

The praise of ignorance: randomness as lack of information

“So you don’t have a unique answer to your questions?”

“Adson, if I had, I would teach theology in Paris.”

“Do they always have a right answer in Paris?”

“Never”, said William, “but there they are quite confident of their errors.”

(Umberto Eco: *The Name of the Rose*)

In the previous chapter the problem of statistical inference was considered from the frequentist’s point of view: the data consist of a sample from a parametric probability density and the underlying parameters are *deterministic* quantities that we seek to estimate based on the data. In this section, we adopt the Bayesian point of view: *randomness simply means lack of information*. Therefore *any* quantity that is not known exactly is regarded as a random variable. The subjective part of this approach is clear: even if we believed that an underlying parameter corresponds to an existing physical quantity that could, in principle, be determined and therefore is conceptually a deterministic quantity, the lack of the *subject’s* information about it justifies modeling it as a random variable. This is the general guiding principle that we will follow, applying it with various degrees of rigor¹.

When applying statistical techniques to inverse problems, the notion of *parameter* needs to be extended and elaborated. In classical statistics, parameters are often regarded as tools, like, for example, the mean or the variance, which identify a probability density. It is not uncommon that even in that context, parameters may have a physical² interpretation, yet they are treated as abstract parameters. In inverse problems, parameters are more often *by*

¹ After all, as tempting as it may sound, we don’t want to end up teaching theology in Paris.

² The word “physical” in this context may be misleading in the sense that it makes us think of physics as a discipline. The use of the word here is more general, almost a synonym of “material” or “of observable nature”, as opposed to something that is purely abstract.

definition physical quantities, but they *appear* as statistical model parameters defining probability densities. Disquisitions about such differences in interpretation may seem unimportant, but these very issues often complicate the dialogue between statisticians and “inversionists”.

EXAMPLE 3.1: Consider the general inverse problem of estimating a quantity $x \in \mathbb{R}^n$ that cannot be observed directly, but for which indirect observations of a related quantity $y \in \mathbb{R}^m$ are available. We may, for example, want to know the concentrations of certain chemical species (variable x) in a gas sample, but for some reason, we cannot measure them directly; instead, we observe spectral absorption lines of light that passes through the specimen (variable y). A mathematical model describing light absorption by a mixture of different chemical compounds ties these quantities together. The fact that the variables that we are interested in are concentration values already carries *a priori* the information that they cannot take on negative values. In addition, knowing where the sample is taken from, regardless of the subsequent measurement, we may have already a relatively good idea of what to expect to be found in the gas sample. In fact, the whole process of measuring may be performed to confirm a hypothesis about the concentrations.

In order to set up the statistical framework we need to express the distribution of y in terms of the parameter x . This is done by constructing the *likelihood model*. The design of the prior model takes care of incorporating any available *prior* information.

As the preliminary example above suggests, the statistical model for inverse problems comprises two separate parts:

- The construction of the likelihood model;
- The construction of the prior model,

both of which make extensive use of *conditioning* and *marginalization*. When several random variables enter the construction of a model, by means of conditioning we can take into consideration one unknown at the time pretending that the others are given. This allows us to construct complicated models step by step. For example, if we want the joint density of x and y , we can write the density of y for fixed x and then deal with the density of x alone,

$$\pi(x, y) = \pi(y \mid x)\pi(x).$$

If, on the other hand, some of the variables appearing in the model are of no interest, we can eliminate them from the density by marginalizing them, that is by integrating them out. For example, if we have the joint density $\pi(x, y, v)$ but we are not interested in v , we can marginalize it as follows:

$$\pi(x, y) = \int \pi(x, y, v) dv.$$

The parameter v of no interest is often referred to as noise or as a nuisance parameter.

As statistics and probability are “common sense reduced to calculations”, there is no universal prescription for the design of priors or likelihoods, although some recipes seem to be used more often than others. These will be discussed next.

3.1 Construction of Likelihood

In the previous section, the parametric likelihood density was viewed as a probability density from which, presumably, the observed data was generated. In the Bayesian context, the meaning of the likelihood is the same, with the only difference that parameters are seen as realizations of random variables. When discussing inverse problems, the unknown parameters always include the variables that we are primarily interested in. Hence, we can think of the likelihood as of the answers to the following question: *If we knew the unknown x and all other model parameters defining the data, how would the measurements be distributed?*

Since the construction of the likelihood starts from the assumption that, if x were known, the measurement y would be a random variable, it is important to understand the source of its randomness. Randomness being synonymous of lack of information, it suffices to analyze what makes the data deviate from the predictions of our observation model. The most common sources of deviations are

1. measurement noise in the data;
2. incompleteness of the observation model.

The probability density of the noise can, in turn, depend on unknown parameters, as will be demonstrated later in the examples. The second source of randomness is more complex, as it includes errors due to discretization, model reduction and more generally, all the shortcomings of a computational model, that is the discrepancy between the model and “reality” – in the heuristic sense of the word³.

EXAMPLE 3.2: In inverse problems, it is very common to use additive models to account for measurement noise, as we did in the previous chapter. Assume that $x \in \mathbb{R}^n$ is the unknown of primary interest, that the observable quantity $y \in \mathbb{R}^m$ is ideally related to x through a functional dependence,

$$y = f(x), \quad f : \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad (3.1)$$

³ Writing a model for the discrepancy between the model and reality implicitly, and arrogantly, assumes that we know the reality and we are able to tell how the model fails to describe it. Therefore, the word “reality” is used in quotes, and should be understood as the “most comprehensive description available”.

and that we are very certain of the validity of the model. The available measurement, however, is corrupted by noise, which we attribute to external sources or to instabilities in the measuring device, hence not dependent on x . Therefore we write the *additive noise model*,

$$Y = f(X) + E,$$

where $E : \Omega \rightarrow \mathbb{R}^m$ is the random variable modeling the noise. Observe that since X and Y are unknown, they are interpreted as random variables – hence upper case letters –, leading to a *stochastic extension* of the deterministic model (3.1).

Let us denote the distribution of the error by

$$E \sim \pi_{\text{noise}}(e).$$

Since we assume that the noise does not depend on X , fixing $X = x$ does not change the probability distribution of E . More precisely

$$\pi(e \mid x) = \pi(e) = \pi_{\text{noise}}(e).$$

If, on the other hand, X is fixed, the only randomness in Y is due to E . Therefore

$$\pi(y \mid x) = \pi_{\text{noise}}(y - f(x)), \quad (3.2)$$

that is, the randomness of the noise is translated by $f(x)$, as illustrated in Figure 3.1.

In this example we assume that the distribution of the noise is known. Although this is a common assumption, in practice the distribution of the noise is seldom known. More typically, the noise distribution itself depends on unknown parameters θ , thus

$$\pi_{\text{noise}}(e) = \pi_{\text{noise}}(e \mid \theta),$$

hence equation (3.2) becomes

$$\pi(y \mid x, \theta) = \pi_{\text{noise}}(y - f(x) \mid \theta).$$

To illustrate this, assume that the noise E is zero mean Gaussian with unknown variance σ^2 ,

$$E \sim \mathcal{N}(0, \sigma^2 I),$$

where $I \in \mathbb{R}^{m \times m}$ is the identity matrix. The corresponding likelihood model is then

$$\pi(y \mid x, \sigma^2) = \frac{1}{(2\pi)^{m/2} \sigma^m} \exp \left(-\frac{1}{2\sigma^2} \|y - f(x)\|^2 \right),$$

with $\theta = \sigma^2$. If the noise variance is assumed known, we usually do not write the dependency explicitly, instead using the notation

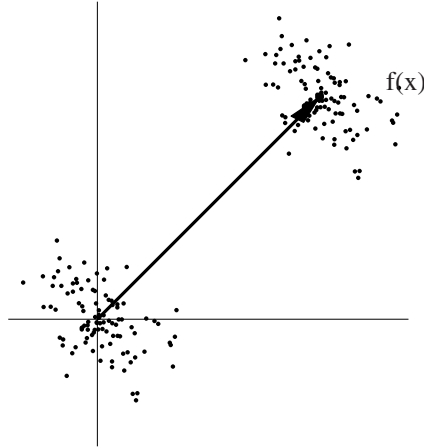


Fig. 3.1. Additive noise: the noise around the origin is shifted to a neighborhood of $f(x)$ without otherwise changing the distribution.

$$\pi(y | x) \propto \exp \left(-\frac{1}{2\sigma^2} \|y - f(x)\|^2 \right),$$

hence ignoring the normalizing constant.

In the previous example we started from an ideal deterministic model and added independent noise. This is not necessarily always the case, as the following example shows: the forward model may be intrinsically probabilistic.

EXAMPLE 3.3: Assume that our measuring device consists of a collector lens and a counter of photons emitted from N sources, with average photon emission per observation time equal to x_j , $1 \leq j \leq N$, and that we want to estimate the total emission from each source over a fixed time interval. We take into account the geometry of the lens by assuming that the total photon count is the weighted sum of the individual contributions. When the device is above the j th source, it collects the photons from its immediate neighborhood. If the weights are denoted by a_k , with the index k measuring the offset to the left of the current position, the *expected* count is

$$\bar{y}_j = \mathbb{E}\{Y_j\} = \sum_{k=-L}^L a_k x_{j-k},$$

where the weights a_j are determined by the geometry of the lens and the index L is related to the width of the lens, as can be seen in Figure 3.2. Here it is understood that $x_j = 0$ if $j < 1$ or $j > N$.

Considering the ensemble of all source points at once, we can write

$$\bar{y} = \mathbb{E}\{Y\} = Ax,$$

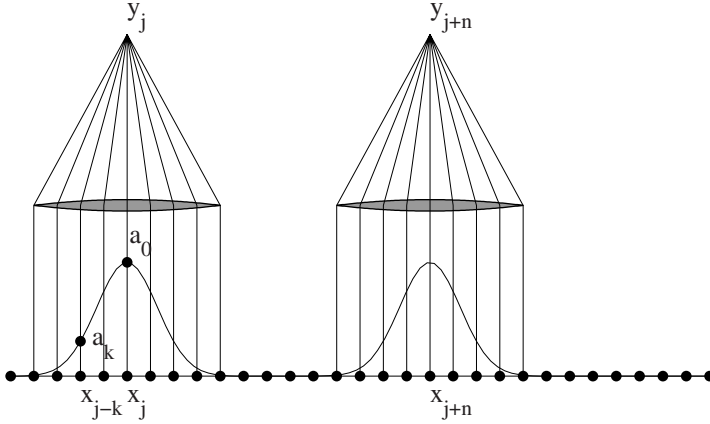


Fig. 3.2. The expected contribution from the different sources.

where $A \in \mathbb{R}^{N \times N}$ is the Toeplitz matrix

$$A = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-L} & & \\ & a_1 & a_0 & & \ddots & \\ \vdots & & \ddots & & & a_{-L} \\ a_L & & & \ddots & & \vdots \\ & \ddots & & & a_0 & a_{-1} \\ & & a_L & \cdots & a_1 & a_0 \end{bmatrix}.$$

The parameter L defines the *bandwidth* of the matrix.

If the sources are weak, the observation model just described is a photon counting process. We may think that each Y_j is a Poisson process with mean \bar{y}_j ,

$$Y_j \sim \text{Poisson}((Ax)_j),$$

that is,

$$\pi(y_j | x) = \frac{(Ax)_j^{y_j}}{y_j!} \exp(-(Ax)_j).$$

Observe that, in general, there is no guarantee that the expectations $(Ax)_j$ are integers. If we assume that consecutive measurements are independent, the random variable $Y \in \mathbb{R}^N$ has probability density

$$\pi(y | x) = \prod_{j=1}^N \pi(y_j | x) = \prod_{j=1}^N \frac{(Ax)_j^{y_j}}{y_j!} \exp(-(Ax)_j).$$

We express this relation simply by writing

$$Y \sim \text{Poisson}(Ax).$$

A model sometimes used in the literature assumes that the photon count is relatively large. In that case, using the Gaussian approximation of the Poisson density discussed in Chapter 1, the likelihood model becomes

$$\begin{aligned} \pi(y \mid x) &\approx \prod_{j=1}^N \left(\frac{1}{2\pi(Ax)_j} \right)^{1/2} \exp \left(-\frac{1}{2(Ax)_j} (y_j - (Ax)_j)^2 \right) \\ &= \left(\frac{1}{(2\pi)^L \det(\Gamma(x))} \right)^{1/2} \exp \left(-\frac{1}{2} (y - Ax)^T \Gamma(x)^{-1} (y - Ax) \right), \end{aligned} \quad (3.3)$$

where

$$\Gamma(x) = \text{diag}(Ax).$$

We could try to find an approximation of the Maximum Likelihood estimate of x by maximizing (3.3). However this maximization is nontrivial, since the matrix Γ appearing in the exponent and in the determinant is itself a function of x . In addition, a typical problem arising with convolution kernels is that the sensitivity of the problem to perturbations in the data is so high that *even if* we were able to compute the maximizer, it may be meaningless. Attempts to solve the problem in a stable way by means of numerical analysis have led to classical regularization techniques. The statistical approach advocated in this book makes an assessment of what we believe of a reasonable solution and incorporates this belief in the form of probability densities.

It is instructive to see what the Poisson noise looks like, and to see how Poisson noise differs from Gaussian noise with constant variance. In Figure 3.3, we have plotted a piecewise linear average signal \bar{x} and calculated a realization of a Poisson process with \bar{x} as its mean. It is evident that the higher the mean, the higher the variance, in agreement with the fact that the mean and the variance are equal. By visual inspection, Poisson noise could be confused with *multiplicative noise*, which will be discussed in the next example.

Before the example, let us point out a fact that is useful when composing probability densities. Assume that we have two random variables X and Y in \mathbb{R}^n that are related via a formula

$$Y = f(X),$$

where f is a differentiable function, and that the probability distribution of Y is known. We write

$$\pi(y) = p(y),$$

to emphasize that the density of Y is given as a particular function p . It would be tempting to deduce that the density of X , denoted by $\pi(x)$, is obtained by substituting $y = f(x)$ in the function p . This, in general, is

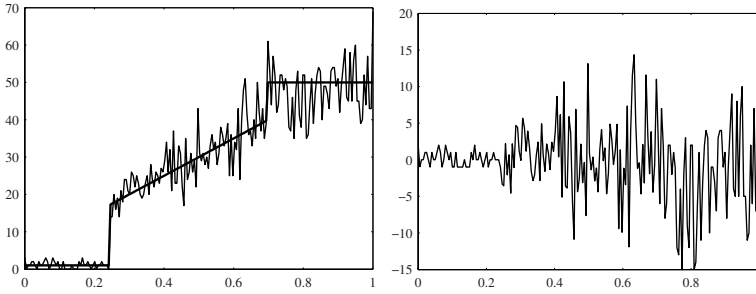


Fig. 3.3. Left panel: the average piecewise linear signal and a realization of a Poisson process, assuming that the values at each discretization point are mutually independent. Right panel: the difference between the noisy signal and the average.

not true, since probability densities represent *measures* rather than functions. The proper way to proceed is therefore to take the Jacobian of the coordinate transformation into consideration, by writing

$$\pi(y)dy = p(y)dy = p(f(x))|\det(Df(x))|dx,$$

where $Df \in \mathbb{R}^{n \times n}$ is the differential of f . Now we may identify the density of X as being

$$\pi(x) = p(f(x))|\det(Df(x))|. \quad (3.4)$$

In the following example, we make use of this formula.

EXAMPLE 3.4: Consider a noisy amplifier that takes in a signal $f(t)$ and sends it out amplified by a constant factor $\alpha > 1$. The ideal model for the output signal is

$$g(t) = \alpha f(t), \quad 0 \leq t \leq T.$$

In practice, however, it may happen that the amplification factor is not constant but fluctuates slightly around a mean value α_0 . We write a discrete likelihood model for the output by first discretizing the signal. Let

$$x_j = f(t_j), \quad y_j = g(t_j), \quad 0 = t_1 < t_2 < \cdots < t_n = T.$$

Assuming that the amplification at $t = t_j$ is a_j , we have a discrete model

$$y_j = a_j x_j, \quad 1 \leq j \leq n,$$

and, replacing the unknown quantities by random variables, we obtain the stochastic extension

$$Y_j = A_j X_j, \quad 1 \leq j \leq n,$$

which we write in vector notation as

$$Y = A.X, \quad (3.5)$$

with the dot denoting componentwise multiplication of the vectors $A, X \in \mathbb{R}^n$. Assume that A as a random variable is independent of X , as is the case, for instance, if the random fluctuations in the amplification are due to thermal phenomena. If A has the probability density

$$A \sim \pi_{\text{noise}}(a),$$

to find the probability density of Y , conditioned on $X = x$, we fix X and write

$$A_j = \frac{Y_j}{x_j}, \quad 1 \leq j \leq n.$$

Applying formula (3.4), we obtain

$$\pi(y | x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\text{noise}}\left(\frac{y \cdot}{x}\right), \quad (3.6)$$

where the dot denotes that the division of the two vectors is componentwise.

Let us consider a special example where we assume that all the variables are positive, and A is *log-normally distributed*, i.e., the logarithm of A is normally distributed. For simplicity, we assume that the components of A are mutually independent, identically distributed, hence

$$W_i = \log A_i \sim \mathcal{N}(w_0, \sigma^2), \quad w_0 = \log \alpha_0.$$

To find an explicit formula for the density of A , we note that if $w = \log a$, where the logarithm is applied componentwise, we have

$$dw = \frac{1}{a_1 a_2 \cdots a_n} da,$$

thus the probability density of A is

$$\begin{aligned} \pi_{\text{noise}}(a) &\propto \frac{1}{a_1 a_2 \cdots a_n} \exp\left(-\frac{1}{2\sigma^2} \|\log a - w_0\|^2\right) \\ &= \frac{1}{a_1 a_2 \cdots a_n} \exp\left(-\frac{1}{2\sigma^2} \left\|\log\left(\frac{a \cdot}{\alpha_0}\right)\right\|^2\right). \end{aligned}$$

By substituting this formula in (3.6), we find that

$$\pi(y | x) \propto \frac{1}{y_1 y_2 \cdots y_n} \exp\left(-\frac{1}{2\sigma^2} \left\|\log\left(\frac{y \cdot}{(\alpha_0 \cdot x)}\right)\right\|^2\right).$$

Before moving onto the design of priors, let's look at an example with two different sources of noise.

EXAMPLE 3.5: In this example, we assume that the photon counting convolution device of Example 3.3 adds a noise component to the collected data. More precisely, we have an observation model of the form

$$Y = Z + E,$$

where $Z \sim \text{Poisson}(Ax)$ and $E \sim \mathcal{N}(0, \sigma^2 I)$.

To write the likelihood model, we begin with assuming that $X = x$ and $Z_j = z_j$ are known. Then

$$\pi(y_j | z_j, x) \propto \exp\left(-\frac{1}{2\sigma^2}(y_j - z_j)^2\right).$$

Observe that x does not appear explicitly here, but is, in fact, a hidden parameter that affects the distribution of Z_j . From the formula for the conditional probability density it follows that

$$\pi(y_j, z_j | x) = \pi(y_j | z_j, x)\pi(z_j | x), \quad \pi(z_j | x) = \frac{(Ax)_j^{z_j}}{z_j!} \exp(-(Ax)_j).$$

Since the value of z_j is not of interest here, we can marginalize it out. In view of the fact that z_j takes on only integer values greater than or equal to zero, we write:

$$\begin{aligned} \pi(y_j | x) &= \sum_{z_j=0}^{\infty} \pi(y_j, z_j | x) \\ &\propto \sum_{z_j=0}^{\infty} \pi(z_j | x) \exp\left(-\frac{1}{2\sigma^2}(y - z_j)^2\right), \end{aligned}$$

which gives us the likelihood as a function of the variable of interest x . It is possible to replace the summation by using the Gaussian approximation for the Poisson distribution, but since the calculations are quite tedious and the final form is not so informative, we do not pursue that here.

3.2 Enter, Subject: Construction of Priors

The prior density expresses what we *know*⁴, or more generally, *believe* about the unknown variable of interest prior to taking the measurements into account. The choice of a prior accounts for the subjective portion of the

⁴ The use of the word *knowing* could be replaced by *being certain of*, conforming with the notion of Wittgenstein on certainty: being certain of something does not imply that things are necessarily that way, which is the tragedy of science and the salvation of the art.

procedure. In fact, what we believe a priori about the parameters biases the search so as to favor solutions which adhere to our expectation. How significantly the prior is guiding our estimation of the unknowns depends in part on the information contents of the measurements: in the absence of good measured data, the prior provides a significant part of the missing information and therefore has a great influence on the estimate. If, on the other hand, the measured data is highly informative, it will play a big role in the estimation of the unknowns, leaving only a lesser role for the prior, unless we want to go deliberately against the observations. We actually use priors more extensively than we are generally ready to admit, and in variety of situations: often, when we claim not to know what to expect, our prior is so strong to rule out right away some of the possible outcomes as meaningless, and, in fact, the concept of meaningless only exists as a counterpart to meaningful, and its meaning is usually provided by a prior.

EXAMPLE 3.6: Here we give a trivial example of hidden priors: assume that you want somebody to pick “a number, any number”. If that somebody is a normal⁵ human being, the answer may be 3, 7, 13 – very popular choices in Judeo-Christian tradition – or 42 – a favorite number in the science fiction subculture⁶. You may think that you did not have any prior on what to expect, until you ask a space alien in disguise who, taking your question very seriously, starts to recite 40269167330954837..., continuing this litany of digits for over a week. Although eventually the stream of digits will end, as the number is finite after all, you start to wonder if a lifetime will be enough to finish a round of the game. Quite soon, probably much before the end of the week, you realize that this is not at all what you had in mind with “picking a number”: the response is definitely against your prior. And in fact, you discover that not only your prior was there, but it was quite strong too: there are infinitely more numbers that would take more than a week – or a lifetime, for that matter – to say, than those you expected as an answer⁷!

Another, more pertinent example comes from the field of inverse problems.

⁵ Preferably not from a math department, where people enjoy to come up with the most intricate answers for the simplest of questions, a tendency which becomes exacerbated when the answer involves numbers.

⁶ In the subculture of mathematicians, even more exotic answers are likely, such as the Hardy–Ramanujan number, known also as the taxicab number, 1729. The point here is to notice that to obtain numbers no more complex than this, it takes some of the best number-minded brains. The human bias towards small numbers is enormous.

⁷ Jose Luis Borges has an elegant short story, *The book of sand*, about a book with an endless number of pages. Opening the book and reading the page number would be tantamount to choosing “any” number. Borges, possibly aware of the problem, leaves the page numbering a mystery.

EXAMPLE 3.7: Your orthopedic surgeon has asked you to take along the MRI slides of your knee in the next visit, but, on your way out of the house, you accidentally took the envelope with the MRI slides of the brain. As soon as the slides are taken out of the envelope, the doctor knows that you have grabbed the wrong ones, in spite of having just told you not to have any idea what to expect to see in the slides.

The prior beliefs that we often have are *qualitative*, and it may be quite a challenge to translate them into *quantitative* terms. How do we express in a prior that we expect to see an MRI of the neck, or how do we describe quantitatively what radiologists mean by “habitus of a malignant tumor”? Clearly, we are now walking along the divide between art and science.

The obvious way to assign a prior is by stating that the unknown parameter of interest follows a certain distribution. Since the prior expresses our subjective belief about the unknown, our belief is a sufficient justification for its being. Note that the reluctance to choosing any specific prior, usually for fear of its subjective nature, *de facto* leads often to a claim that all values are equally likely, which, apart of its mathematical shortcomings, leaves it entirely up to the likelihood to determine the unknown. The results may be catastrophic.

The priors that we apply in day to day life are often inspired by our previous experience in similar situations. To see how we can proceed in the construction of a prior, we consider some examples.

EXAMPLE 3.7: Assume that we want to determine the level x of hemoglobin in blood by near-infrared (NIR) measurements at a patient’s finger. If we have a collection of hemoglobin values measured directly from the patient’s blood,

$$S = \{x_1, \dots, x_N\},$$

we can think of them as *realizations* of a random variable X with an unknown distribution. As explained in Chapter 2, there are two possible ways for extracting information about the underlying distribution from S :

- The *non-parametric* approach looks at a histogram based on S and tries to infer what the underlying distribution is.
- The *parametric* approach proposes a parametric model, then computes the Maximum Likelihood estimate of the model parameters from the sample S .

Let us assume, for example, that a Gaussian model is proposed,

$$X \sim \mathcal{N}(x_0, \sigma^2).$$

We know from Chapter 2 that the Maximum Likelihood estimates for x_0 and σ^2 are,

$$x_{0,\text{ML}} = \frac{1}{N} \sum_{j=1}^N x_j,$$

and

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{j=1}^N (x_j - x_{0,\text{ML}})^2,$$

respectively. We then assume that any future value x of the random variable is a realization from this Gaussian distribution. Thus, we postulate that:

- The unknown X is a random variable, whose probability distribution, called the *prior distribution*, is denoted by $\pi_{\text{prior}}(x)$,
- Guided by our prior experience, and assuming that a Gaussian prior is justifiable, we use the parametric model

$$\pi_{\text{prior}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right),$$

with x_0 and σ^2 determined experimentally from S by the formulas above.

Methods in which prior parameters are estimated empirically, either from previous observations or simultaneously with the unknown from the current data, are called *empirical Bayes methods*.

The prior may be partly based on a physical, chemical or biological model, as in the following example.

EXAMPLE 3.8: Consider a petri dish with a culture of bacteria whose density we want to estimate. For simplicity, let's assume that we have a rectangular array of squares, where each square contains a certain number of bacteria, as illustrated in Figure 3.4, and that we are interested in estimating the density of the bacteria from some indirect measurements⁸. We begin with setting up a model based on our belief about bacterial growth. For example, we may assume that the number of bacteria in a square is approximately the average of bacteria in the neighboring squares,

$$x_j \approx \frac{1}{4}(x_{\text{left},j} + x_{\text{right},j} + x_{\text{up},j} + x_{\text{down},j}), \quad (3.7)$$

see Figure 3.4 for an explanation. Since the squares at the boundary of the petri dish have no neighbors in some directions, we need to modify (3.7) to account for their special status. The way in which we will handle this turns out to be related to the choice of boundary conditions for partial differential equations. For example, we can assume that $x_j = 0$ in pixels outside the square.

Let N be the number of pixels and $A \in \mathbb{R}^{N \times N}$ be the matrix with j th row

⁸ Aside from the fact that direct count of bacteria would be impossible, it would amount to moments of unmatched enjoyment.

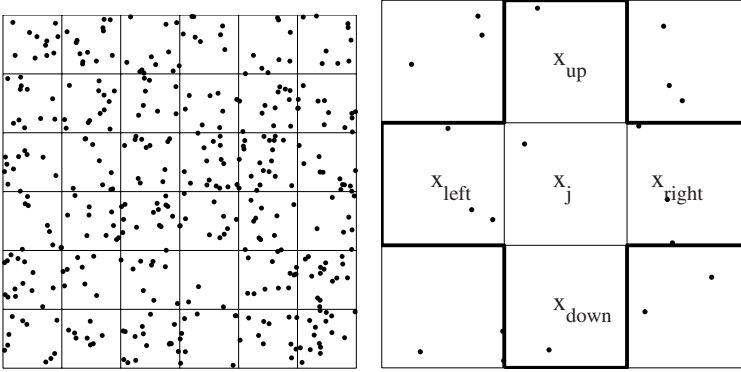


Fig. 3.4. Square array of bacteria. On the right we zoom in on a neighborhood system.

$$A(j, :) = \begin{bmatrix} 0 & \dots & 1/4 & \dots & 1/4 & \dots & 1/4 & \dots & 1/4 & \dots & 0 \end{bmatrix},$$

with the understanding that for boundary and corner pixels, some of the columns may be missing, corresponding to the assumption that no contribution from the outside of the array is coming.

If we were absolutely certain of the validity of the model (3.7), we would replace the approximate sign with strict equality. In other words we would assume that

$$x = Ax. \quad (3.8)$$

which clearly does not work! In fact, after rewriting (3.8) as

$$(I - A)x = 0,$$

it immediately follows that $x = 0$, since $I - A$ is an invertible matrix. Therefore, to relax the model (3.8) by admitting some uncertainty, we define X to be a random variable and write a stochastic model

$$X = AX + R, \quad (3.9)$$

where R expresses the uncertainty of the averaging model. The lack of information about the values of R is encoded in its probability density,

$$R \sim \pi_{\text{mod.error}}(r),$$

and, by writing $R = X - AX$, we define that the prior density for X is

$$\pi_{\text{prior}}(x) \propto \pi_{\text{mod.error}}(x - Ax).$$

Equation (3.9) defines an *autoregressive Markov model*, and in this context R is referred to as an *innovation process*. In particular, if R is Gaussian with mutually independent, identically distributed components,

$$R \sim \mathcal{N}(0, \sigma^2 I),$$

the prior for X is of the form

$$\begin{aligned}\pi_{\text{prior}}(x \mid \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\|x - Ax\|^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}\|Lx\|^2\right),\end{aligned}$$

where

$$L = I - A.$$

We remark that if we have no natural way of fixing the value of σ^2 , estimating it on the basis of observations is part of the larger estimation problem. One possible line of thought for choosing the variance σ^2 could be as follows: suppose that you do expect to find about M bacteria in the whole dish. That means that, in the average, one square contains about $m = M/N$ bacteria. The fluctuation around the postulated average value in a single square is probably not much more than m , so a reasonable choice could be $\sigma \approx m$.

We conclude this example by showing a Matlab code to construct the matrix A . Since this matrix is sparse, it is advisable to construct it as a sparse matrix. Sparse matrices are defined in Matlab by three vectors of equal length. The first one contains the indices of the rows with non-vanishing entries, the second one the column indices and the third one the actual non-vanishing values. These vectors are called `rows`, `cols` and `vals` in the code below.

```
n = 50; % Number of pixels per directions

% Creating an index matrix to enumerate the pixels

I = reshape([1:n^2],n,n);

% Right neighbors

Icurr = I(:,1:n-1);
Ineigh = I(:,2:n);
rows = Icurr(:);
cols = Ineigh(:);
vals = ones(n*(n-1),1);

% Left neighbors

Icurr = I(:,2:n);
Ineigh = I(:,1:n-1);
```

```

rows = [rows;Icurr(:)];
cols = [cols;Ineigh(:)];
vals = [vals;ones(n*(n-1),1)];

% Upper neighbors

Icurr = I(2:n-1,:);
Ineigh = I(1:n-1,:);
rows = [rows;Icurr(:)];
cols = [cols;Ineigh(:)];
vals = [vals;ones(n*(n-1),1)];

% Lower neighbors

Icurr = I(1:n-1,:);
Ineigh = I(2:n,:);
rows = [rows;Icurr(:)];
cols = [cols;Ineigh(:)];
vals = [vals;ones(n*(n-1),1)];

A = 1/4*sparse(rows,cols,vals);
L = speye(n^2) - A;

```

Observe that L is, in fact, a second order finite difference matrix built from the mask

$$\begin{bmatrix} & -1/4 & \\ -1/4 & 1 & -1/4 \\ & -1/4 & \end{bmatrix},$$

hence L is, up to a scaling constant, a discrete approximation of the Laplacian

$$-\Delta = -\frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2}$$

in the unit square. Indeed, if

$$x_j = f(p_j),$$

where p_j is a point in the j th pixel, the finite difference approximation of the Laplacian of f can be written in the form

$$-\Delta f(p_j) \approx \frac{4}{h^2} (Lx)_j,$$

where h is the length of the discretization step. At the boundaries, we assume that f extends smoothly by zero outside the square.

The prior model derived in the last example corresponds to what is often referred to as *second order smoothness prior*. The reason for the name is that

this prior assigns a higher probability to vectors $x \in \mathbb{R}^N$ corresponding to discrete approximations of functions with a small second derivative, since the exponential is larger when the negative exponent is smaller, that is, when Lx has a small norm. This is the case, in general, for vectors x which are discretizations of smooth functions. We remark that a Gaussian smoothness prior does not exclude the occurrence of large jumps between adjacent pixels, but it gives them an extremely low probability.

Before leaving, temporarily, the theme of prior distributions, we want to report on a typical discussion which could occur after a presentation of Bayesian solutions of inverse problems on how to estimate a gray scale image from a blurred and noisy copy. In the image processing literature, this is referred to as a denoising and deblurring problem. The image can be represented as a pixel matrix, each entry assigning a gray scale value to the corresponding pixel. Assume that, after stacking the pixel values in one long vector, a Gaussian prior has been constructed for the image vectors. A typical question is: “Can you really assume that the prior density is Gaussian⁹?” While at first sight such a question may seem reasonable, once properly analyzed, it becomes rather obscure. In fact, the question is based on the Platonic view that *there is a true prior*. But, true prior of what, we may ask. All possible images, maybe? Most certainly not, and in fact *all possible images* is a useless category, as is the category of all possible realizations of all random process, or all possible worlds. Maybe the set should be restricted to all possible images that can be realized in the particular application that we have in mind? But such category is also useless, unless we specify its genesis by describing, for instance, the probability distribution of the images. And at this point we face the classic problem of which came first, the hen or the egg. A more reasonable question would therefore be, what the given prior implies. As we shall see, a possible way of exploring the implications of a prior distribution is to use sampling techniques.

3.3 Posterior Densities as Solutions of Statistical Inverse Problems

Within the previous sections of this chapter we introduced the main actors in our Bayesian play, the prior and the likelihood. From now on, the marginal density of the unknown of primary interest will be identified as prior density, and denoted by π_{prior} . Notice that both the likelihood and the prior may contain parameters whose values we are not very confident about; the natural Bayesian solution is to regard them as random variables, too.

If we let X denote the random variable to be estimated and Y the random variable that we observe, from the *Bayes formula* we have that

⁹ Note that when saying that the prior is Gaussian, we do not intend a prior with independent equally distributed components, so the obvious arguments referring to non-negativity of the image do not necessarily apply here.

$$\pi(x | y) = \frac{\pi_{\text{prior}}(x)\pi(y | x)}{\pi(y)}, \quad y = y_{\text{observed}}. \quad (3.10)$$

The conditional density $\pi(x | y)$, called the *posterior density*, expresses the probability of the unknown given that the observed parameters take on the values given as the data of the problem and our prior belief. In a Bayesian statistical framework, *the posterior density is the solution of the inverse problem*.

Two questions are now pertinent. The first one is how to find the posterior density and the second how to extract information from it in a form suitable for our application. The answer to the first question has already been given: from the prior and the likelihood, the posterior can be assembled via the Bayes formula. The answer to the second question will be given in the remaining chapters. As a prelude to the ensuing discussion, let us consider the following example.

EXAMPLE 3.9: Consider a linear system of equations with noisy data,

$$y = Ax + e, \quad x \in \mathbb{R}^n, \quad y, e \in \mathbb{R}^m, \quad A \in \mathbb{R}^{m \times n},$$

and let

$$Y = AX + E$$

be its stochastic extension, where X , Y and E are random variables. A very common assumption is that X and E are independent and Gaussian,

$$X \sim \mathcal{N}(0, \gamma^2 \Gamma), \quad E \sim \mathcal{N}(0, \sigma^2 I),$$

where we have assumed that both random variables X and E have zero mean. If this is not the case, the means can be subtracted from the random variable, and we arrive at the current case. The covariance of the noise indicates that each component of Y is contaminated by independent identically distributed random noise. We assume that the matrix Γ in the prior is known. The role of the scaling factor γ will be discussed later. The prior density is therefore of the form

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2\gamma^2}x^T \Gamma^{-1}x\right),$$

The likelihood density, assuming that the noise level σ^2 is known, is

$$\pi(y | x) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2\right).$$

It follows from the Bayes formula that the posterior density is

$$\begin{aligned} \pi(x | y) &\propto \pi_{\text{prior}}(x)\pi(y | x) \\ &\propto \exp\left(-\frac{1}{2\gamma^2}x^T \Gamma^{-1}x - \frac{1}{2\sigma^2}\|y - Ax\|^2\right) \\ &= \exp(-V(x | y)), \end{aligned}$$

where

$$V(x | y) = \frac{1}{\gamma^2} x^T \Gamma^{-1} x + \frac{1}{2\sigma^2} \|y - Ax\|^2.$$

Since the matrix Γ is symmetric positive definite, so is its inverse, thus admitting a symmetric, e.g., Cholesky, factorization:

$$\Gamma^{-1} = R^T R.$$

With this notation,

$$x^T \Gamma^{-1} x = x^T R^T R x = \|R x\|^2,$$

and we define

$$T(x) = 2\sigma^2 V(x | y) = \|y - Ax\|^2 + \delta^2 \|R x\|^2, \quad \delta = \frac{\sigma}{\gamma}. \quad (3.11)$$

This functional T , sometimes referred to as the *Tikhonov functional*, plays a central role in the classical regularization theory.

The analogue of the Maximum Likelihood estimator in the Bayesian setting maximizes the posterior probability of the unknowns and is referred to as the *Maximum A Posteriori (MAP) estimator*:

$$x_{\text{MAP}} = \arg \max \pi(x | y).$$

Note that

$$x_{\text{MAP}} = \arg \min V(x | y), \quad V(x | y) = -\log \pi(x | y).$$

In this particular case we have that

$$x_{\text{MAP}} = \arg \min (\|y - Ax\|^2 + \delta^2 \|R x\|^2). \quad (3.12)$$

When the posterior density is Gaussian, the Maximum A Posteriori estimate coincides with the *Conditional Mean (CM)*, or *Posterior Mean* estimate,

$$x_{\text{CM}} = \int x \pi(x | y) dy.$$

It is immediate to check that if we have $R = I$ and let γ increase without bounds, the parameter δ goes to zero and the Maximum A Posteriori estimator reduces formally to the Maximum Likelihood estimator, namely the estimation of x is entirely based on the likelihood density.

In our discussion of the Maximum Likelihood estimation in Chapter 2, Example 2.4, we saw that its calculation can be reduced to solving a system of linear equations

$$Ax = y, \quad (3.13)$$

possibly in the least squares sense. It was also pointed out that even when the matrix A is square, if it is numerically singular the calculation of the Maximum Likelihood estimate becomes problematic. Classical regularization techniques are based on the idea that an ill-posed problem is replaced by a nearby problem that is well-posed. *Tikhonov regularization*, for example, replaces the linear system (4.1) with the minimization problem

$$\min \left\{ \|y - Ax\|^2 + \delta^2 \|Rx\|^2 \right\}, \quad (3.14)$$

where the first term controls the fidelity of the solution to the data and the second one acts as a *penalty*. The matrix R is typically selected so that large $\|Rx\|$ corresponds to an undesirable feature of the solution. The role of the penalty is to keep this feature from growing unnecessarily. A central problem in Tikhonov regularization is how to choose the value of the parameter δ judiciously. A Tikhonov regularized solution is the result of a compromise between fitting the data and eliminating unwanted features. As we have seen, this fits well into the Bayesian framework, since the penalty can be given a statistical interpretation via the prior density.

A question which comes up naturally at this point is if and how Tikhonov regularization, or MAP estimation, can avoid the numerical problems that make Maximum Likelihood estimation generally infeasible for noisy data. The answer can be found by writing the functional to be minimized in (3.14) in the form

$$\|y - Ax\|^2 + \delta^2 \|Rx\|^2 = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \delta R \end{bmatrix} x \right\|^2,$$

which reveals that the Maximum A Posteriori estimate is the least squares solution of the linear system

$$\begin{bmatrix} A \\ \delta R \end{bmatrix} x = \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

While the original problem (3.13) may be ill-posed, the augmentation of the matrix A by δR considerably reduces the ill-posedness of the problem. The quality of the solution depends on the properties of the matrix R .

Exercises

1. A patient with a lump in the breast undergoes a mammography. The radiologist who examines her mammogram believes that the lump is malignant, that is, the result of the mammogram is positive for malignancy. The radiologist's record of true and false positives is shown in the table below:

	Malignant	Benign
Positive mammogram	0.8	0.1
Negative mammogram	0.2	0.9

The patient, with the help of the internet, finds out that in the whole population of females, the probability of having a malignant breast tumor at her age is 0.5 percent. Without thinking much further, she takes this number as her prior belief of having cancer.

- (a) What is the probability of the mammogram result being positive for malignancy?
 - (b) What is the conditional probability of her having a malignant tumor, considering the fact that the mammogram's result was positive?
 - (c) What problems are there with her selection of the prior, others than the possible unreliability of internet?
2. A person claims to be able to guess a number from 1 to 10 which you are asked to think of. He has a record of having succeeded 8 times out of 10. You question the claim and in your mind assign a certain probability x of him having indeed such gift, but decide to give the poor devil a chance and let him try to guess the number between 1 and 10 that you are thinking. He guesses correctly, but you are still not convinced. In other words, even after his successful demonstration, you still think that he is a swindler and that the probability of such an extraordinary gift is less than 0.5. How low must your prior belief x have been for this to happen?
 3. Rederive the result of Example 3.4 in a slightly different way: By taking the logarithm of both sides of equation (3.5), we obtain

$$\log Y = \log X + \log A = \log X + W,$$

where W is normally distributed. Write the likelihood density for $\log Y$ conditioned on $X = x$ and derive the likelihood for Y .

An Introduction to Bayesian Scientific Computing

Ten Lectures on Subjective Computing

Calvetti, D.; Somersalo, E.

2007, XIV, 202 p., Softcover

ISBN: 978-0-387-73393-7