}.$$

- **Remark 9.2.3.** *To assess the predictive capabilities of a computational model, the Bayesian inference framework offers the possibility to compute predictive distributions.*

  The prior predictive distribution $\pi'_{pred}(\mathbf{z})$ is obtained by "averaging" the conditional distribution $\pi(\mathbf{z} \mid \boldsymbol{\theta})$ against the prior distribution $\pi_{prior}(\boldsymbol{\theta})$:

$$\pi'_{pred}(\mathbf{z}) = \int_{\mathcal{P}} \pi(\mathbf{z} \mid \boldsymbol{\theta}) \pi_{prior}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

  However, it is not necessary in practice to explicitly compute the above integral: samples from this distribution are obtained by sampling first $\boldsymbol{\theta}$ according to $\pi_{prior}(\boldsymbol{\theta})$, then sampling the distribution of $\mathbf{Z}|\mathbf{f}(\boldsymbol{\theta}), \boldsymbol{\varepsilon}$ – this corresponds to simply sampling a realization of the discrepancy term $\boldsymbol{\varepsilon}$ and adding it to $\mathbf{f}(\boldsymbol{\theta})$.

  The posterior predictive distribution $\pi''_{pred}(z_{new} \mid \mathbf{z})$ of z_{new} is obtained by "averaging" the conditional distribution $\pi(z_{new} \mid \boldsymbol{\theta})$ over the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$:

$$\pi''_{pred}(z_{new} \mid \mathbf{z}) = \int_{\mathcal{P}} \pi(z_{new} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\boldsymbol{\theta}$$

  A sample from the posterior predictive distribution is obtained by drawing $\boldsymbol{\theta}$ according to $\pi(\boldsymbol{\theta} \mid \mathbf{z})$, then evaluating $\mathbf{f}(\boldsymbol{\theta})$ and adding an independently sampled discrepancy term ε.

- **Remark 9.2.4.** *In practical inverse problems, the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{z})$ can also be an intermediate quantity that is further used for computing the conditional expectation of a certain quantity of interest (QoI) $\Psi : \mathcal{P} \to \mathbb{R}$. This can be anything from a simple analytical function to complex secondary models. This conditional expectation is simply the expectation of $\Psi(\boldsymbol{\theta})$ under the posterior distribution and is computed by the integral*

$$\mathbb{E}\left[\Psi(\boldsymbol{\theta}) \mid \mathbf{z}\right] = \int_{\mathcal{P}} \Psi(\boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{z}) d\boldsymbol{\theta}.$$

The simplest probabilistic model that can be used to describe experimental uncertainties is the Gaussian model, for which the noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma_\varepsilon)$ is normally distributed, with mean $\mathbf{0}$ and covariance matrix Σ_ε. Collecting all the data in the vector $\mathbf{z}$, the likelihood function is

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{z} - \mathbf{f}(\boldsymbol{\theta})\|^2_{\Sigma_\varepsilon^{-1}}\right).$$

If we can also assume a Gaussian model on the *prior* knowledge of the parameter distributions, $\pi_{\mathrm{prior}} \sim \mathcal{N}(\boldsymbol{\theta}_p, \Sigma_p)$, then the posterior PDF will be normally distributed as well,

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}) \propto \exp\left(-\frac{1}{2}\|\mathbf{z} - \mathbf{f}(\boldsymbol{\theta})\|^2_{\Sigma_\varepsilon^{-1}} - \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_p\|^2_{\Sigma_p^{-1}}\right). \tag{9.6}$$

In this case, the maximum *a posteriori* estimator is

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left(\frac{1}{2}\|\mathbf{z}-\mathbf{f}(\boldsymbol{\theta})\|^2_{\Sigma_\varepsilon^{-1}} + \frac{1}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_p\|^2_{\Sigma_p^{-1}}\right);$$

that is, it coincides with the estimator obtained by solving a constrained optimization problem – as it usually happens when dealing with *variational methods*; the second term in this case plays the role of a *regularization term*. If we assume instead that no information is available about the parameter distribution except that it resides in a subset $\mathcal{P}$ of the parameter space $\mathbb{R}^p$, then $\pi_{\text{prior}}(\boldsymbol{\theta}) \sim \mathcal{U}(\mathcal{P})$ is a uniform distribution over $\mathcal{P}$.

Remark 9.2.5. *Classical variational methods, based on PDE-constrained optimization, provide point estimates of the quantities of interest, given a set of observations; in the case of input quantities depending on time, sequential (or filtering) approaches can be used to reach the same goal. Compared with the variational methods, where classical regularization methods yield point estimates by curing the ill-posedness of the problem, statistical inversion aims at removing ill-posedness by recasting the inverse problem in a larger space of probability distributions [22]. This strategy also enables better characterization of the prior information contained in the regularization terms, in the form of a* prior *PDF of the unknown inputs.*

The formulation of the inverse problem in the Bayesian framework is provided by the result of (9.5). However, implementing a way to obtain the posterior distribution might become extremely challenging if the dimensionality p of $\boldsymbol{\theta}$ is large, as is often the case for cases in applied sciences and engineering. Besides few, selected cases in which an analytic integration is possible, we might rely on:

- classical Gaussian quadrature rules, if the number p of parameters is small, say $p = 1, \dots, 4$;
- sparse grid quadrature techniques for larger values of p;
- Monte Carlo techniques for high dimension, involving Markov Chain methods.

Instead of evaluating the posterior PDF at a single point, a Markov Chain Monte Carlo (MCMC) technique is a systematic way of generating a sample which can be used to *explore* the distribution, as well as to perform integration in order to compute the conditional mean or conditional covariance. The goal of a MCMC technique is to construct chains whose stationary distribution is the posterior density. In other words, the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$ plays the role of target probability distribution that we want to explore, and is obtained as a realization of a Markov chain. Before turning to this aspect, let us consider again the example related with the spring model.

Example 9.2.5. (A spring model, revisited). *We assume, in this context, displacement evaluations so that $z(t_i, \boldsymbol{\theta}) = u(t_i, \boldsymbol{\theta})$. We consider $K = 20.5$ to be known and treat $\boldsymbol{\theta} = C$ as the unknown parameter to be estimated. To construct synthetic data, we take $C_0 = 1.5$ and construct iid errors $\varepsilon_i \sim N(0, \sigma_0^2)$ where $\sigma_0 = 0.1$. For this error distribution, the likelihood is given by*

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{z}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - f_i(\boldsymbol{\theta}))^2\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} SS_{\boldsymbol{\theta}}\right).$$

We employ a non-informative prior $\pi_{\text{prior}}(\boldsymbol{\theta}) = \chi_{[0,\infty)}(\boldsymbol{\theta})$ to enforce C to be nonnegative. The posterior density can thus be expressed as

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}) = \frac{\exp^{-SS_{\boldsymbol{\theta}}/2\sigma_0^2}}{\int_0^\infty \exp^{-SS_\xi/2\sigma_0^2} d\xi} = \frac{1}{\int_0^\infty \exp^{-(SS_\xi - SS_{\boldsymbol{\theta}})/2\sigma_0^2} d\xi}.$$

The integral can be approximated, e.g., through the midpoint rule, obtaining

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}) \approx \frac{1}{\sum_{k=1}^{K} \exp^{-(SS_{\xi_k} - SS_{\boldsymbol{\theta}})w_k / 2\sigma_0^2}}$$

where ξ_k and w_k are the quadrature nodes and weights, respectively.

After generating a set of synthetic data z_i, $i = 1, \ldots, 501$, plotted in the figure below along with the model response $f_i(\boldsymbol{\theta}_0)$, we determine the posterior density and in particular the MAP estimate $\theta_{MAP} = 0.1489$ that coincides with the ML estimate since we have employed a non-informative prior. For the assumed error distribution, it also coincides with the LS estimate. However, the posterior density has the same shape of the sampling distribution

$$\hat{\boldsymbol{\theta}}_{OLS} = \hat{C}_{OLS} \sim N(C_0, \sigma_0^2 [\mathbf{J}(C_0)^\top \mathbf{J}(C_0)]^{-1})$$

but the sampling distribution is centered at C_0.

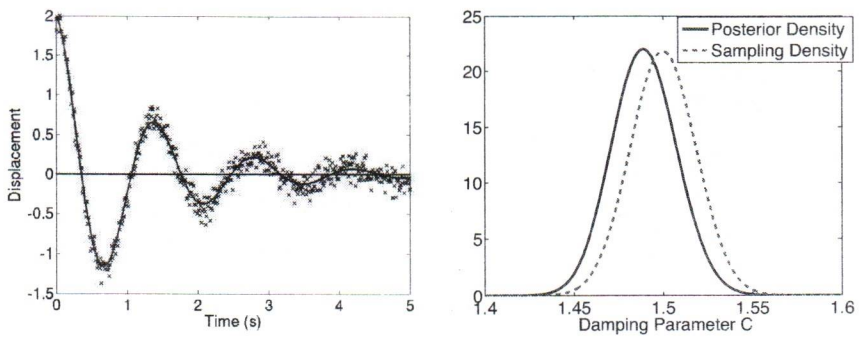


Figure 9.7: Left: synthetic data z_i and model response $f_i(\theta_0)$. Right: posterior density and sampling distribution. Taken from [45].

9.2.3 More on the case of inverse UQ for PDE systems*

Bayesian inverse problems provide a suitable framework for the estimation of parameters which are not function of time. PDE models and outputs can be given by either a set of measurements, acquired at different locations in the spatial domain (in the case of a stationary PDE model), or at different times, or both (in the case of a time-dependent PDE model). In the case of time-dependent model inputs, provided measurements are acquired over time, parameter estimation is performed by relying on filtering techniques and data assimilation methods, still rooted in the Bayesian framework.

Remark 9.2.6. *Very often, inverse UQ is then performed by considering* synthetic *measurements given by an output QoI (depending on the solution of the state system), obtained for a selected value of the unknown model inputs, to which an additive noise is summed.*

Let us assume that, given a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, the state $\mathbf{u}(\boldsymbol{\theta}) \in \mathbb{R}^N$ of a physical system can be obtained as a solution of a set of discretized (non)linear equations arising from the discretization of a stationary, (non)linear PDE, under the form

$$\mathbf{R}(\mathbf{u}; \boldsymbol{\theta}) := \mathbf{R}_0 + A\mathbf{u} + \mathbf{s}(\mathbf{u}; \boldsymbol{\theta}) = \mathbf{0},$$

where $\mathbf{R}_0$ is a source term independent of the state, $A\mathbf{u}$ is the linear part of the state equation assumed independent of the parameters, and $\mathbf{s}(\mathbf{u}; \boldsymbol{\theta})$ the nonlinear part of the state equations

including all the dependence on the parameters. For the sake of simplicity we restrict ourselves to the case of stationary PDEs, the extension to the time-dependent case being also possible. Observations $\mathbf{z}$ of the state $\mathbf{u}$ can be made through the observation equation

$$\mathbf{z} = C(\mathbf{u}(\boldsymbol{\theta}))$$

for a given map $C : \mathbb{R}^N \to \mathbb{R}^{n_z}$. When dealing with parameter estimation for PDEs, given an observed output $\mathbf{z}_{obs} \in \mathbb{R}^{n_z}$, hampered by noise, $\mathbf{z}_{obs} = \mathbf{z} + \boldsymbol{\varepsilon}$, we aim at finding the parameter $\hat{\boldsymbol{\theta}}$ that satisfies $\mathbf{z}_{obs} = C(\mathbf{u}(\hat{\boldsymbol{\theta}}))$. This is an ill-posed problem, since it might happen that:

1. there exists no solution because $\mathbf{z}_{obs} \notin \mathrm{Im}(\mathbf{C} \circ \mathbf{R})$;
2. the solution is not unique because $\mathbf{C} \circ R : \mathbb{R}^p \to \mathbb{R}^{n_z}$ is not injective;
3. The operator $\Phi : \mathrm{Im}(\mathbf{C} \circ \mathbf{R}) \to \mathbb{R}^p$ given by

$$\Phi(\mathbf{z}) := \underset{\boldsymbol{\theta} \in \mathcal{P}}{\operatorname{argmin}} \{\|\mathbf{z} - \mathbf{C}(\mathbf{R}(\mathbf{u}; \boldsymbol{\theta}))\|_2\} \quad \forall \mathbf{z} \in \mathrm{Im}(\mathbf{C} \circ \mathbf{R})$$

   is not continuous with respect to the observations $\mathbf{z}$.

Typical inverse problems are (severely) ill-posed in the sense of the latter condition. A first attempt to rely on a deterministic approach for the solution of the inverse problem: given an observation $\mathbf{z}_{obs}$, solve the least-squares problem

$$\min_{\boldsymbol{\theta} \in \mathcal{P}} \quad \frac{1}{2}\|\mathbf{z}_{obs} - \mathbf{z}\|_2^2$$

subject to state and observation equations

$$\begin{cases} \mathbf{R}(\mathbf{u}; \boldsymbol{\theta}) = \mathbf{0} \\ \qquad \mathbf{z} = C\mathbf{u}; \end{cases}$$

however, this problem still suffers from nonuniqueness and ill-posedness. By introducing the so-called *Tykhonov regularization*, we rather aim at solving the following minimization problem:

$$\min_{\boldsymbol{\theta} \in \mathcal{P}} \quad \frac{1}{2}\|\mathbf{z}_{obs} - \mathbf{z}\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{prior}}\|_R^2$$

subject to the above constraints, where $\boldsymbol{\theta}_{\text{prior}}$ is some prior guess for the parameters and $\alpha > 0$ is a regularization parameter (possibly very small). Here $\|\cdot\|_R$ denotes the norm $\sqrt{(R\cdot, \cdot)}$ for some symmetric and positive definite matrix R. This is a *regularized least-squares* approach. Solution $\hat{\boldsymbol{\theta}}$ is now unique, however for sufficiently large α, and usually strongly depends on the choice of α and $\boldsymbol{\theta}_{\text{prior}}$. Incorrect choices of μ_{prior} and $\|\cdot\|_R$ can introduce bias into the estimate.

For these reasons, and aiming at obtaining a PDF for the model input, rather than a point estimate, we cast the problem in the Bayesian framework, and derive a conditional distribution $\pi_{\boldsymbol{\theta}|\mathbf{z}_{obs}}(\boldsymbol{\theta} \mid \mathbf{z}_{obs})$. Let us consider the following assumptions:

1. the state $\mathbf{u}$ depends linearly on the parameters $\boldsymbol{\theta}$, i.e. $R(\mathbf{u}; \boldsymbol{\theta}) = A\mathbf{u} + S\boldsymbol{\theta} = \mathbf{0}$; hence, the forward model is given by

$$\mathbf{z} = -CA^{-1}S\boldsymbol{\theta};$$

2. parameters follow a Gaussian distribution, i.e. $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_p, \Sigma_0)$, so that

$$\pi_{\text{prior}}(\boldsymbol{\theta}) \propto \exp\left(-(\boldsymbol{\theta} - \boldsymbol{\theta}_p)^T \Sigma_0^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_p)\right);$$

3. observation noise is i.i.d Gaussian with zero mean, i.e. $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, so that

$$\pi_{\text{noise}}(\varepsilon) \propto \exp\left(-\frac{1}{\sigma^2}\|\varepsilon\|_2^2\right),$$

Then, the conditional distribution $\pi(\boldsymbol{\theta} \mid \mathbf{z}_{obs})$ becomes

$$\pi(\mathbf{z}_{obs} \mid \mu) = \pi_{\text{noise}}(\mathbf{z}_{obs} + CA^{-1}S\boldsymbol{\theta}) \propto \exp\left(-\|\mathbf{z}_{obs} + CA^{-1}S\mu\|_2^2\right),$$

and applying the Bayes' theorem we find

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}_{obs}) \propto \exp(-\underbrace{\|\mathbf{z}_{obs} + CA^{-1}S\boldsymbol{\theta}\|_2^2}_{\text{"least-squares fit"}} - \underbrace{\sigma^2(\boldsymbol{\theta} - \boldsymbol{\theta}_p)^T\Sigma_p^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_p)}_{\text{"regularization"}}).$$

Note that if we choose $\boldsymbol{\theta}_{\text{prior}} = \boldsymbol{\theta}_p$, $\alpha = \sigma^{-2}$, and $\|\mu\|_R^2 = \boldsymbol{\theta}^\top\Sigma_p^{-1}\boldsymbol{\theta}$, the solution of the deterministic inverse problem coincides with the MAP estimate

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} := \underset{\boldsymbol{\theta}\in\mathcal{P}}{\text{argmax}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z}_{obs}).$$

Remark 9.2.7. *Recall that $\pi_{prior}(\boldsymbol{\theta})$ expresses our beliefs about the variability of $\boldsymbol{\theta}$ before doing any measurements. Indeed, prior with large variance (no information) will smear the posterior distribution; prior with small variance but incorrect form (wrong information) introduces bias.*

The main challenges in applying a Bayesian approach to inverse UQ for PDEs are:

1. the construction of realistic priors $\pi_{\text{prior}}(\boldsymbol{\theta})$, which is a highly problem-specific task;
2. the fact that the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{z})$ is rarely available in analytical form if state equations are nonlinear or if prior $\pi_{\text{prior}}(\boldsymbol{\theta})$ is not of special form;
3. the need to sample from high-dimensional distributions, such as $\pi(\boldsymbol{\theta} \mid \mathbf{z})$;
4. the need to rely, sometimes, on the entire PDF of the input parameters,

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\mathbb{R}^{n_z}} \pi(\boldsymbol{\theta} \mid \mathbf{z})\,\pi(\mathbf{z})\,d\mathbf{z}$$

so that marginalizing the output requires an additional n_z-dimensional integral.

In principle, the posterior distribution gives the solution to the parameter estimation problem in a fully probabilistic sense. We can find the peak of the probability density, and determine, for instance, the 95% credibility regions for the parameters. However, working with the posterior density directly is challenging, since we need to compute the normalization constant in the Bayes formula. In most cases this cannot be computed analytically, and classical numerical integration methods (sparse grid, Monte Carlo, ...) also become infeasible, if the number of parameters is larger than a few. The difficulty of this approach is however exacerbated by the fact that the support of the density is often part of the information we are seeking. The same issue arises whenever interested to compute statistics (posterior mean, covariance, ...) of the posterior distribution.

9.3 A non rigorous introduction to MCMC techniques

With the so called Markov chain Monte Carlo (MCMC) methods, statistical inference for the model parameters can be done without explicitly computing this difficult integral. MCMC methods aim at generating a sequence of random samples $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N, \ldots)$ whose distribution asymptotically approaches the posterior distribution as N increases. That is, the posterior density is not used directly, but samples from the posterior distribution are produced instead[5].

[5]Rather than using quadrature or MC algorithms to specify parameter values at which we evaluate the density, we can use attributes of the density to specify parameter values that are more suitable to explore the geometry of the distribution.

The samples are generated so that each new point $\boldsymbol{\theta}_{k+1}$ only depends on the previous point $\boldsymbol{\theta}_k$, and the samples therefore form a Markov Chain. Markov Chain theory can be used to show that the distribution of the resulting samples approaches the correct target (posterior) – more rigorously, the stationary distribution of the constructed *Markov chain* is the posterior density. By evaluating realizations of the chain, we thus sample the posterior and hence obtain a density for the parameter values based on observed measurements. This is the idea of the MCMC techniques, that are first introduced in a (hopefully) intuitive way below, then further analyzed later on.

9.3.1 Metropolis algorithm

One of the most widely used MCMC algorithms is the *random walk Metropolis* algorithm introduced already in 1950s in statistical physics literature. The Metropolis algorithm is very simple: it works by generating candidate parameter values from a proposal distribution and then either accepting or rejecting the proposed value according to a simple rule. The idea behind this method is not so different from the one of Acceptance-Rejection sampling.

Metropolis algorithm (to sample from the *target* distribution $\pi(\boldsymbol{\theta})$)

1. choose a starting point $\boldsymbol{\theta}_1$, set $k = 1$;
2. choose a new candidate $\boldsymbol{\theta}^*$ from a suitable *proposal distribution* $q(\cdot \mid \boldsymbol{\theta}_k)$, that may depend on the previous point of the chain;
3. accept the candidate with probability

$$\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_k)}\right).$$

and set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}^*$, $k \to k+1$. If rejected go back to step 2.

Note that the accept-reject step can be implemented by generating a uniform random number $u \sim \mathcal{U}(0,1)$ and accepting if

$$u \leq \pi(\boldsymbol{\theta}^*)/\pi(\boldsymbol{\theta}_k).$$

Note also that in the Metropolis algorithm we only need to compute ratios of target (in our case, posterior) densities, so that the normalization constant (the nasty integral appearing at the denominator in the expression of the posterior distribution) cancels out. This is what makes MCMC computationally feasible in multidimensional parameter estimation problems.

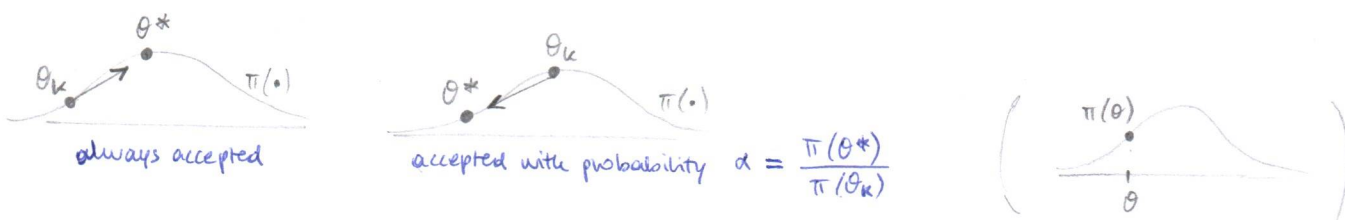
The Metropolis algorithm assumes a symmetric proposal distribution q, that is, the probability density of moving from the current point to the proposed point is the same as moving backwards from the proposed point to the current point:

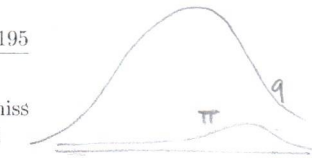
$$q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_n) = q(\boldsymbol{\theta}_n \mid \boldsymbol{\theta}^*).$$

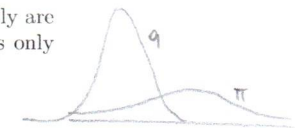
A simple extension to non-symmetric proposal distributions is given by the Metropolis-Hastings algorithm, that will be addressed later on.

Remark 9.3.1. *In the Metropolis algorithm the candidate points that give a better posterior density value than the previous point (points where $\pi(\boldsymbol{\theta}^*) > \pi(\boldsymbol{\theta}_n)$), or moves 'upward' in the posterior density function, are always accepted. However, moves 'downward' may also be accepted, with probability given by the ratio of the posterior density values at previous and proposed points.*

The problem remaining in the implementation of the Metropolis algorithm is defining the proposal distribution q. The proposal should be chosen so that it is easy to sample from and as 'close' to the underlying target distribution (posterior) as possible. An unsuitable proposal can lead to inefficient implementation (see Figure 9.8):







- *wide proposal*: if the variance of the proposal is too *large*, the new candidates mostly miss the essential support π, they are chosen at points where $\pi \approx 0$ and only rarely accepted;
- *narrow proposal*: if the variance of the proposal is too *small*, the new candidates mostly are accepted, but from a small neighborhood of the previous point. So the chain moves only slowly, and may, in finite number of steps, not cover the target π.

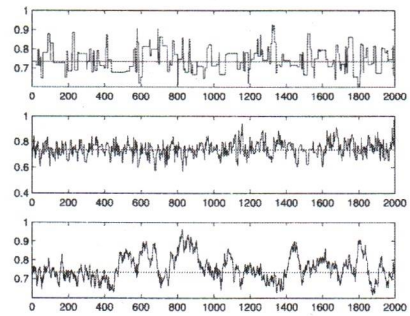


Figure 9.8: Examples of MCMC chains for one parameter. The upper picture tells that the proposal is too wide - the chain stays still for long periods. The lowest picture presents narrow proposal - the sampler explores the distribution slowly. The chain in the middle shows a good 'mixing'.

When dealing (as it is often the case for our purposes) with multidimensional continuous parameter spaces, a multivariate Gaussian distribution is a common choice for a proposal distribution. In the commonly used random walk Metropolis algorithm, the current point in the chain is taken as the center point of the Gaussian proposal. The problem then is to find a suitable covariance matrix so that the size and the shape (orientation) of the proposal matches as well as possible with the target density (posterior) to produce efficient sampling. The covariance matrix selection issue is slightly postponed.

The form of the posterior density depends on the case. Our most typical application is MCMC for standard nonlinear parameter estimation. Typically, the prior information we have are some bound constraints for the parameters, and within the bounds we use a uniform, *flat* prior,

$$\pi_{\text{prior}}(\boldsymbol{\theta}) \propto 1.$$

Assuming in addition independent measurement error with a known constant variance σ^2, the posterior density can be written as

$$\pi(\boldsymbol{\theta} \mid \mathbf{z}) \propto \pi(\mathbf{z} \mid \boldsymbol{\theta})\pi_{\text{prior}}(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\sigma^2}SS(\boldsymbol{\theta})\right)$$

where

$$SS(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - f_i(\boldsymbol{\theta}))^2$$

is the sum of squares function that we minimize when the LS estimate is computed.

Using this notation, the acceptance ratio in the Metropolis algorithm reduces to

$$\alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_n)}\right) = \min\left(1, \exp\left(-\frac{1}{2}(SS(\boldsymbol{\theta}^*) - SS(\boldsymbol{\theta}_n))\right)\right).$$

Using these assumptions and this notation, the Metropolis algorithm with a multivariate Gaussian proposal distribution, with covariance matrix $\mathbf{C}$ and initial point $\boldsymbol{\theta}_{old} = \boldsymbol{\theta}_0$, becomes:

Random walk Metropolis Algorithm, case of uniform prior distribution

1. Generate a candidate value $\boldsymbol{\theta}_{new} \sim N(\boldsymbol{\theta}_{old}, \mathbf{C})$ and compute $SS(\boldsymbol{\theta}_{new})$.
2. Accept the candidate if $u < \exp\left(-\frac{1}{2\sigma^2}(SS(\boldsymbol{\theta}_{new}) - SS(\boldsymbol{\theta}_{old}))\right)$ where $u \sim \mathcal{U}(0,1)$.
   - If accepted, add $\boldsymbol{\theta}_{new}$ to the chain, set $\boldsymbol{\theta}_{old} := \boldsymbol{\theta}_{new}$ and $SS(\boldsymbol{\theta}_{old}) := SS(\boldsymbol{\theta}_{new})$.
   - If rejected, repeat $\boldsymbol{\theta}_{old}$ in the chain.
3. Go to step 1 until a desired chain length is achieved.

Hands-On 9 focuses on the implementation of the Random walk Metropolis Algorithm.

Remark 9.3.2. *Note that although we assume here a* flat *prior, it is easy to add possible prior information about the parameters. Implementing simple bound constraints is easy: if the proposed parameter is out of bounds, it is simply rejected.*

Remark 9.3.3. *A possible initialization of the chain in the MCMC algorithm can be obtained by computing the OLS estimate by minimizing the sum of squares.*

9.3.2 Selecting the Proposal distribution

Selecting the proposal distribution is one of the main factors that affects the performance of an MCMC algorithm. In our setting so far, selecting the proposal means specifying the covariance matrix $\mathbf{C}$ of the multivariate Gaussian proposal distribution. A good starting point for selecting $\mathbf{C}$ is to use the approximation of the covariance matrix obtained via linearization of the model, which was developed in Section 9.1.4. That is, we perform a Gaussian approximation of the posterior distribution at the LS estimate and choose the proposal covariance matrix $\mathbf{C} = \sigma^2(\mathbf{J}(\boldsymbol{\theta}_{LS})^\top \mathbf{J}(\boldsymbol{\theta}_{LS}))^{-1}$. This proposal can better match with the orientation of the target distribution. In addition to orientation, scale of the proposal distribution is important; it has been found that, for Gaussian targets, an efficient scaling is $s_p = 2.4^2/p$, where p is the dimension of the problem (number of parameters). This result can be used as a rule of thumb also for non-Gaussian targets. That is, utilizing the Jacobian-based covariance matrix, we can use proposal covariance matrix

$$\mathbf{C} = s_p \sigma^2 (\mathbf{J}(\boldsymbol{\theta}_{LS})^\top \mathbf{J}(\boldsymbol{\theta}_{LS}))^{-1}. \tag{9.7}$$

We'll see that keeping $C = s_p \sigma^2 (\ldots)^{-1}$ fixed during the algorithm is not a good idea. An adaptive version of Metropolis is the one in which C is updated during the generation of the chain.

9.3.3 Adaptive MCMC

The bottleneck in MCMC computations is usually selecting a proposal distribution that matches well with the target distribution. The proposal covariance matrix using the linearization of the model discussed in the previous section is a good starting point, but does not always lead to efficient sampling. For instance, some of the parameters might be badly identified by the available data, which can result in a nearly singular Jacobian matrix and inefficient sampling. The purpose of adaptive MCMC methods is to tune the proposal 'on the run' as the sampling proceeds, using the information of the previously sampled points.

A simple way to implement adaptive MCMC is to simply compute the empirical covariance matrix of the points sampled so far, and use that as a proposal covariance matrix. Note that now the sampled points depend on the earlier history of the chain, not just the previous point, and the chain is therefore no longer Markovian. However, if the adaptation is based on an increasing part of the history, so that the number of previous points that is used to compute the empirical covariance matrix increases constantly as the sampling proceeds, it can be shown that the algorithm gives correct (ergodic) results.

Adaptive Metropolis

In the Adaptive Metropolis (AM) algorithm, the proposal is taken to be Gaussian, centered at the current point, and the proposal covariance matrix is taken to be the empirical covariance matrix computed from the history. More precisely, if we have sampled points $(\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{n-1})$, we propose the next candidate using the covariance $\mathbf{C}_n = s_p(Cov(\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{n-1}) + \varepsilon \mathbf{I})$, where $\varepsilon > 0$ is a regularization parameter that ensures that the proposal covariance matrix stays positive definite (in practice, it can often be chosen to be very small or even set to zero).

In order to start the adaptation procedure, an arbitrary strictly positive definite initial covariance $\mathbf{C}_0$ is chosen, according to a priori knowledge (which may be quite poor). A time index $n_0 > 0$ defines the length of the initial non-adaptation period, which is often called the *burn-in* period in the literature, after which we use the empirical covariance matrix as the proposal:

$$\mathbf{C}_n = \begin{cases} \mathbf{C}_0 & n \leq n_0 \\ s_p(Cov(\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{n-1}) + \varepsilon \mathbf{I}) = & n > n_0 \end{cases}$$

where we can use the scaled Jacobian-based covariance matrix (9.7) as the initial proposal – or even a simpler (e.g., diagonal) $\mathbf{C}_0$ that is small enough so that the sampler gets moving, and let the adaptation tune the proposal. The empirical covariance matrix does not have to be recomputed every time, since it is possible to express

$$Cov(\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_k) = \frac{1}{k}\left(\sum_{i=0}^{k} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top - (k+1)\bar{\boldsymbol{\theta}}_k \bar{\boldsymbol{\theta}}_k^\top\right)$$

where

$$\bar{\boldsymbol{\theta}}_k = \frac{1}{k+1}\sum_{i=0}^{k} \boldsymbol{\theta}_i$$

is the sample mean. Hence,

$$\mathbf{C}_{n+1} = \frac{n-1}{n}\mathbf{C}_n + \frac{s_p}{n}\left(n\bar{\boldsymbol{\theta}}_{n-1}\bar{\boldsymbol{\theta}}_{n-1}^\top - (n+1)\bar{\boldsymbol{\theta}}_n \bar{\boldsymbol{\theta}}_n^\top + \boldsymbol{\theta}_n \boldsymbol{\theta}_n^\top + \varepsilon \mathbf{I}\right)$$

which permits the calculation of the covariance update with little computational cost. Note that the effect of adaptation goes down as $1/n$; this is often called diminishing adaptation: in the long run, AM goes back to usual non-adaptive sampling, since new sampled points affect the proposal less and less as the sampling proceeds. This form of adaptation can proved to be ergodic.

- **Remark 9.3.4.** *An even more efficient algorithm is the so-called Delayed Rejection Adaptive Metropolis (DRAM) algorithm. Upon rejection, instead of repeating the previous value in the chain, $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$, a second stage move $\tilde{\boldsymbol{\theta}}_k^{(2)}$ is proposed from a (possibly) different proposal distribution q_2. The second stage proposal is allowed to depend not only on the current position of the chain but also on what we have just proposed and rejected: $q_2(\cdot \mid \boldsymbol{\theta}_k, \tilde{\boldsymbol{\theta}}_k)$. An ergodic chain is then created, if the second stage proposal is accepted with suitably modified acceptance probability. The process of delaying rejection can be iterated to try sampling from further proposals. In practice, often only a 2-stage version is used, where the second stage proposal is a downscaled version of the first stage proposal. That is, upon rejection, we try a new candidate value closer to the current point. The delayed rejection method can be combined with the adapting proposal covariance matrix, yielding the so-called Delayed Rejection Adaptive Metropolis (DRAM).*

9.3.4 The case of a general prior distribution

Let us now start to highlight the role of the prior and the likelihood function in the Metropolis algorithm, thus looking at the target distribution $\pi(\boldsymbol{\theta})$ as the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$.

In the case of a general prior distribution $\pi_{prior}(\boldsymbol{\theta})$, the Metropolis algorithm with a multivariate Gaussian proposal distribution, with covariance matrix $\mathbf{C}$ and initial point $\boldsymbol{\theta}_0$, becomes:

Random walk Metropolis Algorithm, case of a generic prior distribution

1. Initialization: choose an initial parameter value $\boldsymbol{\theta}_0$ such that $\pi(\boldsymbol{\theta}_0 \mid \mathbf{z}) > 0$.
2. For $k = 1, \ldots, M$
   (a) Sample $z \sim N(0,1)^p$ and construct the candidate $\boldsymbol{\theta}^* = \boldsymbol{\theta}_{k-1} + \mathbf{R}z$, where $\mathbf{R}$ is the Cholesky factor of $\mathbf{C}$
   (b) Compute the ratio

$$r(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_{k-1}) = \frac{\pi(\boldsymbol{\theta}^* \mid \mathbf{z})}{\pi(\boldsymbol{\theta}_{k-1} \mid \mathbf{z})} = \frac{\pi(\mathbf{z} \mid \boldsymbol{\theta}^*)\pi_{prior}(\boldsymbol{\theta}^*)}{\pi(\mathbf{z} \mid \boldsymbol{\theta}_{k-1})\pi_{prior}(\boldsymbol{\theta}_{k-1})}$$

   (c) Set

$$\boldsymbol{\theta}_k = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \alpha = \min(1, r) \\ \boldsymbol{\theta}_{k-1} & \text{otherwise.} \end{cases}$$

(Recall) sampling from a multivariate gaussian can be done by sampling from a standard multivariate gaussian and then scaling by the cholesky factor of the covariance matrix and summing the shift

In the case of a uniform prior and normally distributed errors,

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} SS(\boldsymbol{\theta})\right)$$

so that we find, once again, the relationship

$$r(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_{k-1}) = \exp\left(-\frac{1}{2}(SS(\boldsymbol{\theta}^*) - SS(\boldsymbol{\theta}_{k-1}))\right).$$

The effect of a proposal distribution that is too *narrow* or *wide* is shown, regarding the example of the spring model, in Figure 9.9.

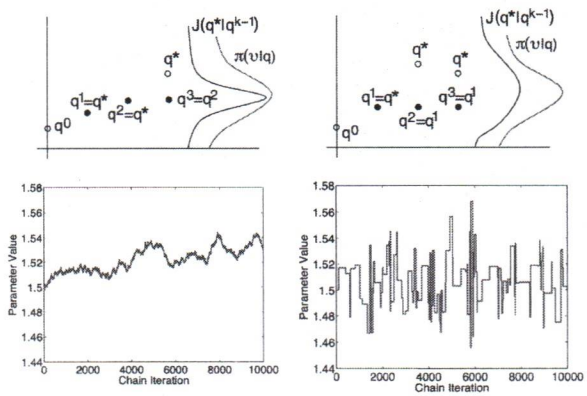


Figure 9.9: Top: generation of candidates $\boldsymbol{\theta}^*$ (here denoted as q^*) based on (left) narrow and (right) wide proposal distributions $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_{k-1}$. Bottom: resulting chains. Taken from [45].

9.3.5 Metropolis-Hastings Algorithm and Gibbs sampling

In the Metropolis algorithm presented so far, the proposal distribution was assumed to be symmetric, that is, for two parameter values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ we have

$$q(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1) = q(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2),$$

where $q(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$ denotes the density for proposing a move from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$. However, the algorithm can be extended for non-symmetric proposal distributions: the algorithm is otherwise the same as the Metropolis algorithm, but the acceptance probability is slightly modified to account for the non-symmetry.

In the *Metropolis-Hastings (MH) algorithm*, the probability of accepting the move from $\boldsymbol{\theta}_k$ to $\boldsymbol{\theta}^*$ is given by

$$\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_k)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_k)}\right).$$

Comparing to the acceptance probability of the Metropolis algorithm, one can see that the only difference is the inclusion of the ratio of the proposal densities, $q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}^*)/q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_k)$.

In the Metropolis algorithm presented above, candidate values for all parameters are proposed at the same time. However, sometimes, especially in high-dimensional problems, it may be difficult to find a good multivariate proposal distribution for all parameters simultaneously. The idea in *Gibbs sampling* is to reduce the sampling to one dimensional distributions: each parameter is sampled in turn, while the other parameters are kept fixed. In more detail, the parameter vector $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_p)$ is updated in sweeps, by updating one coordinate at a time. This may be done if the one-dimensional conditional distributions $\pi(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{\sim i})$ are known. In many cases these reduce to simple known densities which are easy to sample from. Often, however – in non- linear problems – the conditional distributions are not known, and they must be approximated by computing 'sufficiently' many values in the 1D directions. In particular, there is no accept-reject procedure here and the point taken from the 1D distribution is always accepted.

- **Remark 9.3.5.** *Instead of sampling directly from the one-dimensional conditional distributions, as in Gibbs sampling, one can perform component-wise Metropolis sampling. The proposal distribution of each component is, for instance, a normal distribution with the present point as the center point and with a given variance, separate for each coordinate. The coordinates are updated in loops, similarly as in Gibbs sampling.*

9.3.6 Sample-based error variance and MCMC estimation of σ^2

In our typical applications, the likelihood distribution reads as

$$\pi(\mathbf{z} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} SS(\boldsymbol{\theta})\right) \tag{9.8}$$

where $SS(\boldsymbol{\theta})$ is the sum of squares function. In practice, we need to specify the value for the measurement error variance σ^2. So far, we have given a fixed value for σ^2, estimated from the residuals of the model fit or from repeated measurements. However, instead of fixing σ^2 to a specific point estimate, we can also regard σ^2 as a random variable and treat it in a Bayesian way by sampling it along with the model parameters in the MCMC algorithm.

We often have a rather good idea about the level of the measurement error, and we therefore would like to specify a prior distribution for it. A computationally convenient choice for the prior is obtained using the conjugacy property[6] Looking at the Gaussian likelihood (9.8), and considering it as a function of σ^2 (with fixed $\boldsymbol{\theta}$), we see that the inverse variance $1/\sigma^2$ has a Gamma type distribution[7]. The conjugate prior for the Gamma distribution is also a Gamma distribution.

[6] If we set the prior so that the posterior is of the same form as the prior, the prior is called a conjugate prior.

[7] For $X \sim Gamma(\alpha, \beta)$, the density is

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0.$$

That is, if we specify a Gamma prior for $1/\sigma^2$, the conditional posterior $\pi(\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\theta})$ will also be Gamma distributed. More specifically, the prior for σ^{-2} can be defined as

$$\sigma^{-2} \sim \Gamma\left(\frac{n_0}{2}, \frac{n_0}{2} S_0^2\right).$$

That is, we define the prior for the measurement error variance with two parameters, n_0 and S_0 (hyper-parameters). This parameterization is chosen because the prior parameters are easy to interpret: S_0^2 gives the mean value for σ^2 and n_0 defines how accurate we think the value S_0^2 is. The higher n_0, the more peaked the prior distribution is around S_0^2, and the more informative the prior is (see Figure 9.10).

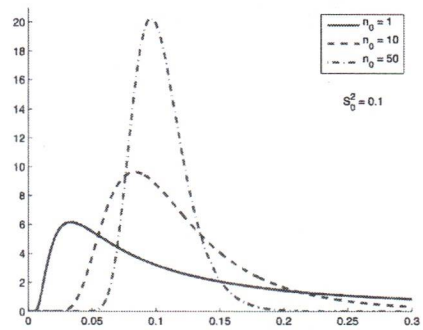


Figure 9.10: Prior densities for error variance with $S_0^2 = 0.1$ and different values for n_0.

With the conjugate prior, we can derive a Gamma posterior for the conditional posterior for the error variance, $\pi(\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\theta})$. Without going into the details, the conditional posterior for the posterior of σ^{-2} can be written as

$$\pi(\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\theta}) \sim \Gamma\left(\frac{n_0 + n}{2}, \frac{n_0 S_0^2 + SS(\boldsymbol{\theta})}{2}\right).$$

Now we have an analytical expression for the conditional distribution of σ^{-2}, and we can build a Gibbs sampler that first samples $\boldsymbol{\theta}$ as usual and then samples a new value for σ^2 from the above density by iterating the following steps:

1. sample a new $\boldsymbol{\theta}$ value from $\pi(\boldsymbol{\theta} \mid \sigma^2, \mathbf{z})$;
2. sample a new σ^2 value from $\pi(\sigma^{-2} \mid \mathbf{z}, \boldsymbol{\theta})$.

Hands-On 9 focuses on the application of MCMC algorithms implemented in the `mcmcstat` `Matlab` package.

9.3.7 A further example of application: a simply supported beam

To show a further example of application of the Bayesian framework, let us consider a case arising from structural mechanics. This example is part of the `UQLab` library available for `Matlab`. In `UQLab`, currently, four MCMC samplers are implemented: Metropolis Hastings (MH), Adaptive-Metropolis (AM), Hamiltonian Monte Carlo (HMC) and affine invariant ensemble sampler (AIES).

Then, $Y = X^{-1}$ has an inverse Gamma distribution whose density is

$$f_Y(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/y}, \quad y > 0.$$

In this example, the measured deflections of a simply supported beam under a uniform load p are used to calibrate the Young's Modulus E of the beam material. Basic uncertainty on the applied p is assumed. The beam has a known rectangular cross-section of width b and height h and a known span of length L (see Figure 9.11). In the experiments, the beam is subject to a constant distributed load p and the mid-span deflection V_{mid} is measured. A set of $N = 5$ independent experiments are carried out with this beam, with the goal of inferring the Young's modulus E.

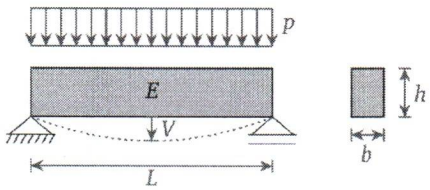


Figure 9.11: Beam deflection: configuration

The measured beam deflections $z_1, \ldots, z_N$ are reported in Table 9.1. Due to measurement error, the measured deflections vary across experiments.

Experiment	1	2	3	4	5
V_{mid} (mm)	12.84	13.12	12.13	12.19	12.67

Table 9.1: Measurements of the beam deflection

The forward model computes the mid-span deflection V_{mid}, whose analytical expression, according to the standard beam theory, is

$$V = \frac{5pL^4}{32Ebh^3};$$

this simple equation serves as the forward model and relates the unknown Young's modulus E to the measurable mid-span deflection. Note that, for more general constitutive laws or structural configurations, evaluating the mid-span deflection would require the solution of a steady partial differential equation.

For the case at hand, in particular, the geometrical dimensions b (beam width), h (beam height) and L (beam length) are perfectly known: $b = 0.15$ m, $h = 0.3$ m, and $L = 5$ m. The applied load p is known up to some Gaussian measurement noise:

$$p \sim \mathcal{N}(\mu_p = 1.2 \times 10^{-2}, \sigma_p = 6 \times 10^{-4}) \qquad (kN/m);$$

on the other hand, the Young's modulus E, target of the calibration experiment, is given a lognormal prior distribution:

$$E \sim \mathcal{LN}(\mu_E = 3 \times 10^4, \sigma_E = 4.5 \times 10^3) \qquad (MPa).$$

An independent and identically distributed discrepancy $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ between the observations and the predictions for each data is assumed, where the variance σ^2 of the discrepancy term is by default given a uniform prior distribution:

$$\pi(\sigma^2) = \mathcal{U}(0, \mu_Z^2), \quad \text{with} \quad \mu_Z = \frac{1}{N}\sum_{j=1}^{N} z_j \quad \text{(here equal to 0.01259)}.$$

Below we report the prior distributions of (E, p, σ^2) and the posterior distributions obtained through a MCMC algorithm (here Metropolis-Hastings), see Figure 9.12 The mean a posteriori and the standard deviation a posteriori for each parameter, along with credibility intervals, are reported in Table 9.2. Finally, we also visualize the posterior predictive distribution $\pi''_{pred}(z_{new} \mid \mathbf{z})$ of new measurements given the data, with its mean and the data, in Figure 9.13.

Parameter	Mean	Std	(0.05-0.95) CI
E	$2.4 \cdot 10^4$	$2.1 \cdot 10^3$	$(2.2 \cdot 10^4, 2.8 \cdot 10^4)$
p	0.012	0.00059	$(0.011, 0.013)$
σ^2	$4.2 \cdot 10^6$	$1.3 \cdot 10^5$	$(1.3 \cdot 10^7, 1.9 \cdot 10^5)$

Table 9.2: Posterior Marginals

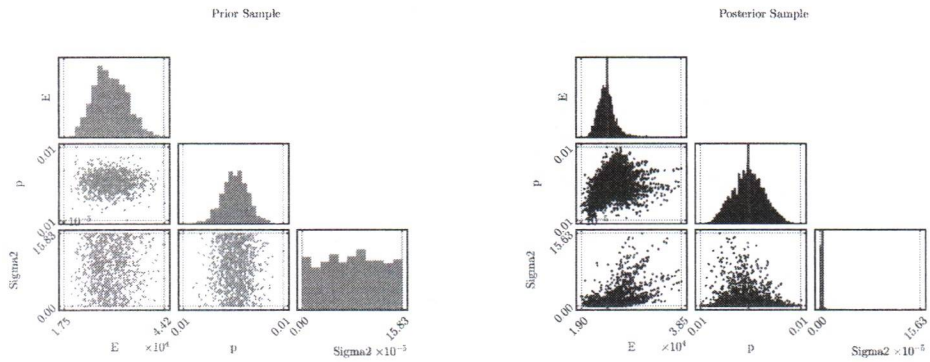


Figure 9.12: Left: samples from the prior distribution. Right: samples from the computed posterior distribution.

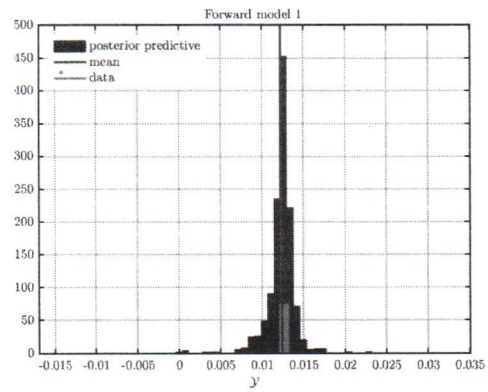


Figure 9.13: Posterior predictive distribution, its mean, and the observed data.

9.4 The theory behind MCMC Algorithms

Let us now try to make things more rigorous, and explain *why* the Metropolis-Hastings (and the random walk Metropolis) algorithms indeed provide a *good* approximation of the posterior distribution. Before doing that, let us recall what a Markov chain is. In this section notation will be slightly different than the one used so far, also to highlight the generality of some results and procedures.

9.4.1 What is a Markov chain?

A *Markov chain* is a particular type of *stochastic process*, that is a collection of random variables $\{\theta^{(t)} : t \in T\}$ with respect to the set of T indices. The process is valued in the $\mathcal{S}$ state space,

which can be finite or infinite. We assume that:

- The set T is always countable, that is, we will always consider discrete-time stochastic processes; it will be assumed without loss of generality $T = \mathbb{N}$, since for our purposes T must simply represent the successive iterations in a simulation scheme;
- the set $\mathcal{S}$ is generally a subset of $\mathbb{R}^p$, as it represents the support of the vector of the parameters of interest.

A Markov chain is a stochastic process in which, known the present state, past and future are independent; formally, this concept is expressed by the fact that the stochastic process enjoys the so-called *Markov property*, that is,

$$P(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \dots, \theta^{(0)} \in A_0) = P(\theta^{(n+1)} \in A | \theta^{(n)} = x), \quad (9.9)$$

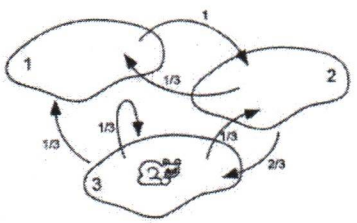
for all sets $A_0, \dots, A_{n-1}, A \in \mathcal{S}$.

The probability (9.9) depends on x, A and n; when it does not depend on n – i.e., the transition probabilities are the same for all times – then the chain is called *homogeneous*. In this case, it is possible to define a *transition (or Markov) kernel* $P(x, A)$ based on the following properties:

1. $P(x, \cdot)$ is a probability distribution on $\mathcal{S}$, for every $x \in \mathcal{S}$;
2. the function $x \mapsto P(x, A)$ can be evaluated, for each $A \subset \mathcal{S}$.

Hence, a Markov chain is characterized by a state space $\mathcal{S}$, an initial distribution p^0 and a transition kernel. The state space is the range of all random variables, that is, it is the set of all possible realizations.

- **Example 9.4.1.** *A frog hopping among lily pads can be described through a Markov chain.*



The state space is $\mathcal{S} = \{1, 2, 3\}$, *representing the pads; the initial distribution is* $p^0 = [1/2, 1/4, 1/4]$ *and the probability transition matrix is, in this case,*

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/3 & 0 & 2/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \longrightarrow [3 \to 1,\ 3 \to 2,\ 3 \to 3]$$

defines the probabilities of jumping from one state to another one. The frog chooses its initial position X_0 *according to the initial distribution* p^0*; the state after the first junp is the value of the random variable* X_1*, and so on...keeping on iterating, we would realize that the long range predictions are independent from the starting state (somehow, the frog forgets the initial state). Does this happen for all chains?*

Let us first consider the case in which $\mathcal{S}$ is given by a finite number k of discrete states, so $\mathcal{S} = \{x_1, \dots, x_k\}$. The initial distribution $p^0 = [p_1^0, \dots, p_k^0]$ quantifies the starting configuration for the chain, that is,

$$p_i^0 = P(X_0 = x_i), \qquad i = 1, \dots, k,$$

whereas the transition kernel quantifies the probability of transitioning from state x_i to state x_j, hence it establishes how the chain evolves. For a homogeneous Markov chain, let us denote by p_{ij} the probability of moving from x_i to x_j in one step, so that

$$p_{ij} = P(X_{n+1} = x_j | X_n = x_i) = (P)_{ij}, \qquad i, j = 1, \dots, k$$

and arrange all the entries into a matrix $P \in \mathbb{R}^{k \times k}$. The m-step transition matrix, encoding the probabilities of transitioning between states in m steps,

$$(P_m)_{ij} = P(X_{n+m} = x_j | X_n = x_i) = ((P)_{ij})^m,$$

yields the matrix $P_m = P^m$. All the entries of p^0 and the rows of P are nonnegative and sum to unity (P is a row-stochastic matrix).

Given an initial distribution and transition kernel, the distribution after 1 step is

$$p^1 = p^0 P$$

and, in general,

$$p^n = p^{n-1} P = p^0 P^n$$

after n steps. We ask if p^n converges in distribution to some distribution π for $n \to \infty$, that is,

$$\lim_{n \to \infty} p^n = \pi.$$

If the limit exists, it must fulfill

$$\pi = \lim_{n \to \infty} p^0 P^n = \lim_{n \to \infty} p^0 P^{n+1} = \left(\lim_{n \to \infty} p^0 P^n \right) P = \pi P.$$

For a Markov chain with transition kernel P, distributions π that satisfy $\pi = \pi P$ are called *stationary* distributions for the chain. For every finite Markov chain, there exists at least one stationary distribution. Uniqueness can be stated for irreducible and aperiodic Markov chains, where:

- a Markov chain is irreducible if any state x_j can be reached from any other state x_i in a finite number of steps, that is, $p_{ij}^{(m)} > 0$ for all state in finite m;
- a Markov chain is periodic if parts of the state space are visited at regular intervals. More formally, a state i has period m if any return to state i must occur in multiples of m time steps. The period of state i is defined as $m = \gcd\{n > 0 : \Pr(X_n = i \mid X_0 = i) > 0\}$ where gcd is the greatest common divisor, provided that this set is not empty. If $m = 1$, the state is said to be aperiodic[8], otherwise it is periodic with period m.

A fundamental result is the

Theorem 9.4.1. *(Basic Limit Theorem)*
Every homogeneous Markov chain defined on a finite set of states $\mathcal{S}$ that is irreducible and aperiodic, with transition matrix P, has a unique stationary distribution $\pi = [\pi_1, \dots, \pi_k]$ (with all its k component positive), and the chain converges in the sense of distributions, for every initial distribution p^0, that is,

$$\lim_{n \to \infty} p_{ij}^{(n)} = \pi_j, \qquad \forall i, j \in \mathcal{S}.$$

[8] A Markov chain is aperiodic if every state is aperiodic. An irreducible Markov chain only needs one aperiodic state to imply all states are aperiodic.

Remark 9.4.1. *An alternative formulation of the result above is the following one: for any* ergodic *Markov chain, there is a unique stationary distribution. A Markov chain is ergodic if there exists $T_0 > 0$ such that for all pairs of states x_i, x_j in the Markov chain, if it is started at time 0 in state x_i then for all $t > T_0$ the probability of being in state x_j at time t is greater than 0. For a Markov chain to be ergodic, the two conditions of irreducibility and aperiodicity must be satisfied.*

Every finite state Markov chain has at least one stationary distribution; a trivial example with infinitely many stationary distribution is the one in which P is the identity matrix, in which case all distributions are stationary. However, it is often difficult (if not impossible) to solve for π the equation

$$\pi P = \pi \quad \text{s.t.} \quad \sum_i \pi_i = 1$$

that is, to find a (normalized multiple of a left) eigenvector of the transition matrix P associated to an eigenvalue equal to 1.

The so-called *(detailed) balance* condition provides an alternative straightforward to implement in MCMC methods, where the goal is to construct Markov chains whose stationary distribution π is the posterior distribution for parameters. A chain with transition matrix P and stationary distribution π is reversible if the *detailed balance condition*

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j = 1, \ldots, k$$

is satisfied. Since

$$\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$$

it follows immediately that $\pi P = \pi$, so that reversibility implies stationarity. Hence, if the chains are irreducible and aperiodic, they will uniquely tend, at the limit, to this specified stationary distribution.

9.4.2 From a finite to an infinite state space*

This section can be skipped without compromising the comprehension of the following ones. Let us now turn to the case in which the state space $\mathcal{S}$ is a subset of $\mathbb{R}^p$, representing the parameter space, rather than a finite set of values, as we did when introducing Markov chains. This fact requires to slightly modify our point of view concerning the definition of the transition probability from a state to another in the state space.

Indeed, in the case where the state space is *continuous*, it is no longer possible to build a transition matrix; we will deal instead with a *transition kernel*. Since $P(x, \cdot)$ defines a probability distribution, we can denote by

$$P(x, y) = P(\theta^{(n+1)} \leq y | \theta^{(n)} = x) = P(\theta^{(1)} \leq y | \theta^{(0)} = x), \text{ for } x, y \in \mathcal{S},$$

so that the conditional density

$$p(x, y) = \frac{\partial P(x, y)}{\partial y}, \quad \text{for } x, y \in \mathcal{S}$$

can be used to define the transition core of the chain. From the previous definitions we easily get the transition core at m steps:

$$p^m(x, y) = \frac{\partial P^m(x, y)}{\partial y}, \quad \text{for } x, y \in \mathcal{S},$$

and the so-called *Chapman-Kolmogorov equations in the continuum*,

$$p^{n+m}(x, y) = \int_{-\infty}^{+\infty} p^n(x, z) P^m(z, y) dz, \; m, n \geq 0. \tag{9.10}$$

In particular, a distribution is stationary for the chain whose transition kernel is denoted by $p(x, y)$ if

$$\int_{-\infty}^{+\infty} \pi(x)p(x, y)dx = \pi(y) \qquad \forall y \in \mathcal{S}.$$

Moreover, let us denote by $P_x(A)$ the probability of any set $A \subset \mathcal{S}$ for a Markov chain starting at x. Furthermore, the *hitting time* of A is defined by $T_A = \min\{n \geq 1 : \theta^{(n)} \in A\}$, if $\theta^{(n)} \in A$ for some $n \geq 1$. The definitions of *irreducible* and *periodic* Markov chain also hold in the case of an infinite state space. In particular:

- a Markov chain is said to be irreducible if it is possible to get to any state from any state – that is, if all states communnicate. More formally, given a probability distribution ν, a chain is said to be ν-irreducible if for a set A such that $P_\nu(A) > 0$, we have that $P_x(T_A < +\infty) > 0$, $\forall x \in \mathcal{S}$. The chain is said to be *irreducible* if there is at least one distribution ν that makes it ν-irreducible.

- Recurrent sets $A \in \mathcal{S}$ are guaranteed (with probability 1) to have a finite hitting time T_A. If the chain starts from a recurrent[9] set A, the random variable T_A is almost certainly finite, and we can evaluate its expected value μ_A: if $\mu_A < +\infty$, then the set A is called *(positive) recurrent*. A set A is said to be transient if, given that we start in A, there is a non-zero probability that we will never return to A.

- The period d_A of a set A is given by $d_A = \gcd\{n > 0 : p_\nu^n(A) > 0\}$. A set A such that $p_\nu(A) > 0$ is aperiodic if $d_A = 1$. A set A is said to be ergodic if it is aperiodic and positive recurrent. If all states in an irreducible Markov chain are ergodic, then the chain is said to be ergodic[10].

It is clear from the definition of stationary distribution that if there exists a *limit distribution* of the chain, i.e. a distribution π such that

$$\lim_{n\to\infty} p^n(x, A) = \pi(A),$$

it must be a stationary distribution. However, there are situations where the stationary distribution exists, but it is not a limit distribution for the chain. This cannot happen if the chain is ergodic:

An irreducible Markov chain has a stationary distribution if and only if the Markov chain is ergodic. If the Markov chain is ergodic, the stationary distribution is unique.

This is stated by the

Theorem 9.4.2. *(***Basic Limit Theorem***, case of an infinite state space)*
Every homogeneous Markov chain defined on an infinite set of states $\mathcal{S}$ that is irreducible and recurrent, with transition kernel p, has a unique stationary distribution π. Moreover, if it is ergodic, it admits a limit distribution, which coincides with π, that is,

$$\lim_{n\to\infty} p^{(n)}(x, A) = \pi(A) \qquad \forall x \in \mathcal{S}.$$

[9] Transience and recurrence describe the likelihood of a process beginning in some state of returning to that particular state. There is some possibility (a nonzero probability) that a process beginning in a transient state will never return to that state. There is a guarantee that a process beginning in a recurrent state will return to that state.

[10] On the other hand, a finite state irreducible Markov chain is ergodic if it has at least an aperiodic state. More generally, a Markov chain is ergodic if there is a number N such that any state can be reached from any other state in any number of steps less or equal to a number N. In case of a fully connected transition matrix, where all transitions have a non-zero probability, this condition is fulfilled with $N = 1$.

Finally, let $(\boldsymbol{\theta}^{(n)})_{n\geq 0}$ be a continuous homogeneous Markov chain with transition kernel $p(x,y)$ and stationary distribution π. Suppose we want to study the sequence of states in reverse order: $\boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^{(n-1)}, \ldots$: we can prove that this sequence also defines a Markov chain.

Suppose that the inverse chain has a transition kernel $p(x,y)$ such that $p(x,y) = p(y,x), \ \forall\, x,y \in \mathcal{S}$. Then the Markov chain is said to be *reversible*, and the following detailed balance condition holds

$$\pi(x)p(x,y) = \pi(y)p(y,x), \qquad \forall\, x,y \in \mathcal{S}. \tag{9.11}$$

Why to consider reversible Markov chains? If there is a distribution π that satisfies (9.11) for an irreducible Markov chain, then the chain is also *positive recurrent*, and the distribution π is a *stationary* distribution. Once we have then verified that the chain is *aperiodic*, we can conclude that π is also the limit distribution. The construction of a Markov chain with limit distribution π therefore boils down to finding suitable transition kernels $p(x,y)$ that satisfy (9.11).

Reversible Markov chains are common in Markov chain Monte Carlo (MCMC) approaches because the detailed balance condition for a desired distribution π necessarily implies that the Markov chain has been constructed so that π is a stationary distribution.

9.4.3 Markov Chains + Monte Carlo integration = MCMC methods

Let us now come back to MCMC methods, that can be seen as general sampling methods, necessary of other purposes than estimating the posterior distribution in Bayesian inversion.

Definition 9.4.1. *A Markov Chain Monte Carlo (MCMC) method for estimating the expectation* $\mathbb{E}[\Psi(X)]$ *is a numerical method based on the approximation*

$$\mathbb{E}[\Psi(X)] \approx \frac{1}{N}\sum_{i=1}^{N} \Psi(X^{(i)})$$

where $\{X^{(i)}\}_{i\in\mathbb{N}}$ *is a Markov chain with the distribution of* X *as its stationary distribution.*

Where Monte Carlo methods uses independent samples, MCMC methods use samples which can have direct dependence on the immediately preceding value. We remark that our use of MCMC methods is meant to address two fundamental questions:

- generate sample from the posterior distribution, to *visualize* the distribution of the parameters informed by the data, and subsequently to *use* it, for instance to compute the MAP estimator $\boldsymbol{\theta}_{\text{MAP}}$ by maximizing it, or
- compute any statistics of interest related to the posterior distribution (such as the conditional mean $\boldsymbol{\theta}_{\text{CM}}$) relying on suitable Monte Carlo methods that exploit those samples.

Since samples generated from a Markov chain are not independent by construction, to ensure that those strategies are well-posed we need to extend to this case the basic results such as the *law of large numbers*, and the *central limit theorem*. The validity of these results is ensured provided that the Markov chain is ergodic.

The ergodic mean of a function $\Psi(X)$ defined over the state space $\mathcal{S}$ is given by

$$\overline{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n} \Psi(X^{(i)}),$$

where $X^{(i)}$ are the states of the Markov chain.

Theorem 9.4.3. *(Ergodic Theorem)*
For an ergodic Markov chain, provided that the expected value with respect to the (unique) limit distribution π is $\mathrm{E}_\pi[\Psi(X)] < \infty$, it holds that, for any bounded function $\Psi : \mathbb{R} \to \mathbb{R}$, when $n \to \infty$ the ergodic mean of Ψ is such that

$$\bar{\Psi}_n \xrightarrow{p} \mathrm{E}_\pi[\Psi(X)]. \tag{9.12}$$

This means that the sample mean of the states of the chain is a consistent estimator of the expected value of the limit distribution π, although the states are dependent from each other.

Theorem 9.4.4. *(Central Limit Theorem for Markov Chains)*
If a Markov chain is ergodic, provided that $\mathrm{E}_\pi[\Psi^2(X)] < \infty$, then

$$\sqrt{n}\frac{\bar{\Psi}_n - \mathrm{E}_\pi[\Psi(X)]}{\sigma_g} \rightharpoonup \nu \tag{9.13}$$

where ν is $N(0,1)$ and $\sigma_g^2 = \sigma^2\,(1 + 2\sum_{k=1}^{\infty}\rho_k)$, being

$$\rho_k = \frac{\gamma_k}{\sigma^2}, \qquad \gamma_k = \mathrm{Cov}_\pi(\Psi(X^{(n)}), \Psi(X^{(n+k)})), \qquad \sigma^2 = \gamma_0.$$

9.4.4 Metropolis Hastings and Gibbs Sampler revisited

MCMC methods can be employed to draw a sample from a distribution that approximates a given, target distribution π, as well as to draw a sample for the sake of evaluating (sample) estimates of expectations/moments of a given function of the random variable, this latter being distributed according to the target distribution π. Heuristically, an MCMC method consists in the construction of an irreducible and aperiodic *Markov chain*, having the target distribution π as *stationary distribution*; for n sufficiently large, a realization of the chain is *equivalent* to sampling from π. In particular:

- as for MC integration, the *accuracy* depends on the sample size, and the *approximation error* can be checked in probability;
- unlike MC techniques, as the size of the problem grows, convergence rate and computational burden do not slow down.

At each iteration, the Metropolis-Hastings algorithm picks a candidate for the next sample value based on the current sample value. Then, with some probability, the candidate is either accepted (in which case the candidate value is used in the next iteration) or rejected (in which case the candidate value is discarded, and current value is reused in the next iteration) with some probability of acceptance.

The transition kernel $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ of the chain is such that π results its stationary distribution. According to the detailed balance condition (9.11), a possible option is to choose p such that

$$\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})p(\boldsymbol{\phi}, \boldsymbol{\theta}) \qquad \forall\ (\boldsymbol{\theta}, \boldsymbol{\phi}); \tag{9.14}$$

indeed, the resulting chain is reversible, which is a sufficient condition for π to be a stationary distribution. In particular, we choose a kernel $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ under the form

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\ \alpha(\boldsymbol{\phi}, \boldsymbol{\theta}), \quad \text{if } \boldsymbol{\theta} \neq \boldsymbol{\phi}, \tag{9.15}$$

where $q(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ is an arbitrary transition kernel (from $\boldsymbol{\phi}$ to $\boldsymbol{\theta}$) called *proposal density* or *jumping distribution*, and $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})$ is an *acceptance probability*. This latter is then chosen so that the chain is reversible, that is, the probability of accepting the move from $\boldsymbol{\phi}$ to $\boldsymbol{\theta}$ is given by

$$\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta})q(\boldsymbol{\phi} \mid \boldsymbol{\theta})}{\pi(\boldsymbol{\phi})q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\right\}. \tag{9.16}$$

Any MCMC algorithm based on Markov chains with transition kernels under the form (9.15) and acceptance probability (9.16) is called **Metropolis – Hastings**; the choice of the kernel $q(\cdot, \cdot)$ is arbitrary, thus making the algorithm extremely flexible.

Metropolis–Hastings Algorithm.

1) set $k = 1$ and the initial value $\boldsymbol{\theta}_1 \in \mathbb{R}^p$ of the chain;

2) draw $\boldsymbol{\theta}^* \in \mathbb{R}^p$ from a proposal distribution $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_k)$ and calculate

$$\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}_k)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_k)}\right)$$

3) accept $\boldsymbol{\theta}^*$ with probability $\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*)$ by

- drawing an independent sample from $u \sim \mathcal{U}[0, 1]$;
  - if $\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) \geq u$, *accept* the step, and set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}^*$,
  - else, if $\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) < u$, *refuse* the step, and set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$,

4) set $k \leftarrow k + 1$ and go back to step 2), until convergence.

If we attempt to move to a point that is more probable than the existing point (i.e. a point in a higher-density region of $\pi(\boldsymbol{\theta})$), we will always accept the move. However, if we attempt to move to a less probable point, we will sometimes reject the move, and the more the relative drop in probability, the more likely we are to reject the new point. Thus, we will tend to stay in (and return large numbers of samples from) high-density regions of $\pi(\boldsymbol{\theta})$, and only occasionally visiting low-density regions — this is why this algorithm returns samples following the target $\pi(\boldsymbol{\theta})$.

Remark 9.4.2. *Note that the Metropolis-Hastings algorithm can draw samples from any probability distribution $\pi(\boldsymbol{\theta})$, provided that we know a function $\tilde{\pi}(\boldsymbol{\theta})$ proportional to the density of $\pi(\boldsymbol{\theta})$ and the values of $\tilde{\pi}(\boldsymbol{\theta})$ can be calculated. In this way,*

$$\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}^*) = \min\left(1, \frac{\tilde{\pi}(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}^*)}{\tilde{\pi}(\boldsymbol{\theta}_k)q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_k)}\right).$$

The requirement that $\tilde{\pi}(\boldsymbol{\theta})$ must only be proportional to the density, rather than exactly equal to it, makes the Metropolis-Hastings algorithm particularly useful, because calculating the necessary normalization factor is often extremely difficult in practice.

Why does the MH algorithm work? We sketch the proof in the simpler case of a finite dimensional set of states of the chain. Note that

$$\begin{aligned} p_{k-1,k} &= P(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}) = P(\text{proposing } \boldsymbol{\theta}_k)P(\text{accepting } \boldsymbol{\theta}_k) \\ &= q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})\,\alpha(\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k) \\ &= q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}) \min\left(1, \frac{\pi(\boldsymbol{\theta}_k)\, q(\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k)}{\pi(\boldsymbol{\theta}_{k-1})\, q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})}\right). \end{aligned}$$

= $\mathbb{P}$(transitioning $\vec{\theta}_{k-1} \to \vec{\theta}_k$)

We can show that the acceptance probability fulfills the detailed balance condition $\iff \pi P = \pi$

$$\pi(\boldsymbol{\theta}_{k-1})p_{k-1,k} = \pi(\boldsymbol{\theta}_k)p_{k,k-1}$$

namely: $\pi_i p_{ij} = \pi_j p_{ji}$

Indeed, from the relation

$$v \min\left(1, \frac{x}{v}\right) = \min(x, v) = x \min\left(1, \frac{v}{x}\right)$$

we obtain that

$$\begin{aligned}\pi(\boldsymbol{\theta}_{k-1}) p_{k-1,k} &= \pi(\boldsymbol{\theta}_{k-1}) q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1}) \min\left(1, \frac{\pi(\boldsymbol{\theta}_k)\, q(\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k)}{\pi(\boldsymbol{\theta}_{k-1})\, q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})}\right) \\ &= \pi(\boldsymbol{\theta}_k)\, q(\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k) \min\left(1, \frac{\pi(\boldsymbol{\theta}_{k-1})\, q(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_{k-1})}{\pi(\boldsymbol{\theta}_k)\, q(\boldsymbol{\theta}_{k-1} \mid \boldsymbol{\theta}_k)}\right) = \pi(\boldsymbol{\theta}_k) p_{k,k-1}.\end{aligned}$$

In the end we have discovered why the acceptance-rejection rule had to be expressed as $\alpha = \min(1, \ldots)$, because this way of expressing acceptance/rejection is what allows us to satisfy the detailed balance condition (the condition that ensures us that the target is the only stationary distribution)

Remark 9.4.3. *Although the Markov chain eventually converges to the desired distribution, the initial samples may follow a very different distribution, especially if the starting point is in a region of low density. As a result, a* burn-in *period is typically necessary, where an initial number of samples (e.g. the first 1000 or so) are thrown away.*

Gibbs sampler*

When the number of dimensions is high, finding the suitable proposal distribution to use can be difficult, as the different individual dimensions behave in very different ways. An alternative approach that often works better in such situations, known as *Gibbs sampling*, involves choosing a new sample for each dimension separately from the others, rather than choosing a sample for all dimensions at once. This is especially applicable when the multivariate distribution is composed of a set of individual random variables in which each variable is conditioned on only a small number of other variables. The individual variables are then sampled one at a time, with each variable conditioned on the most recent values of all the others. The *Gibbs sampling* is a convenient option provided the *full conditional* distributions $\pi_i(\theta_i)$ related to each component $\theta_i \mid \boldsymbol{\theta}_{\sim i}$ are easy to draw samples from. In this case, a cyclic update scheme of $\boldsymbol{\theta} \in \mathbb{R}^p$ can be considered, with:

- a proposal distribution $q^i(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ which proposes a new value for the i-th component of $\boldsymbol{\theta}$, $i = 1, \ldots, p$;
- the target distribution can be written as $\pi(\boldsymbol{\theta}) = \pi_i(\theta_i)\pi(\boldsymbol{\theta}_{\sim i})$, where π_i is the so-called *full conditional* distribution of θ_i;
- the shift proposed by q^i only refers to the component θ_i, hence $\boldsymbol{\theta}_{\sim i} = \boldsymbol{\phi}_{\sim i}$ and the ratio $\pi(\boldsymbol{\phi})/\pi(\boldsymbol{\theta})$ in (9.16) can be rewritten as $\pi_i(\phi_i)/\pi_i(\theta_i)$;
- since the proposal q^i only acts on θ_i, the term appearing in q^i and depending on $\boldsymbol{\theta}_{\sim i}$ can be simplified with the term involving $\boldsymbol{\phi}_{\sim i}$: hence,

$$\alpha_i(\phi_i, \theta_i) = \min\left(1, \frac{\pi_i(\theta_i) q^i(\phi_i \mid \theta_i)}{\pi_i(\phi_i) q^i(\theta_i \mid \phi_i)}\right); \tag{9.17}$$

- choosing as proposal the full conditional distribution of the component of $\boldsymbol{\theta}$ that has to be updated, the proposed value θ_i is drawn from its full conditional distribution, and is accepted with probability 1 – this can be seen by replacing $q^i(\theta_i \mid \phi_i) = \pi_i(\theta_i)$, $q^i(\phi_i \mid \theta_i) = \pi_i(\phi_i)$ in (9.17).

9.5 An Ideal UQ workflow

Summarizing the discussion so far, given a model, the ideal UQ workflow for predictions under uncertainty would consist of the following steps[11]:

[11] Bare in mind that this ideal workflow however is missing one step, i.e., the *identifiability analysis*, which is a crucial preliminary analysis to perform – here, for the sake of space, we will not take it into account.

Ideal UQ Workflow

1. Choose a prior distribution for the model parameters (literature, expert opinion);
2. Compute Sobol indices to assess which parameters are more influential and can be inferred from data. Fix the remaining parameters to some reasonable value;
3. Perform the inverse UQ analysis to adjust the prior distribution to the data evidence;
4. Perform the forward UQ analysis based on the posterior distribution to obtain statistical information about the quantities of interest of the model (e.g. expected value, variance, full pdf of the outputs).

↑ sensitivity analysis

The fundamental assumption underlying the inversion approach proposed in this Chapter is that there exists a certain set of parameters and hyper-parameters $\boldsymbol{\theta}_{true}$ that generated the observed data from a given differential system. We have then embraced the fact that data are noisy, and that this noise might prevent us from correctly determining the values $\boldsymbol{\theta}_{true}$; we therefore give up on giving a "one-shot" estimate of $\boldsymbol{\theta}_{true}$, and rather content ourselves with computing a posterior PDF, which quantifies our degree of belief on each possible value of $\boldsymbol{\theta}_{true}$. We have then computed the posterior PDF, and ideally three scenarios might then occur:

1. the posterior PDF is actually close to Gaussian;
2. the posterior PDF is unimodal but it departs from Gaussian in that it might show "heavy tails" and/or some degree of skewness. This would indicate that we might be introducing a bias that leads to over/underestimates;
3. the posterior PDF is multi-modal. This would mean that the inversion procedure is suggesting a few "likely" combinations of parameters $\boldsymbol{\theta}$, each corresponding to one peak of the posterior PDF: in this case the heights of the peaks represent our belief on the plausibility that such $\boldsymbol{\theta}$ is the "true one". The crucial point that one has to address is: can we guarantee that there is a unique set $\boldsymbol{\theta}_{true}$ that generates the observed outputs? Equivalently, is the inverse problem well-posed? If not, the MCMC approach also needs to be handled with care.

To show a practical example of how this ideal UQ workflow could be performed, let us consider the case of a SIR model, already addressed in Hands on 8. The *SIR model* (for $t > 0$)

$$\begin{cases} \dot{S} = & -\dfrac{\beta}{N_{pop}} IS \\ \dot{I} = & \dfrac{\beta}{N_{pop}} IS - rI \\ \dot{R} = & rI \end{cases}$$

is the simplest epidemiological model, and describes the time-evolution of three compartments: individuals (S)usceptible to the disease, individuals (I)nfected with the disease, and finally individuals (R)emoved from the disease dynamics (either because they recovered, assuming immunity after having contracted the disease, or died). The total number of individuals in the population $N_pop = S + I + R$ is supposed constant, and individuals transition from one compartment to the next one with certain transition rates β, r.

SIR-like models can be written as ODE systems for a state vector X with N_{states} components. The evolution of the system depends on N_{coef} coefficients $\mathbf{p} = (p_1, \ldots, p_{N_{coef}})$ – in the case of the SIR model, $\mathbf{p} = (\beta, r)$ – and on the N_{states} initial conditions $\mathbf{q} = (q_1, \ldots, q_{N_{states}})$. Our goal is to monitor some related quantities Y (Quantities of Interest) which can be derived from X by an

observation operator G, that in turn might depend on some hyper-parameters $\mathbf{h}$:

$$\begin{cases} \dot{X} = f(X, \mathbf{p}), & t \in (0, T) \\ X(0) = \mathbf{q} \\ Y(t) = G(X(t), \mathbf{h}) \end{cases}$$

Possible quantities of interest might be:

- the peak-time of number of infected persons, $G(X) = \arg\max_{t\in[0,T]} I(t)$;
- the cumulative number of infected persons, $G(X) = \int_0^T I(t)dt$, also called incidence data (as opposed to prevalence (instantaneous) data $G(X(t)) = I(t)$);
- the peak-time of the new infected persons in a time-window of length Δ, $G(X, \Delta) = \arg\max \int_t^{t+\Delta} \frac{\beta}{N_{pop}} S(s)I(s)ds$.

In particular, we consider a SIR model with initial conditions

$$S(0) = 0.95, \qquad I(0) = 0.05, \qquad R(0) = 0.$$

and choose these ranges for the parameters:

$$\beta \in [0.25, 0.35], \qquad r \in [0.06, 0.18].$$

We assume that a-priori we have no knowledge that any value of β, r is more plausible than others; therefore, we assume that β, r are uniform independent random variables. These ranges for the parameters are suggested from a survey of the literature. Moreover, we solve the SIR system with Matlab's `ode45` up to final time $T = 150$.

In the case we have no further information from recorded data, we can compute the SIR dynamics by 100 Monte Carlo samples, obtained by sampling β and r independently, plot the obtained trajectories of S, I and B, highlighting the sample average trajectories (see Figure 9.14).

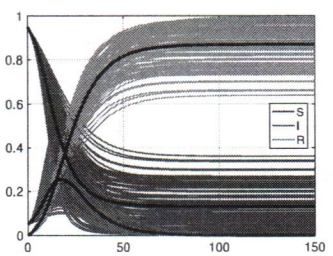


Figure 9.14: 100 realizations of the SIR dynamics and MC means of S, I and R obtained sampling uniformly the parameter space.

Moreover, we evaluate the PDFs of the following quantities of interest: SIR states at $T = 30$ (after the average peak position) and $T = 100$ (when the dynamics is over), see Figure 9.15; peak time for I; peak value for I, see Figure 9.16.

We can also compute the Sobol' indices for the SIR compartments S, I and R, as functions of time. First-order and total effect indices behave very similarly, showing that the interaction between the two parameters is quite limited. We highlight that the Sobol' indices are not constant in time, and behave differently for the different compartments. More specifically, the asymptotic regime is mostly dictated by r for all the compartments, while β impacts more in the transient regime, especially in the case of the compartment R. This has an impact on the inversion procedure. In particular, severe difficulties in the estimation of β can be encountered if the data of R are missing or too noisy. Moreover, note that the influence of r is larger, in general. Further

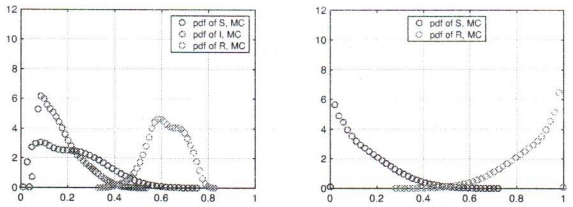


Figure 9.15: Left: PDFs of S, I and R at $t = 30$. Right: PDFs of S and R at $t = 100$.

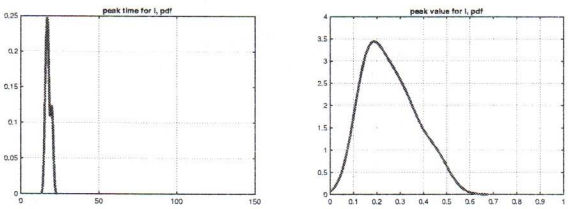


Figure 9.16: PDFs of the peak time (left) and the peak value (right).

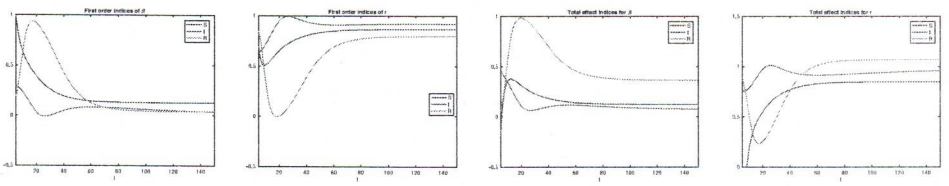


Figure 9.17: From top, left to bottom, right: First order indices of β on S, I, R as a function of time; first order indices of r on S, I, R as a function of time; total effect indices of β on S, I, R as a function of time; total effect indices of r on S, I, R as a function of time.

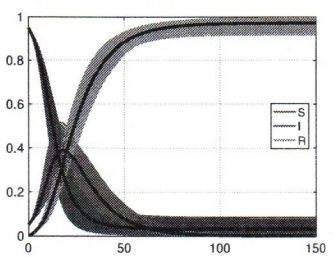


Figure 9.18: SIR trajectories obtained by reducing the range of variability of r.

evidence of this is shown in Figure 9.18, where we show the variability of the trajectories if the range of r is reduced to $[0.06, 0.1]$. The overall variability is greatly reduced, as expected.

So far, we have not taken into account data. Suppose now that we can rely on N_{meas} measurements of the I state and N_{meas} measurements of the R state, at equispaced times $t_i = i\Delta t$, $i = 1, 2, 3, \ldots, N_{meas}$. In total we have $2N_{meas}$ data, $\mathcal{D} = \{I_1^*, I_2^*, \ldots, R_1^*, R_2^*, \ldots\}$. We assume that these data correspond to some values $\boldsymbol{\theta}_{true}$ of coefficients of the SIR system, and that the prior PDF for $\boldsymbol{\theta}$ is uniform. Moreover, we assume that measurements are under-reported by a factor K (constant in time, known), that is, it is possible to measure only a fraction of the actual compartments. Data are affected by some random errors ε_i^I, ε_i^R, modeled by independent Gaussian random variables with zero mean and standard deviation σ – equal for I and R, and unknown.

The likelihood function becomes, in this case,

$$\pi(\mathcal{D} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^n} \prod_{i=1}^{N_{meas}} e^{-\frac{1}{2\sigma^2}\left(\frac{1}{K} I(t_i;\boldsymbol{\theta}) - I_i^*\right)^2} \prod_{i=1}^{N_{meas}} e^{-\frac{1}{2\sigma^2}\left(\frac{1}{K} R(t_i;\boldsymbol{\theta}) - R_i^*\right)^2}$$

so that the posterior PDF of the parameters reads as

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \mathcal{D}) \propto \; & \pi_{prior}(\boldsymbol{\theta})\pi(\mathcal{D} \mid \boldsymbol{\theta}) \\ = \; & \frac{1}{(2\pi\sigma^2)^n} \prod_{i=1}^{N_{meas}} e^{-\frac{1}{2\sigma^2}\left(\frac{1}{K} I(t_i;\boldsymbol{\theta}) - I_i^*\right)^2} \prod_{i=1}^{N_{meas}} e^{-\frac{1}{2\sigma^2}\left(\frac{1}{K} R(t_i;\boldsymbol{\theta}) - R_i^*\right)^2} \pi_{prior}(\boldsymbol{\theta}). \end{aligned}$$

Following [37], we show the results of the inversion procedure using artificial/synthetic data (see Figure 9.19), by fixing the values of the parameters to $\boldsymbol{\theta}_{true} = [0.29, 0.09]$, adding numerical Gaussian noise with $\sigma = 0.025$, discount factor to $K = 3$, considering data collected at $t = 1, 2, \ldots, 30$, and verifying the results of the inversion procedure.

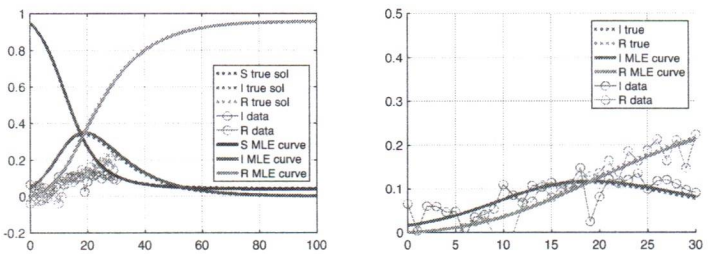


Figure 9.19: Trajectories corresponding to $\boldsymbol{\theta} = \boldsymbol{\theta}_{true}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_{MLE}$, and synthetic data obtained dividing the trajectories for $\boldsymbol{\theta} = \boldsymbol{\theta}_{true}$ by the under-reporting factor K and adding the Gaussian noise; zoom on the data, and trajectories for $\boldsymbol{\theta} = \boldsymbol{\theta}_{true}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}_{MLE}$ rescaled by K.

Regarding the inversion procedure, compared to the paper [37], we use instead the DRAM algorithm provided in the `mcmcstat` package, similarly to what we have done in Hands On 9. In particular, we use a Gaussian proposal distribution, with adapted covariance matrix, and a uniform prior distribution, starting the chains from the maximum likelihood estimate,

$$\hat{\beta} = 0.28482, \qquad \hat{r} = 0.086111, \qquad \hat{\sigma} = 0.027919.$$

After generating 10000 samples, from the DRAM algorithm we obtain the sample reported in Figure 9.20. Posterior distributions are compared to the prior distributions in Figure 9.21. The mean and the standard deviation of the chain, and MCMC convergence diagnostic (see Hands on 9) are reported below:

	mean	std	MC_err	tau	geweke
beta	0.28572	0.009401	0.00024387	6.7517	0.99703
r	0.086337	0.0052684	0.00012146	6.6667	0.99362

We then perform the forward UQ based on the posterior PDF. We clearly see that the uncertainty in the prediction is smaller than what would be obtained by using the prior information only (see Figure 9.22), and the expected values of the quantities of interest is closer to the true values when using the posterior PDF than when using the prior (see Figures 9.23–9.24).

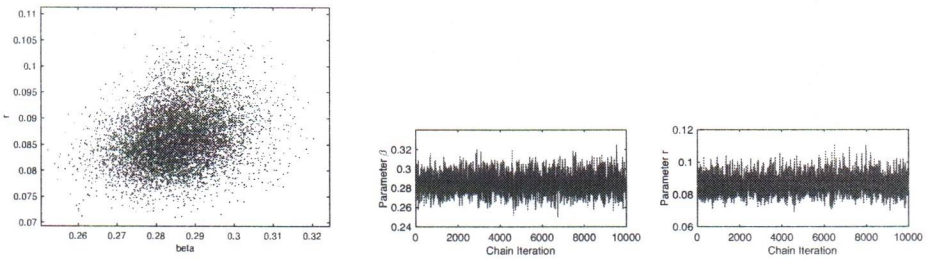


Figure 9.20: Computed posterior distribution of (β, r) through the DRAM MCMC algorithm.

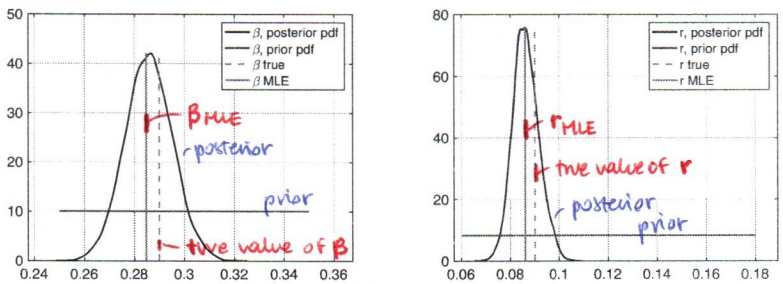


Figure 9.21: Computed posterior distribution of (β, r) through the DRAM MCMC algorithm, compared to the prior distribution, and the value $\boldsymbol{\theta}_{true}$ of the parameters.

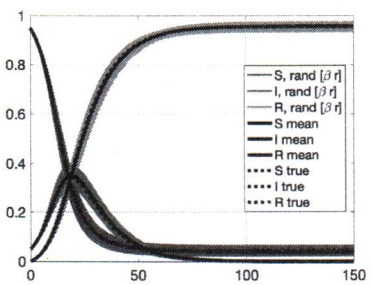


Figure 9.22: 100 realizations of the SIR dynamics and MC means of S, I and R obtained sampling the posterior distribution of (β, r).

9.6 Further readings, code and material

General references for an introduction to inverse UQ in the Bayesian framework are the review paper by A. Stuart [46], and the book by Kaipio and Somersalo [22]. The book by R. Smith [45], and the textbook by T.J. Sullivan [47], also provide an in-depth survey of the topic, as well as of several extensions of the MCMC techniques introduced in this chapter. See also, e.g., [4, Chapter 13], [9], [24]; the bible of Monte Carlo simulation is [39]. Some chapters of the Handbook of Uncertainty Quantification [17] are also devoted to inverse UQ.

Regarding code libraries and packages, the list - as usual - is endless. Two list of libraries can be found at:

- `https://uqworld.org/t/various-uncertainty-quantification-software-tools/137`
- `https://en.wikipedia.org/wiki/List_of_uncertainty_propagation_software`

Summary of this workflow: (UNCERTAINTY QUANTIFICATION WORKFLOW)

- forward UQ and sensitivity analysis to understand what is the impact of the parameters on the output
- MCMC estimation through a delayed rejection adaptive Metropolis algorithm to acquire informations from data and identify parameters
- propagate uncertainty through the (new) posterior parameters distribution to better inform the phenomenon through the combination of model and the data

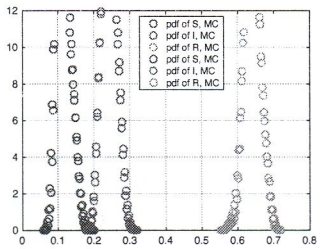


Figure 9.23: PDFs of S, I and R at $t = 30$ obtained by sampling (β, r) from the posterior distribution.

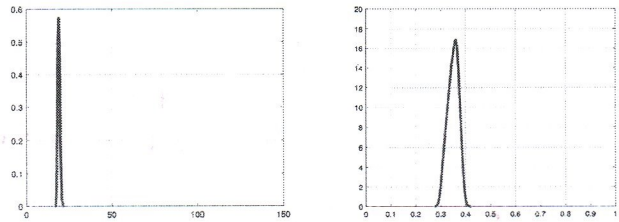


Figure 9.24: PDFs of the peak time (left) and the peak value (right) obtained by sampling (β, r) from the posterior distribution.

- whereas a list of implementations of MCMC algorithms in Python can be found at `https://github.com/Gabriel-p/pythonMCMC`.

An useful stand-alone Matlab implementation of MCMC algorithms is the `mcmcstat` package which can be found at `https://mjlaine.github.io/mcmcstat/`; a recent adaptation in Python can be found at `https://github.com/prmiles/pymcmcstat`. `UQlab` implements in Matlab several methods for inverse UQ, among others. Check also `https://github.com/Zaijab/DREAM/` for a Matlab implementation of DREAM MCMC algorithms.

Possible Python libraries for MCMC and inverse UQ can be found at

- `UQpy`, `https://github.com/SURGroup/UQpy`
- `MUQ`, `http://muq.mit.edu`
- `UQ Toolkit (UQTk)`, `http://www.sandia.gov/UQToolkit/`
- pymc3, Probabilistic Programming in Python: Bayesian Modeling and Probabilistic Machine Learning, `https://github.com/pymc-devs/pymc3`
- `https://mc-stan.org/users/interfaces/`
- `http://edwardlib.org`

BAYESIAN → parametric framework

frequentista vs. Bayesiana

- probabilità (oggettive/non)
- parametri (fissi/non)
- var. aleatorie (stimatore vs. parametro)

parametri → var. aleatorie con densità proprie : da Bayes' $\pi(\vec{\theta}|\vec{z}) = \frac{\pi(\vec{z}|\vec{\theta})\pi(\vec{\theta})}{\pi_{\vec{z}}(\vec{z})}$ (posterior; likelihood · prior)

prior: informative/un-informative, proper/non

likelihood: • i dati informano la posterior tramite la likelihood
• $\mathbb{P}$(dati | parametro fissato)

$\pi_{\vec{z}}(\vec{z}) = \int \dots$ calcolabile solo in casi particolari

Ex. evento $\sim Be(\theta)$

$\pi(\vec{z}|\theta) = \prod_{i=1}^{N} \theta^{z_i}(1-\theta)^{1-z_i} = \theta^{N_1}(1-\theta)^{N_0}$ } $\Rightarrow \pi(\theta|\vec{z}) \overset{d}{=}$ gamma

$\pi_0(\theta) \overset{d}{=} \mathcal{U}([0,1])$ uninformative

$\pi_0(\theta) \overset{d}{=} N(\mu, \sigma^2)$ ma μ sbagliata → peggio che $\mathcal{U}([0,1])$

$\Rightarrow$ piuttosto che una prior informative sbagliata è meglio una prior non-informative

prior-posterior coniugate $\pi_0(\vec{\theta}) \overset{d}{=} \pi(\vec{\theta}|\vec{z})$

(i parametri di $\pi_0(\vec{\theta})$ e $\pi(\vec{\theta}|\vec{z})$ sono detti IPERPARAMETRI)

Ex. evento $\sim N(\mu, \sigma^2)$, $\mu \sim N(\mu_0, \sigma_0^2)$

$\Rightarrow \pi(\vec{z}|\mu) \overset{d}{=} N$ } $\Rightarrow \pi(\mu|\vec{z}) \propto \pi(\vec{z}|\mu)\cdot\pi_0(\mu) \overset{d}{=} N$

$\pi_0(\mu) \overset{d}{=} N$

$\mu \sim N(\mu_0, \sigma_0^2)$ $\mu|\vec{z} \sim N(\mu_1, \sigma_1^2)$

$\mu_1 = w_0\mu_0 + w_1 \overline{\text{evento}}_n$ (media campionaria)

0	1	se abbiamo tanta incertezza nella prior / se n è grande
1	0	se abbiamo poche osservazioni

la prior è aggiustata "verso" la media campionaria secondo un gain.

$\mu_1 = \mu_0 + \underbrace{\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}}_{\text{gain}}(\overline{\text{evento}}_n - \mu_0)$

(lo stesso gain è sfruttato per la varianza)

Bayesian framework:

$f(\cdot)$ = computational forward model that links observations and param.

$\vec{\theta}$ = input parameters (with their own densities / or congiunta $\pi_{\vec{\theta}}(\cdot)$)

$\vec{z}$ = experimental data

$\vec{z} = f(\vec{\theta}) + \vec{\varepsilon}$ discrepancy (additive, mutually independent (from $\vec{\theta}$))

$$\vec{z} = f(\vec{\theta}) + \vec{\varepsilon}$$

$$\pi(\vec{z}|\vec{\theta}) = L(\vec{\theta}|\vec{z}) = \text{likelihood function}$$
$$= \pi_{noise}(\vec{z} - f(\vec{\theta}))$$

dato un valore di $\vec{\theta}$, la distribuzione di $\vec{z}|\vec{\theta}$ è uguale a quella di $\vec{\varepsilon}$ shiftato (di $f(\vec{\theta})$).
⟹ la likelihood riflette le nostre assunzioni sull'errore

Nel caso di problemi stazionari (parametri ⊥ tempo)
il goal è:

$\pi_{prior}(\vec{\theta})$, $\pi(\vec{z}|\vec{\theta})$ → $\pi_{posterior}(\vec{\theta}|\vec{z})$ →

evaluate a point estimator
- MAP (maximum aposteriori)
  $\hat{\theta}_{MAP} = \arg\max_{\vec{\theta}} \pi(\vec{\theta}|\vec{z})$
- CM (conditional mean)
  $\hat{\theta}_{CM} = \mathbb{E}[\vec{\theta}|\vec{z}] = \int \vec{\theta}\,\pi(\vec{\theta}|\vec{z})\,d\vec{\theta}$

/ evaluation of variability
→ $Cov(\vec{\theta}|\vec{z})$
→ confidence regions

→ predictive distributions
$\pi_{pred}(z_{new}|\vec{z}) =$
$= \int \pi(z_{new}|\vec{\theta})\,\pi(\vec{\theta}|\vec{z})\,d\vec{\theta}$

Caso più semplice: $\vec{\varepsilon} \sim N(\vec{0}, \Sigma_\varepsilon)$
$\pi_{prior} \sim N(\theta_p^{\vec{}}, \Sigma_p)$

Se p (dim($\vec{\theta}$)) è grande → implementare un metodo per ottenere la posterior $\pi(\vec{\theta}|\vec{z})$ può diventare challenging

⟹ al posto di valutare $\pi(\vec{\theta}|\vec{z})$, la tecnica MCMC è un modo per produrre sistematicamente campioni per esplorare la distribuzione $\pi(\vec{\theta}|\vec{z})$.
(e può poi essere sfruttata per calcolare conditional means/variance)

→ Il goal delle MCMC è di costruire una catena la cui distribuzione stazionaria è la posterior - in altre parole, la posterior $\pi(\vec{\theta}|\vec{z})$ è la target probability distribution che vogliamo esplorare ed è ottenuta tramite realizzazioni di catene di Markov.

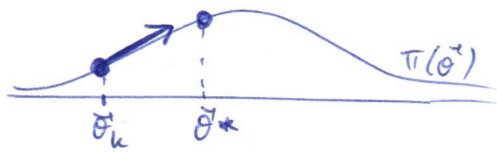
→ the posterior is not used directly. The stationary distr. of the constructed MC is the posterior.

→ the chain is s.t. $\vec{\theta}_{k+1}$ depends only on $\vec{\theta}_k$ (e quindi forma una MC).

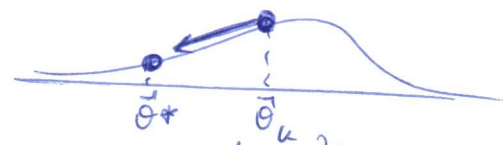
METROPOLIS con target $\pi(\vec{\theta})$

1. choose $\vec{\theta}_1$ and set $k=1$
2. campiona $\vec{\theta}^*$ dalla proposal $q(\cdot \mid \vec{\theta}_k)$
3. accetta con probabilità $\alpha(\vec{\theta}_k, \vec{\theta}^*) = \min\left\{1, \frac{\pi(\vec{\theta}^*)}{\pi(\vec{\theta}_k)}\right\}$ e fissa $\vec{\theta}_{k+1} = \vec{\theta}^*$ e $k = k+1$, altrimenti 2.

- la costante di normalizzazione di $\pi(\vec{\theta})$ non serve
- $q(\cdot \mid \vec{\theta}_k)$ deve essere simmetrica: $q(\vec{\theta}^* \mid \vec{\theta}_k) = q(\vec{\theta}_k \mid \vec{\theta}^*)$
  → METROPOLIS - HASTINGS
  estensione per $q(\cdot \mid \vec{\theta}_k)$ non simmetrica

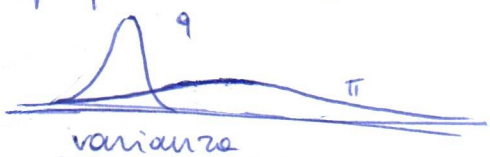


always accepted

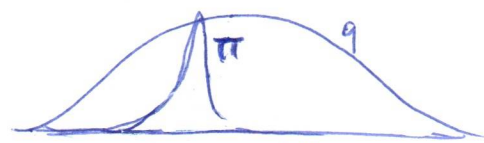


accepted with prob. $\alpha = \frac{\pi(\vec{\theta}^*)}{\pi(\vec{\theta}_k)}$

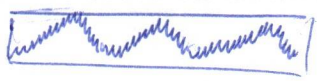
- proposal: facile da campionare e vicina a $\pi(\vec{\theta})$



varianza della proposal troppo piccola → poco esplorativa



varianza della proposal troppo grande → si ritrovano the essential support di π (si generano tanti tentativi che si rifiutano)



- multivariate case: gaussian proposal centrata in $\vec{\theta}_1$. Il problema è capire direzione e covarianza $\mathbf{C}$.
  (orientamento)

$$\pi_{prior}(\vec{\theta}) \propto 1$$
$$\pi(\vec{\theta} \mid \vec{z}) \propto \pi(\vec{z} \mid \vec{\theta}) \cdot \pi_{prior}(\vec{\theta}) \propto e^{-\frac{SS(\vec{\theta})}{2\sigma^2}}$$
$$SS(\vec{\theta}) = \sum_{i=1}^{n} (z_i - f_i(\vec{\theta}))^2$$
$$\Longrightarrow \alpha(\vec{\theta}_k, \vec{\theta}^*) = \min\left\{1, e^{-\frac{SS(\vec{\theta}^*) - SS(\vec{\theta}_k)}{2}}\right\}$$

METROPOLIS, uniform prior

1. Genera $\vec{\theta}_{new} \sim N(\vec{\theta}_{old}, C)$ e calcola $SS(\vec{\theta}_{new})$
2. $u \sim U([0,1])$, accetta $\vec{\theta}_{new}$ se $u < e^{-(SS(\vec{\theta}_{new}) - SS(\vec{\theta}_{old}))/2}$
   → se accettato: $\vec{\theta}_{old} = \vec{\theta}_{new}$, $SS(\vec{\theta}_{old}) = SS(\vec{\theta}_{new})$
   → se rifiutato ⇒ repeat $\vec{\theta}_{old}$ nella chain
3. Torna a 1. fino a che non si ha una catena lunga abb.

$\mathbf{C}$?
- → approx. of the covariance ottenuta via linearization
  $$\mathbf{C} = s_p \cdot \sigma^2 \left(J(\vec{\theta}_{LS})^T J(\vec{\theta}_{LS})\right)^{-1}$$
- → adaptive version of $\mathbf{C}$: si aggiorna C in real time usando le info del punto appena sampled
  ⇒ ADAPTIVE METROPOLIS: $\mathbf{C}_k$

(• gaussian multivariate proposal)

METROPOLIS - generic prior

1. Choose $\vec{\theta}_0$: $\pi(\vec{\theta}_0 | \vec{z}) > 0$
2. $k = 1, .., M$:
   - $z \sim N(0,1)^p$: $\vec{\theta}^* = \vec{\theta}_{k-1} + Rz$ (R = cholesky fact. di C) ← sampling da una multivariate gaussian
   - accetto con prob. $\alpha = ..$