Statistical Methods in Imaging

Daniela Calvetti and Erkki Somersalo

Contents

D. Calvetti (✉)
Department of Mathematics and Department of Cognitive Science, Case Western Reserve University, Cleveland, OH, USA
e-mail: daniela.calvetti@case.edu

E. Somersalo
Department of Mathematics, Case Western Reserve University, Cleveland, OH, USA
e-mail: erkki.somersalo@case.edu

O. Scherzer (ed.), *Handbook of Mathematical Methods in Imaging*,
DOI 10.1007/978-1-4939-0790-8_53

Abstract

The theme of this chapter is statistical methods in imaging, with a marked emphasis on the Bayesian perspective. The application of statistical notions and techniques in imaging requires that images and the available data are redefined in terms of random variables, the genesis and interpretation of randomness playing a major role in deciding whether the approach will be along frequentist or Bayesian guidelines. The discussion on image formation from indirect information, which may come from non-imaging modalities, is coupled with an overview of how statistics can be used to overcome the hurdles posed by the inherent ill-posedness of the problem. The statistical counterpart to classical inverse problems and regularization approaches to contain the potentially disastrous effects of ill-posedness is the extraction and implementation of complementary information in imaging algorithms. The difficulty in expressing quantitative and uncertain notions about the imaging problem at hand in qualitative terms, which is a major challenge in a deterministic context, can be more easily overcome once the problem is expressed in probabilistic terms. An outline of how to translate some typical qualitative traits into a format which can be utilized by statistical imaging algorithms is presented. In line with the Bayesian paradigm favored in this chapter, basic principles for the construction of priors and likelihoods are presented, together with a discussion of numerous computational statistics algorithms, including maximum likelihood estimators, maximum a posteriori and conditional mean estimators, expectation maximization, Markov chain Monte Carlo, and hierarchical Bayesian models. Rather than aiming to be a comprehensive survey, the present chapter hopes to convey a wide and opinionated overview of statistical methods in imaging.

1 Introduction

Images, alone or in sequences, provide a very immediate and effective way of transferring information, as the human eye–brain complex is extremely well adapted at extracting quickly their salient features, let them be edges, textures, anomalies, or movement. While the amount of information that can be compressed in an image is tremendously large and varied, the image processing ability of the human eye is so advanced to outperform the most advanced of algorithms. One of the reasons why the popularity of statistical tools in imaging continues to grow is the flexibility that this modality offers when it comes to utilizing qualitative attributes of the images or to recover them from indirect, corrupt specimens. The utilization of qualitative clues to augment scarce data is akin to the process followed by the eye–brain system.

Statistics, which according to Pierre–Simon Laplace, is "common sense expressed in terms of numbers," is well suited for quantifying qualitative attributes. The opportunity to augment poor quality data with complementary information which may be based on our preconception of what we are looking for or on

information coming from sources other than the data makes statistical methods particularly attractive in imaging applications.

In this chapter, we present a brief overview of some of the key concepts and most popular algorithms in statistical imaging, highlighting the similarity and the differences with the closest deterministic counterparts. A particular effort is made to demonstrate that the statistical methods lead to new ideas and algorithms that the deterministic methods do not give.

2 Background

Images in the Statistical Setting

The mathematical vessel that we will use here to describe a black and white image is a matrix with nonnegative entries, each representing the light intensity at one pixel of the discretized image. Color images can be thought of as the result of superimposing a few color intensity matrices; in most application, a color image is represented by three matrices, for example, encoding the red, green, and blue intensity at each pixel. While color imaging applications can also be approached with statistical methods, here we will only consider gray-scale images. Thus, an image X is represented as a matrix

$$\mathsf{X} = \left[x_{ij}\right], \quad 1 \le i \le n,\ 1 \le j \le m,\ x_{ij} \ge 0.$$

In our treatment, we will not worry about the range of the image pixel values, assuming that, if necessary, the values are appropriately normalized. Notice that this representation tacitly assumes that we restrict our discussion to rectangular images discretized into rectangular arrays of pixels. This hypothesis is neither necessary nor fully justified, but it simplifies the notation in the remainder of the chapter. In most imaging algorithms, the first step consists of storing the image into a vector by reshaping the rectangular matrix. We use here a columnwise stacking, writing

$$\mathsf{X} = \left[x^{(1)}\ x^{(2)}\ \dots\ x^{(m)}\right], \quad x^{(j)} \in \mathbb{R}^n,\ 1 \le j \le m,$$

and further

$$x = \mathrm{vec}\,(\mathsf{X}) = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(m)} \end{bmatrix} \in \mathbb{R}^N, \quad N = n \times m.$$

Images can be either directly observed or represent a function of interest, as is, for example, the case for tomographic images.

Randomness, Distributions, and Lack of Information

We start this section by introducing some notations. A multivariate random variable $X : \Omega \to \mathbb{R}^N$ is a measurable mapping from a probability space Ω equipped with a σ-algebra and a probability measure P. The elements of $\mathbb{R}^N$, as well as the realizations of X, are denoted by lowercase letters, that is, for $\omega \in \Omega$ given, $X(\omega) = x \in \mathbb{R}^N$. The probability distribution μ_X is the measure defined as

$$\mu_X(B) = \mathsf{P}\left(X^{-1}(B)\right), \quad B \subset \mathbb{R}^N \text{ measurable.}$$

If μ_X is absolutely continuous with respect to the Lebesgue measure, there is a measurable function π_X, the Radon–Nikodym derivative of μ_X with respect to the Lebesgue measure such that

$$\mu_X(B) = \int_B \pi_X(x)dx.$$

For the sake of simplicity, we shall assume that all the random variables define probability distributions which are absolutely continuous with respect to the Lebesgue measure.

Consider two random variables $X : \Omega \to \mathbb{R}^N$ and $Y : \Omega \to \mathbb{R}^M$. The joint probability density is defined first over Cartesian products,

$$\mu_{X,Y}(B \times D) = \mathsf{P}\left(X^{-1}(B) \cap Y^{-1}(D)\right),$$

and then extended to the whole product σ-algebra over $\mathbb{R}^N \times \mathbb{R}^M$. Under the assumption of absolute continuity, the joint density can be written as

$$\mu_{X,Y}(B \times D) = \int_B \int_D \pi_{X,Y}(x, y)dydx,$$

where $\pi_{X,Y}$ is a measurable function. This definition extends naturally to the case of more than two random variables.

Since the notation just introduced here gets quickly rather cumbersome, we will simplify it by dropping the subscripts, writing $\pi_{X,Y}(x, y) = \pi(x, y)$, that is, letting x and y be at the same time variables and indicators of their parent uppercase random variables. Furthermore, since the ordering of the random variables is irrelevant – indeed, $\mathsf{P}\left(X^{-1}(B) \cap Y^{-1}(D)\right) = \mathsf{P}\left(Y^{-1}(D) \cap X^{-1}(B)\right)$ – we will occasionally interchange the roles of x and y in the densities, without assuming that the probability densities should be symmetric in x and y. In other words, we will use π as a generic symbol for "probability density."

With these notations, given two random variables X and Y, define the marginal densities

$$\pi(x) = \int_{\mathbb{R}^M} \pi(x, y)dy, \quad \pi(y) = \int_{\mathbb{R}^N} \pi(x, y)dx,$$

which express the probability densities of X and Y, respectively, on their own, while the other variable is allowed to take on any value. By fixing y, and assuming that $\pi(y) \neq 0$, we have that

$$\int_{\mathbb{R}^N} \frac{\pi(x, y)}{\pi(y)} dx = 1;$$

hence, the nonnegative function

$$x \mapsto \pi(x \mid y) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(y)} \tag{1}$$

defines a probability distribution for X referred to as the conditional density of X, given $Y = y$. Similarly, we define the conditional density of Y given $X = x$ as

$$\pi(y \mid x) \stackrel{\text{def}}{=} \frac{\pi(x, y)}{\pi(x)}. \tag{2}$$

This rather expedite way of defining the conditional densities does not fully explain why this interpretation is legitimate; a more rigorous explanation can be found in textbooks on probability theory [8, 18].

The concept of probability measure does not require any further interpretation to yield a meaningful framework for analysis, and this indeed is the viewpoint of theoretical probability. When applied to real-world problems, however, an interpretation is necessary, and this is exactly where the opinions of statisticians start to diverge. In frequentist statistics, the probability of an event is its asymptotic relative frequency of occurrence as the number of repeated experiments tend to infinity, and the probability density can be thought of as a limit of histograms. A different interpretation is based on the concept of information. If the value of a quantity is either known or it is at potentially retrievable from the available information, there is no need to leave the deterministic realm. If, on the other hand, the value of a quantity is uncertain in the sense that the available information is insufficient to determine it, to view it as a random variable appears natural. In this interpretation of randomness, it is immaterial whether the lack of information is contingent ("imperfect measurement device, insufficient sampling of data") or fundamental ("quantum physical description of an observable"). It should also be noted that the information, and therefore the concept of probability, is subjective, as the value of a quantity may be known to one observer and unknown to another [14, 18]. Only in the latter case the concept of probability is needed. The interpretation of probability in this chapter follows mostly the subjective, or Bayesian tradition, although most of the time the distinction is immaterial. Connections to non-Bayesian statistics are made along the discussion.

Most imaging problems can be recast in the form of a statistical inference problem. Classically, inverse problems are stated as follows: *Given an observation of a vector* $y \in \mathbb{R}^M$*, find an estimate of the vector* $x \in \mathbb{R}^N$*, based on the forward*

model mapping x to y. Statistical inference, on the other hand, is concerned with identifying a probability distribution that the observed data is presumably drawn from. In the frequentist statistics, the observation y is seen as a realization of a random variable Y, the unknown x being a deterministic parameter that determines the underlying distribution $\pi(y \mid x)$, or *likelihood density*, and hence the estimation of x is the object of interest. In contrast, in the Bayesian setting, both variables x and y are first extended to random variables, Y and X, respectively, as discussed in more detail in the following sections. The marginal density $\pi(x)$, which is independent of the observation y, is called the *prior density* and denoted by $\pi_{\text{prior}}(x)$, while the likelihood is the conditional density $\pi(y \mid x)$. Combining the formulas (1) and (2), we obtain

$$\pi(x \mid y) = \frac{\pi_{\text{prior}}(x)\pi(y \mid x)}{\pi(y)},$$

which is the celebrated Bayes' formula [3]. The conditional distribution $\pi(x \mid y)$ is the *posterior distribution* and, in the Bayesian statistical framework, the solution of the inverse problem.

Imaging Problems

A substantial body of classical imaging literature is devoted to problems where the data consists of an image, represented here as a vector $y \in \mathbb{R}^M$ that is either a noisy, blurred, or otherwise corrupt version of the image $x \in \mathbb{R}^N$ of primary interest. The canonical model for this class of imaging problems is

$$y = \mathsf{A}x + \text{“noise,”} \tag{3}$$

where the properties of the matrix A depend on the imaging problem at hand. A more general imaging problems is of the form

$$y = F(x) + \text{“noise,”} \tag{4}$$

where the function $F : \mathbb{R}^N \mapsto \mathbb{R}^M$ may be a nonlinear function and the data y need not even represent an image. This is a common setup in medical imaging applications with a nonlinear forward model.

In classical, nonstatistical framework, imaging problems, and more generally, inverse problems, are often, somewhat arbitrarily, classified as being linear or nonlinear, depending on whether the forward model F in (4) is linear or nonlinear. In the statistical framework, this classification is rather irrelevant. Since probability densities depend not only on the forward map but also on the noise and, in the Bayesian case, the prior models, even a linear forward map can result in a nonlinear

estimation problem. We review some widely studied imaging problems to highlight this point.

1. *Denoising*: Denoising refers to the problem of removing noise from an image which is otherwise deemed to be a satisfactory representation of the information. The model for denoising can be identified with (3), with $M = N$ and the identity $\mathsf{A} = \mathsf{I} \in \mathbb{R}^{N\times N}$ as forward map.
2. *Deblurring*: Deblurring is the process of removing a blur, due, for example, to an imaging device being out of focus, to motion of the object during imaging ("motion blur"), or to optical disturbances in atmosphere during image formation. Since blurred images are often contaminated by exogenous noise, denoising is an integral part of the deblurring process. Given the image matrix $\mathsf{X} = [x_{ij}]$, the blurring is usually represented as

$$y_{ij} = \sum_{k,\ell} a_{ij,k\ell} x_{k\ell} + \text{"noise."}$$

   Often, but not without loss of generality, the blurring matrix can be assumed to be a convolution kernel,

$$a_{ij,k\ell} = a_{i-k,j-\ell},$$

   with the obvious abuse of notations. It is a straightforward matter to arrange the elements, so that the above problem takes on the familiar matrix–vector form $y = \mathsf{A}x$, and in the presence of noise, the model coincides with (3).
3. *Inpainting*: Here, it is assumed that part of the image x is missing due to an occlusion, a scratch, or other damages. The problem is to paint in the occlusion based on the visible part of the image. In this case, the matrix A in the linear model (3) is a sampling matrix, picking only those pixels of $x \in \mathbb{R}^N$ that are present in $y \in \mathbb{R}^M$, $M < N$.
4. *Image formation*: Image formation is the process of translating data into the form of an image. The process is common in medical imaging, and the description of the forward model connecting the sought image to data may involve linear or nonlinear transformations. An example of a linear model arises in tomography: The image is explored one line at the time, in the sense that the data consist of line integrals indirectly measuring the amount of radiation absorbed in the trajectory from source to detector or the number of photons emitted at locations along the trajectory between pairs of detectors. The problem is of the form (3). An example of a nonlinear imaging model (4) arises in near-infrared optical tomography, in which the object of interest is illuminated by near-infrared light sources, and the transmitted and scattered light intensity is measured in order to form an image of the interior optical properties of the body.

Some of these examples will be worked out in more details below.

3 Mathematical Modeling and Analysis

Prior Information, Noise Models, and Beyond

The goal in Bayesian statistical methods in imaging is to identify and explore probability distributions of images rather than looking for single images, while in the non-Bayesian framework, one seeks to infer on deterministic parameter vectors defining the distribution that the observations are drawn from. The main player in non-Bayesian statistics is the likelihood function, in the notation of section "Randomness, Distributions and Lack of Information," $\pi(y \mid x)$, where $y = y_{\text{observed}}$. In Bayesian statistics, the focus is on the posterior density $\pi(x \mid y)$, $y = y_{\text{observed}}$, the likelihood function being a part of it as indicated by Bayes' formula.

We start the discussion with the Bayesian concept of prior distribution, the non-Bayesian modeling paradigm being discussed in connection with the likelihood function.

Accumulation of Information and Priors

To the question, what should be in a prior for an imaging problem, the best answer is whatever can be built using available information about the image which can supplement the measured data. The information to be accounted by the prior can be gathered in many different ways. Any visually relevant characteristic of the sought image is suitable for a prior, including but not limited to texture, light intensity, and boundary structure. Although it is often emphasized that in a strict Bayesian framework the prior and the likelihood must be constructed separately, in several imaging problems, the setup may be impractical, and the prior and likelihood need to be set up simultaneously. This is the case, for example, when the noise is correlated with the signal itself. Furthermore, some algorithms may contain intermediate steps that formally amount to updating of the a priori belief, a procedure that may seem dubious in the traditional formal Bayesian setting but can be justified in the framework of hierarchical models. For example, in the restoration of images with sharp contrasts from severely blurred, noisy copies, an initially very vague location of the gray-scale discontinuities can be made more precise by extrapolation from intermediate restorations, leading to a Bayesian learning model.

It is important to understand that in imaging, the use of complementary information to improve the performance of the algorithms at hand is a very natural and widespread practice and often necessary to link the solution of the underlying mathematical problem to the actual imaging application. There are several constituents of an image that are routinely handled under the guidance of a priori belief even in fully deterministic settings. A classical example is the assignment of

boundary conditions for an image, a problem which has received a lot of attention over the span of a couple of decades (see, e.g., [21] and references therein). In fact, since it is certainly difficult to select the most appropriate boundary condition for a blurred image, ultimately the choice is based on a combination of a priori belief and algorithmic considerations. The implementation of boundary conditions in deterministic algorithms can therefore be interpreted as using a prior, expressing an absolute belief in the selected boundary behavior. The added flexibility which characterizes statistical imaging methodologies makes it possible to import in the algorithm the postulated behavior of the image at the boundary with a certain degree of uncertainty.

The distribution of gray levels within an image and the transition between areas with different gray-scale intensities are the most likely topics of a priori beliefs, hence primary targets for priors. In the nonstatistical imaging framework, a common choice of regularization, for the underlying least squares problems is a regularization functional, which penalizes growth in the norm of the derivative of the solution, thus discouraging solutions with highly oscillatory components. The corresponding statistical counterpart is a Markov model, based, for example, on the prior assumption that the gray-scale intensity at each pixel is a properly weighted average of the intensities of its neighbors plus a random innovation term which follows a certain statistical distribution. As an example, assuming a regular quadrilateral grid discretization, the typical local model can be expressed in terms of probability densities of pixel values X_j conditioned on the values of its neighboring pixels labeled according to their relative position to X_j as X_{up}, X_{down}, X_{left}, and X_{right}, respectively. The conditional distribution is derived by writing

$$\begin{aligned} X_j \,\big|\, (X_{\text{up}} = x_{\text{up}}, X_{\text{down}} = x_{\text{down}}, X_{\text{left}} = x_{\text{left}}, X_{\text{right}} = x_{\text{right}}) \\ = \frac{1}{4}(x_{\text{up}} + x_{\text{down}} + x_{\text{left}} + x_{\text{right}}) + \Phi_j, \end{aligned} \tag{5}$$

where Φ_j is a random innovation process. For boundary pixels, an appropriate modification reflecting the a priori belief of the extension of the image outside the field of view must be incorporated. In a large variety of application, Φ_j is assumed to follow a normal distribution

$$\Phi_j \sim \mathcal{N}\left(0, \sigma_j^2\right),$$

the variance σ_j^2 reflecting the expected deviation from the average intensity of the neighboring pixels. The Markov model can be expressed in matrix–vector form as

$$\mathsf{L}X = \Phi,$$

where the matrix L is the five-point stencil discretization of the Laplacian in two dimensions and the vector $\Phi \in \mathbb{R}^N$ contains the innovation terms Φ_j. As we assume the innovation terms to be independent, the probability distribution of Φ is

$$\Phi \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_N^2 \end{bmatrix},$$

and the resulting prior model is a *second-order Gaussian smoothness prior*,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2}\|\Sigma^{-1/2}\mathsf{L}x\|^2\right).$$

Observe that the variances σ_j^2 allow a spatially inhomogeneous a priori control of the texture of the image. Replacing the averaging weights $1/4$ in (5) by more general weights p_k, $1 \le k \le 4$ leads to a smoothness prior with directional sensitivity. Random draws from such anisotropic Gaussian priors are shown in Fig. 1, where each pixel with coordinate vector r_j in a quadrilateral grid has eight neighboring pixels with coordinates r_j^k, and the corresponding weights p_k are chosen as

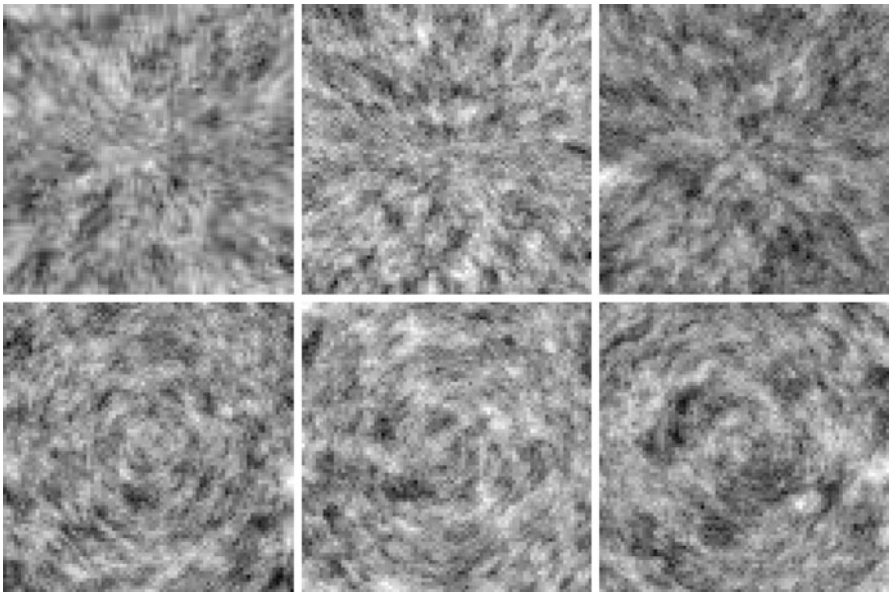

Fig. 1 Random draws from anisotropic Markov models. In the *top row*, the Markov model assumes stronger dependency between neighboring pixels in the radial than in angular direction, while in the *bottom row*, the roles of the directions are reversed. See text for a more detailed discussion

$$p_k = \frac{1}{\tau} \frac{\left(v_j^{\mathsf{T}} \left(r_j - r_j^k \right) \right)^2}{\left| r_j - r_j^k \right|^2}, \quad \tau = 1.1,$$

and the unit vector v_j is chosen either as a vector pointing out of the center of the image (top row) or in a perpendicular direction (bottom row). The former choice thus assumes that pixels are more strongly affected by the adjacent values in the radial direction, while in the latter case, they have less influence than those in the angular direction. The factor τ is added to make the matrix diagonally dominated.

The just described construction of the smoothness prior is a particular instance of priors based on the assumption that the image is a *Markov random field*, (MRF). Similarly to the four-point average example, Markov random fields assume that the conditional probability distribution of a single pixel value X_j conditioned on the remaining image depends only on the neighbors of X_j,

$$\pi \left(x_j \mid x_k, k \neq j \right) = \pi \left(x_j \mid x_k \in N_j \right),$$

where N_j is the list of neighbor pixels of X_j, such as the four adjacent pixels in the model (5). In fact, the *Hammersley–Clifford theorem* (see [5]) states that prior distributions of MRF models are of the form

$$\pi_{\text{prior}}(x) \propto \exp\left(-\sum_{j=1}^{N} V_j(x) \right),$$

where the function $V_j(x)$ depends only on x_j and its neighbors. The simplest model in this family is a Gaussian white noise prior, where $N_j = \emptyset$ and $V_j(x) = x_j^2/(2\sigma^2)$, that is,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2\sigma^2} \|x\|^2 \right).$$

Observe that this prior assumes mutual independency of the pixels, which has qualitative repercussions on the images based on it.

There is no theoretical reason to restrict the MRFs to Gaussian fields, and in fact, some of the non-Gaussian fields have had a remarkable popularity and success in the imaging context. Two non-Gaussian priors are particularly worth mentioning here, the ℓ^1-prior, where $N_j = \emptyset$ and $V_j(x) = \alpha |x_j|$, that is,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\alpha \|x\|_1 \right), \quad \|x\|_1 = \sum_{j=1}^{N} |x_j|,$$

and the closely related total variation (TV) prior,

$$\pi_{\text{prior}}(x) \propto \exp\left(-\alpha \text{TV}(x)\right), \quad \text{TV}(x) = \sum_{j=1}^{N} V_j(x),$$

with

$$V_j(x) = \frac{1}{2} \sum_{k \in N_j} |x_j - x_k|.$$

The former is suitable for imaging sparse images, where all but few pixels are believed to coincide with the background level that is set to zero. The latter prior is particularly suitable for blocky images, that is, for images consisting of piecewise smooth simple shapes. There is a strong connection to the recently popular concept of *compressed sensing*, see, for example, [11].

MRF priors, or priors with only local interaction between pixels, are by far the most commonly used priors in imaging. It is widely accepted and to some extent demonstrated (see [6] and the discussion in it) that the posterior density is sensitive to local properties of the prior only, while the global properties are predominantly determined by the likelihood. Thus, as far as the role of priors is concerned, it is important to remember that until the likelihood is taken into account, there is no connection with the measured data, hence no reason to believe that the prior should generate images that in the large scale resemble what we are looking for. In general, priors are usually designed to carry very general often qualitative and local information, which will be put into proper context with the guidance of the data through the integration with the likelihood. To demonstrate the local structure implied by different priors, in Fig. 2, we show some random draws from the priors discussed above.

Likelihood: Forward Model and Statistical Properties of Noise

If an image is worth a thousand words, a proper model of the noise corrupting it is worth at least a thousand more, in particular when the processing is based on the statistical methods. So far, the notion of noise has remained vague, and its role unclear. It is the noise, and in fact its statistical properties, that determines the likelihood density. We start by considering two very popular noise models.

Additive, nondiscrete noise: An additive noise model assumes that the data and the unknown are in a functional relation of the form

$$y = F(x) + e, \tag{6}$$

where e is the noise vector. If the function F is linear, or it has been linearized, the problem simplifies to

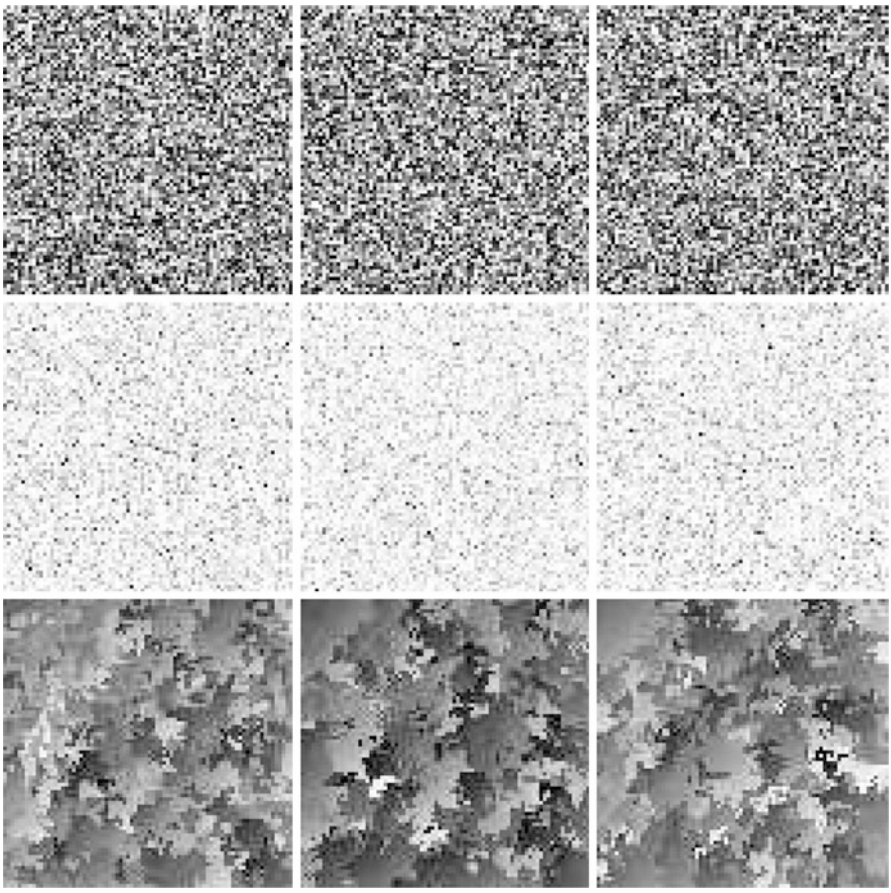

Fig. 2 Random draws from various MRF priors. *Top row*: white noise prior. *Middle row*: sparsity prior or ℓ^1-prior with positivity constraint. *Bottom row*: total variation prior

$$y = \mathsf{A}x + e. \tag{7}$$

The stochastic extension of (6) is

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

where Y, X, and E are multivariate random vectors.

The form of the likelihood is determined not only by the assumed probability distributions of Y, X, and E but also by the dependency between pairs of these variables. In the simplest case, X and E are assumed to be mutually independent and the probability density of the noise vector known,

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

resulting in a likelihood function of the form

$$\pi(y \mid x) \propto \pi_{\text{noise}}\,(y - F(x))\,,$$

which is one of the most commonly used in applications. A particularly popular model for additive noise is a Gaussian noise,

$$E \sim \mathcal{N}(0, \Sigma),$$

where the covariance matrix Σ is positive definite. Therefore, if we write $\Sigma^{-1} = \mathsf{D}^{\mathsf{T}}\mathsf{D}$, where D can be the Cholesky factor of Σ^{-1} or $\mathsf{D} = \Sigma^{-1/2}$, the likelihood can be written as

$$\begin{aligned}\pi(y \mid x) &\propto \exp\left(-\frac{1}{2}(y - F(x))^{\mathsf{T}}\Sigma^{-1}(y - F(x))\right)\\ &= \exp\left(-\frac{1}{2}\|\mathsf{D}(y - F(x))\|^2\right).\end{aligned} \tag{8}$$

In the general case where X and E are not independent, we need to specify the joint density

$$(X, E) \sim \pi(x, e)$$

and the corresponding conditional density

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

In this case, the likelihood becomes

$$\pi(y \mid x) \propto \pi_{\text{noise}}(y - F(x) \mid x).$$

This clearly demonstrates the problems which may arise if we want to adhere to the claim that "likelihood should be independent of the prior." Because the interdependency of the image x and the noise is much more common than we might be inclined to believe, the independency of noise and signal is often in conflict with reality. An instance of such situation occurs in electromagnetic brain imaging using magnetoencephalography (MEG) or electroencephalography (EEG), when the eye muscle during a visual task acts as noise source but can hardly be considered as independent from the brain activation due to a visual stimulus. Another example related to boundary conditions will be discussed later on. Also, since the noise term should account not only for the exogenous measurement noise but also for the shortcomings of the model, including discretization errors, the interdependency is in fact a ubiquitous phenomenon too often neglected.

Most additive noise models assume that the noise follows a Gaussian distribution, with zero mean and given covariance. The computational advantages of a Gaussian likelihood are rather formidable and have been a great incentive to use Gaussian approximations of non-Gaussian densities. While it is commonplace and somewhat justified, for example, to approximate Poisson densities with Gaussian densities when the mean is sufficiently large [14], there are some important imaging applications where the statistical distribution of the noise must be faithfully represented in the likelihood.

Counting noise: The weakness of a signal can complicate the deblurring and denoising problem, as is the case in some image processing applications in astronomy [49, 57, 63], microscopy [45, 68], and medical imaging [29, 60]. In fact, in the case of weak signals, a charge-coupled device (CCD), instead of recording an integrated signal over a time window, counts individual photons or electrons. This leads to a situation where the noise corrupting the recorded signal is no longer exogenous but rather an intrinsic property of the signal itself, that is, the input signal itself is a random process with an unpredictable behavior. Under rather mild assumptions – stationarity, independency of increments, and zero probability of coincidence – it can be shown (see, e.g., [62]) that the counting signal follows a Poisson distribution. Consider, for example, the astronomical image of a very distant object, collected with an optical measurement device whose blurring is described by a matrix A. The classical description of such data would follow (7), with the error term collecting the background noise and the thermal noise of the device. The corresponding counting model is

$$y_j \sim \text{Poisson}\left((\mathsf{A}x)_j + b\right),\ y_j, y_k \text{ independent if } j \neq k,$$

or, explicitly,

$$\pi(y \mid x) = \prod_{j=1}^{m} \frac{((\mathsf{A}x)_j + b)^{y_j}}{(y_j)!} \exp\left(-(\mathsf{A}x)_j + b\right),$$

where $b \geq 0$ is a background radiation level, assumed known. Observe that while the data are counts, therefore integer numbers, the expectation need not to be.

Similar or slightly modified likelihoods can be used to model the positron emission tomography (PET) and single-photon emission computed tomography (SPECT) signals; see [29, 54].

The latter example above demonstrates clearly that the description of imaging problems as linear or nonlinear, without a specification of the noise model, in the context of statistical methods, does not play a significant role: Even if the expectation is linear, traditional algorithms for solving linear inverse problems are useless, although they may turn out to be useful within iterative solvers for solving locally linearized steps.

Maximum Likelihood and Fisher Information

When switching to a parametric non-Bayesian framework, the statistical inference problem amounts to estimating a deterministic parameter that identifies the probability distribution from which the observations are drawn. To apply this framework in imaging problems, the underlying image x, which in the Bayesian context was itself a random variable, can be thought of as a parameter vector that specifies the likelihood function,

$$f(y;x) = \pi(y \mid x),$$

as implied by the notation $f(y;x)$ also.

In the non-Bayesian interpretation, a measure of how much information about the parameter x is contained in the observation is given in terms of the *Fisher information matrix* J,

$$\mathsf{J}_{j,k} = \mathsf{E}\left\{\frac{\partial \log f}{\partial x_j}\frac{\partial \log f}{\partial x_k}\right\} = \int \frac{\partial \log f(y;x)}{\partial x_j}\frac{\partial \log f(y;x)}{\partial x_k} f(y;x)dy. \tag{9}$$

In this context, the observation y only is a realization of a random variable Y, whose probability distribution is entirely determined by the distribution of the noise. The gradient of the logarithm of the likelihood function is referred to as the *score*, and the Fisher information matrix is therefore the covariance of the score.

Assuming that the likelihood is twice continuously differentiable and regular enough to allow the exchange of integration and differentiation, it is possible to derive another useful expression for the information matrix. It follows from the identity

$$\frac{\partial \log f}{\partial x_k} = \frac{1}{f}\frac{\partial f}{\partial x_k}, \tag{10}$$

that we may write the Fisher information matrix as

$$\mathsf{J}_{j,k} = \int \frac{\partial \log f}{\partial x_j}\frac{\partial f}{\partial x_k}dy = \frac{\partial}{\partial x_k}\int \frac{\partial \log f}{\partial x_j} f dy - \int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy.$$

Using the identity (10) with k replaced by j, we observe that

$$\int \frac{\partial \log f}{\partial x_j} f dy = \int \frac{\partial f}{\partial x_j} dy = \frac{\partial}{\partial x_j}\int f dy = 0,$$

since the integral of f is one, which leads us to the alternative formula

$$\mathsf{J}_{j,k} = -\int \frac{\partial^2 \log f}{\partial x_j \partial x_k} f dy = -\mathsf{E}\left\{\frac{\partial^2 \log f}{\partial x_j \partial x_k}\right\}. \tag{11}$$

The Fisher information matrix is closely related to non-Bayesian estimation theory. This will be discussed later in connection with maximum likelihood estimation.

Informative or Noninformative Priors?

Not seldom the use of priors in imaging applications is blamed for biasing the solution in a direction not supported by the data. The concern of the use of committal priors has led to the search of "noninformative priors" [39] or weak priors that would "let the data speak."

The strength or weakness of a prior is a rather elusive concept, as the importance of the prior in Bayesian imaging is in fact determined by the likelihood: the more information we have about the image in data, the less has to be supplied by the prior. On the other hand, in imaging applications where the likelihood is built on very few data points, the prior needs to supply the missing information, hence has a much more important role. As pointed out before, it is a common understanding that in imaging applications, prior should carry small-scale information about the image that is missing from the likelihood that in turn carries information about the large-scale features and in that sense complements the data.

Adding Layers: Hierarchical Models

Consider the following simple denoising problem with additive Gaussian noise,

$$Y = X + N, \quad N \sim \mathcal{N}(0, \Sigma),$$

with noise covariance matrix Σ presumed known, whose likelihood model is tantamount to saying that

$$Y \mid X = x \sim \mathcal{N}(x, \Sigma).$$

From this perspective, the denoising problem is reduced to estimating the mean of a Gaussian density in the non-Bayesian spirit, and the prior distribution is a *hierarchical* model, expressing the degree of uncertainty of the mean x.

Parametric models are common when defining the prior densities, but similarly to the above interpretation of the likelihood, the parameters are often poorly known. For example, when introducing a prior

$$X \sim \mathcal{N}(\theta, \Gamma)$$

with unknown θ, we are expressing a qualitative prior belief that "X differs from an unknown value by an error with a given Gaussian statistics," which says very little about the values of X itself unless information about θ is provided. Similarly as in the denoising problem, it is natural to augment the prior with another layer of

information concerning the parameter θ. This layering of the inherent uncertainty is at the core of *hypermodels*, or Bayesian hierarchical models. Hierarchical models are not restricted to uncertainties in the prior, but can be applied to lack of information of the likelihood model as well.

In hierarchical models, both the likelihood and the prior may depend on additional parameters,

$$\pi(y \mid x) \to \pi(y \mid x, \gamma), \quad \pi_{\text{prior}}(x) \to \pi_{\text{prior}}(x \mid \theta),$$

with both parameters γ and θ poorly known. In this case, it is natural to augment the model with *hyperpriors*. Assuming for simplicity that the parameters γ and θ are mutually independent so that we can define the hyperprior distributions $\pi_1(\gamma)$ and $\pi_2(\theta)$, the joint probability distribution of all the unknowns is

$$\pi(x, y, \theta, \gamma) = \pi(y \mid x, \gamma)\pi_{\text{prior}}(x \mid \theta)\pi_1(\gamma)\pi_2(\theta).$$

From this point on, the Bayesian inference can proceed along different paths. It is possible to treat the hyperparameters as nuisance parameters and marginalize them out by computing

$$\pi(x, y) = \int\int \pi(x, y, \theta, \gamma) d\theta d\gamma$$

and then proceed as in a standard Bayesian inference problem. Alternatively, the hyperparameters can be included in the list of unknowns of the problem and their posterior density

$$\pi(\xi \mid y) = \frac{\pi(x, y, \theta, \gamma)}{\pi(y)}, \quad \xi = \begin{bmatrix} x \\ \theta \\ \gamma \end{bmatrix}$$

needs to be explored. The estimation of the hyperparameters can be based on the optimization or on the *evidence*, as will be illustrated below with a specific example.

To clarify the concept of a hierarchical model itself, we consider some examples where hierarchical models arise naturally.

Blind deconvolution: Consider the standard deblurring problem defined in section "Imaging Problems." Usually, it is assumed that the blurring kernel A is known, and the likelihood, with additive Gaussian noise with covariance Σ, becomes

$$\pi(y \mid x) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x)\right). \tag{12}$$

In some cases, although A is poorly known, its parametric expression is known and the uncertainty only affects the values of some parameters, as is the case when the shape of the continuous convolution kernel $a(r - s)$ is known but the actual width is not. If we express the kernel a as a function of a width parameter,

$$a(r-s) = a_\gamma(r-s) = \frac{1}{\gamma} a_1(\gamma(r-s)), \quad \gamma > 0,$$

and denote by A_γ the corresponding discretized convolution matrix, the likelihood becomes

$$\pi(y \mid x, \gamma) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}_\gamma x)^{\mathsf{T}} \Sigma^{-1} (y - \mathsf{A}_\gamma x)\right),$$

and additional information concerning γ, for example, bound constraints, can be included via a hyperprior density.

The procedure just outlined can be applied to many problems arising from adaptive optics imaging in astronomy [52]; while the uncertainty in the model is more complex than in the explanatory example above, the approach remains the same.

Conditionally Gaussian hypermodels: Gaussian prior models are often criticized for being a too restricted class, not being able to adequately represent prior beliefs concerning, for example, the sparsity or piecewise smoothness of the solution. The range of qualitative features that can be expressed with normal densities can be considerably expanded by considering *conditionally Gaussian* families instead. As an example, consider the problem of finding a sparse image from linearly blurred noisy copy of it. The likelihood model in this case may be written as in (12). To set up an appropriate prior, consider a conditionally Gaussian prior

$$\begin{aligned} \pi_{\text{prior}}(x \mid \theta) &\propto \left(\frac{1}{\theta_1 \cdots \theta_N}\right)^{1/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{N} \frac{x_j^2}{\theta_j}\right) \\ &= \exp\left(-\frac{1}{2}\sum_{j=1}^{N}\left[\frac{x_j^2}{\theta_j} + \log\theta_j\right]\right). \end{aligned} \tag{13}$$

If $\theta_j = \theta_0 =$ constant, we obtain the standard white noise prior which cannot be expected to favor sparse solutions. On the other hand, since θ_j is the variance of the pixel X_j, sparse images correspond to vectors θ with most of the components close to zero. Since we do not know a priori which of the variances should significantly differ from zero, when choosing a stochastic model for θ, it is reasonable to select a hyperprior that favors sparsity without actually specifying the location of the outliers. Two distributions that are particularly well suited for this are the *gamma distribution*,

$$\theta_j \sim \text{Gamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{k-1} \exp\left(-\frac{\theta_j}{\theta_0}\right),$$

and the *inverse gamma distribution*,

$$\theta_j \sim \text{InvGamma}(k, \theta_0), \quad k, \theta_0 > 0, \quad \pi(\theta_j) = \theta_j^{-k-1} \exp\left(-\frac{\theta_0}{\theta_j}\right).$$

The parameters k and θ_0 are referred to as the shape and the scaling, respectively. The inverse gamma distribution corresponds to assuming that the *precision*, defined as $1/\theta_j$, is distributed according to the gamma distribution $\text{Gamma}(k, 1/\theta_0)$. The computational price of introducing hyperparameters is that instead of one image x, we need to estimate the image x and its variance image θ. Fortunately, for conditionally Gaussian families, there are efficient algorithms for computing these estimates, which will be discussed in the section concerning algorithms.

The hyperprior based on the gamma distribution, in turn, contains parameters (k and θ_0) to be determined. Nothing prevents us from defining another layer of hyperpriors concerning these values. It should be noted that in hierarchical models, the selection of the parameters higher up in the hierarchy tends to have less direct effect on the parameters of primary interest. Since this last statement has not been formally proved to be true, it should be considered as a piece of computational folklore.

Conditionally Gaussian hypermodels have been successfully applied in machine learning [66], in electromagnetic brain activity mapping [16], and in imaging applications for restoring blocky images [15]. Recently, their use in compressed sensing has been proposed [40].

4 Numerical Methods and Case Examples

The solution of an imaging inverse problem in the statistical framework is the posterior probability density. Because this format of the solution is not practical for most applications, it is common to summarize the distribution in one or a few images. This leads to the challenging problem of exploring the posterior distributions and finding single estimators supported by the distribution.

Estimators

In this section, we review some of the commonly used estimators and subsequently discuss some of the popular algorithms suggested in the literature to compute the corresponding estimates.

Prelude: Least Squares and Tikhonov Regularization

In the case where the forward model is linear, the problem of estimating an image from a degraded, noisy recording is equivalent in a determinist setting to looking for a solution of a linear system of equations of the form

$$\mathsf{A}x = y, \tag{14}$$

where the right-hand side is corrupt by noise. When A is not a square matrix and/or it is ill conditioned, one needs to specify what a "solution" means. The most straightforward way is to specify it as a least squares solution.

There is a large body of literature, and a wealth of numerical algorithms, for the solution of large-scale least squares problems arising from problems similar to imaging applications (see, e.g., [9]). Since dimensionality alone makes these problems computationally very demanding, they may require an unreasonable amount of computer memory and operations unless a compact representation of the matrix A can be exploited. Many of the available algorithms make additional assumptions about either the underlying image or the structure of the forward model regardless of whether there is a good justification.

In a determinist setting, the entries of the least squares solution of (14) with a right-hand side corrupted by noise are not necessarily in the gray-scale range of the image pixels. Moreover, the inherent ill conditioning of the problem, which varies with the imaging modality and the conditions under which the observations were collected, usually requires requlariza tion, see, for example, [4, 33, 34, 41]. A standard regularization method is to replace the original ill-posed least squares problem by a nearby well-posed problem by introducing a penalty term to avoid that the computed solution is dominated by amplified noise components, reducing the problem to minimizing a functional of the form

$$T(x) = \|\mathsf{A}x - y\|^2 + \alpha J(x), \tag{15}$$

where $J(x)$ is the penalty functional and $\alpha > 0$ is the regularization parameter. The minimizer of the functional (15) is the *Tikhonov regularized solution*. The type of additional information used in the design of the penalty term may include upper bounds on the norm of the solution or of its derivatives, nonnegative constraints for its entries, or bounds on some of the components. Often, expressing characteristics that are expected of the sought image in qualitative terms is neither new nor difficult: the translation of these beliefs into mathematical terms and their implementation is a more challenging step.

Maximum Likelihood and Maximum A Posteriori

We begin with the discussion of the maximum likelihood estimator in the framework of non-Bayesian statistics and denote by x a deterministic parameter determining the likelihood distribution of the data, modeled as a random variable. Let $\hat{x} = \hat{x}(y)$ denote an estimator of x, based on the observations y. Obviously, $\hat{x}$ is also a random variable, because of its dependency on the stochastic observations y; moreover, it is an *unbiased estimator* if

$$\mathsf{E}\{\hat{x}(y)\} = x,$$

that is, if, in the average, it returns the exact value. The covariance matrix C of an unbiased estimator therefore measures the statistical variation around the true value,

$$\mathsf{C}_{j,k} = \mathsf{E}\left\{(\hat{x}_j - x_j)(\hat{x}_k - x_k)\right\},$$

thus the name mean square error. Evidently, the smaller the mean square error, for example, in the sense of quadratic forms, the higher the expected fidelity of the estimator. The Fisher information matrix (9) gives a lower bound for the covariance matrix of all unbiased estimators. Assuming that J is invertible, the *Cramér–Rao lower bound* states that for an unbiased estimator,

$$\mathsf{J}^{-1} \leq \mathsf{C}$$

in the sense of quadratic forms, that is, for any vector

$$u^{\mathsf{T}}\mathsf{J}^{-1}u \leq u^{\mathsf{T}}\mathsf{C}u.$$

An estimator is called *efficient* if the error covariance reaches the Cramér–Rao bound.

The maximum likelihood estimator $\hat{x}_{\mathrm{ML}}(y)$ is the maximizer of the function $x \mapsto f(x; y)$, and in practice, it is found by locating the zero(s) of the score,

$$\nabla_x \log f(x; y) = 0 \Rightarrow x = \hat{x}_{\mathrm{ML}}(y).$$

Notice that in the non-Bayesian context, likelihood refers solely to the likelihood of the observations y, and the maximum likelihood estimation is a way to choose the underlying parametric model so that the observations become as likely as possible.

The popularity of the maximum likelihood estimator, in addition to being an intuitively obvious choice, stems from the fact that it is asymptotically efficient estimator in the sense that when the number of independent observations of the data increases, the covariance of the estimator converges toward the inverse of the Fisher information matrix, assuming that it exists. More precisely, assuming a sequence $y^1, y^2, \ldots$ of independent observations and defining $\hat{x}^n = \hat{x}\left(y^1, \ldots, y^n\right)$ as

$$\hat{x}^n = \mathrm{argmax}\left\{\frac{1}{n}\sum_{j=1}^{n} f\left(x, y^j\right)\right\},$$

asymptotically the probability distribution of $\hat{x}^n$ approaches a Gaussian distribution with mean x and covariance J^{-1}.

The assumption of the regularity of the Fisher information matrix limits the use of the ML estimator in imaging applications. To understand this claim, consider the simple case of linear forward model and additive Gaussian noise,

$$Y = \mathsf{A}x + E, \quad E \sim \mathcal{N}(0, \Sigma).$$

The likelihood function in this case is

$$f(x;y) = \left(\frac{1}{2\pi|\Sigma|}\right)^{1/N} \exp\left(-\frac{1}{2}(y-\mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y-\mathsf{A}x)\right),$$

from which it is obvious that by formula (11),

$$\mathsf{J} = \mathsf{A}^{\mathsf{T}}\Sigma^{-1}\mathsf{A}.$$

In the simplest imaging problems such as of denoising, the invertibility of J is not an issue. However, in more realistic and challenging applications such as deblurring, the ill conditioning of A renders J singular, and the Cramér–Rao bound becomes meaningless. It is not uncommon to regularize the information matrix by adding a diagonal weight to it which, from the Bayesian viewpoint, is tantamount to adding prior information but in a rather uncontrolled manner.

For further reading of mathematical methods in estimation theory, we refer to [17, 46, 50].

We consider the maximum likelihood estimator in the context of regularization and Bayesian statistics. In the case of a Gaussian additive noise observation model, under the assumption that the noise at each pixel is independent of the signal and that the forward map is linear, $F(x) = \mathsf{A}x$, the likelihood (8) is of the form

$$\pi(y \mid x) \propto \exp\left(-\frac{1}{2}\|\mathsf{D}(\mathsf{A}x-y)\|^2\right),$$

where Σ is the noise covariance matrix and $\mathsf{D}^{\mathsf{T}}\mathsf{D} = \Sigma^{-1}$ is the Cholesky decomposition of its inverse. The maximizer of the likelihood function is the solution of the minimization problem

$$x_{\mathrm{ML}} = \operatorname{argmin}\left\{\|\mathsf{D}(\mathsf{A}x-y)\|^2\right\},$$

which, in turn, is the least squares solution of the linear system

$$\mathsf{D}\mathsf{A}x = \mathsf{D}y.$$

Thus, we can reinterpret least squares solutions as maximum likelihood estimates under an additive, independent Gaussian error model. Within the statistical framework, the maximum likelihood estimator is defined analogously for any error model which admits a maximizer for the likelihood, but in the general case, the computation of the minimizer cannot be reduced to the solution of a linear least squares problem.

In a statistical framework, the addition of a penalty terms to keep the solution of the least squares problem from becoming dominated by amplified noise components is tantamount to using a prior to augment the likelihood. If the observation model is linear, the prior and the likelihood are both Gaussian,

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

and the noise is independent of the signal, the corresponding posterior is of the form

$$\pi(x \mid y) \propto \exp\left(-\frac{1}{2}\left(\|\mathsf{D}(\mathsf{A}x - y)\|^2 + \|\mathsf{R}x\|^2\right)\right),$$

where R satisfies $\mathsf{R}^{\mathsf{T}}\mathsf{R} = \Gamma^{-1}$, so typically it is the Cholesky factor of Γ^{-1} or alternatively, $\mathsf{R} = \Gamma^{-1/2}$.

The maximizer of the posterior density, or the maximum a posteriori (MAP) estimate, is the minimizer of the negative exponent, hence the solution of the minimization problem

$$\begin{aligned} x_{\text{MAP}} &= \text{argmin}\{\|\mathsf{D}(\mathsf{A}x - y)\|^2 + \|\mathsf{R}x\|^2\} \\ &= \text{argmin}\left\{\left\|\begin{bmatrix}\mathsf{DA}\\ \mathsf{R}\end{bmatrix}x - \begin{bmatrix}\mathsf{D}y\\ 0\end{bmatrix}\right\|^2\right\}, \end{aligned}$$

or, equivalently, the Tikhonov solution (15) with penalty $J(x) = \|\mathsf{R}x\|^2$ and regularization parameter $\alpha = 1$. Again, it is important to note that the direct correspondence between the Tikhonov regularization and the MAP estimate only holds for linear observation models and Gaussian likelihood and prior. The fact that the MAP estimate in this case is the least squares solution of the linear system

$$\begin{bmatrix}\mathsf{DA}\\ \mathsf{R}\end{bmatrix}x = \begin{bmatrix}\mathsf{D}y\\ 0\end{bmatrix} \tag{16}$$

is a big incentive to stay with Gaussian likelihood and Gaussian priors as long as possible.

As in the case of the ML estimate, the definition of MAP estimate is independent of the form of the posterior, hence applied also to non-Gaussian, nonindependent noise models, with the caveat that in the general case, the search for a maximizer of the posterior may require much more sophisticated optimization tools.

Conditional Means

The recasting in statistical terms of imaging problems effectively shifts the interest from the image itself to its probability density. The ML and MAP estimators discussed in the previous section suffer from the limitations, which come from summarizing an entire distribution with one realization. The ML estimator is known to suffer from instabilities due to the typical ill conditioning of the forward map in imaging problems, and it will not be discussed further here. The computed MAP estimate, on the other hand, may correspond to an isolated spike in the probability density away from the bulk of the mass of the density, and its computation may suffer from numerical complications. Furthermore, a conceptually more serious

limitation is the fact that MAP estimators do not carry information about the statistical dispersion of the distribution. A tight posterior density suggests that any ensemble of images which are in statistical agreement with the data and the given prior show little variability; hence, any realization from that ensemble can be thought of as very representative of the entire family. A wide posterior, on the other hand, suggests that there is a rather varied family of images that are in agreement with the data and the prior, hence lowering the representative power of any individual realization.

In the case where either the likelihood or the prior is not Gaussian, the mean of the posterior density, often referred to as conditional mean (CM) or posterior mean, may be a better choice because it is the estimator with least variance (see [3, 41]). Observe, however, that in the fully Gaussian case, the MAP and CM estimate coincides.

The CM estimate is, by definition,

$$x_{\mathrm{CM}} = \int_{\mathbb{R}^N} x\pi(x \mid y)dx,$$

while the a posteriori covariance matrix is

$$\Gamma_{\mathrm{CM}} = \int_{\mathbb{R}^N} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathsf{T}}\pi(x \mid y)dx,$$

hence requiring the evaluation of the high-dimensional integrals. When the integrals have no closed form solution, as is the case for many imaging problems where, for example, the a priori information contains bounds on pixel values, a numerical approximation of the integral must be used to estimate x_{CM} and Γ_{CM}. The large dimensionality of the parameter space, which easily is of the order of hundreds of thousands when x represents an image, rules out the use of standard numerical quadratures, leaving Monte Carlo integration the only currently known feasible alternative.

The conceptual simplicity of Monte Carlo integration, which estimates the integral value as the average of a large sample of the integrand evaluated over the support of the integration, requires a way of generating a large sample from the posterior density. The generation of a sample from a given distribution is a well-known problem in statistical inference, which has inspired families of sampling schemes generically referred to as Markov chain Monte Carlo (MCMC) methods, which will be discussed in section "Markov Chain Monte Carlo Sampling."

Once a representative sample from the posterior has been generated, the CM estimate is approximately the sample mean. By definition, the CM estimate must be near the bulk of the density, although it is not necessarily a highly probable point. In fact, for multimodal distributions, the CM estimate may fall between the modes of the density and even belong to a subset of $\mathbb{R}^N$ with probability zero, although such a situation is rather easy to detect. There is evidence, however, that in some imaging applications the CM estimate is more stable than the MAP estimate; see

[23]. While the robustness of the CM estimate does not compensate for the lack of information about the width of the posterior, the possibility of estimating the posterior covariance matrix via sampling is an argument for the sampling approach, since the sample can also be used to estimate the posterior width.

Algorithms

The various estimators based on the posterior distribution are simple to define, but the actual computation may be a major challenge. In the case of Gaussian likelihood and prior, combined with linear forward map, the MAP and CM estimates coincide and an explicit formula exists. If the problem is very high dimensional, even this case may be computationally challenging. Before going to specific algorithms, we review the linear Gaussian theory.

The starting point is the linear additive model

$$Y = \mathsf{A}X + E, \quad X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma).$$

Here, we assume that the mean of X and the noise E both vanish, an assumption that is easy to remove. Above, X and E need not be mutually independent, and we may postulate that they are jointly Gaussian and the cross-correlation matrix

$$\mathsf{C} = \mathsf{E}\left\{XE^{\mathsf{T}}\right\} \in \mathbb{R}^{N\times M}$$

may not vanish. The joint probability distribution of X and Y is also Gaussian, with zero mean and variance

$$\begin{aligned}\mathsf{E}\left\{\begin{bmatrix}X\\Y\end{bmatrix}\begin{bmatrix}X^{\mathsf{T}} & Y^{\mathsf{T}}\end{bmatrix}\right\} &= \mathsf{E}\left\{\begin{bmatrix}XX^{\mathsf{T}} & X(\mathsf{A}X+E)^{\mathsf{T}}\\ (\mathsf{A}X+E)X^{\mathsf{T}} & (\mathsf{A}X+E)(\mathsf{A}X+E)^{\mathsf{T}}\end{bmatrix}\right\}\\ &= \begin{bmatrix}\Gamma & \Gamma\mathsf{A}^{\mathsf{T}}+\mathsf{C}\\ \mathsf{A}\Gamma+\mathsf{C}^{\mathsf{T}} & \mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}}+\Sigma\end{bmatrix}.\end{aligned}$$

Let $\mathsf{L} \in \mathbb{R}^{(N+M)\times(N+M)}$ denote the inverse of the above matrix, assuming that it exists, and write a partitioning of it in blocks according to the dimensions N and M,

$$\mathsf{L} = \begin{bmatrix}\Gamma & \Gamma\mathsf{A}^{\mathsf{T}}+\mathsf{C}\\ \mathsf{A}\Gamma+\mathsf{C}^{\mathsf{T}} & \mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}}+\Sigma\end{bmatrix}^{-1} = \begin{bmatrix}\mathsf{L}_{11} & \mathsf{L}_{12}\\ \mathsf{L}_{21} & \mathsf{L}_{22}\end{bmatrix}.$$

With this notation, the joint probability distribution of X and Y is

$$\pi(x, y) \propto \exp\left(-\frac{1}{2}(x^{\mathsf{T}}\mathsf{L}_{11}x + x^{\mathsf{T}}\mathsf{L}_{12}y + y^{\mathsf{T}}\mathsf{L}_{21}x + y^{\mathsf{T}}\mathsf{L}_{22}y)\right).$$

To find the posterior density, one completes the square in the exponent with respect to x,

$$\pi(x \mid y) \propto \exp\left(-\frac{1}{2}\left(x - \mathsf{L}_{11}^{-1}\mathsf{L}_{12}y\right)^{\mathsf{T}}\mathsf{L}_{11}\left(x - \mathsf{L}_{11}^{-1}\mathsf{L}_{12}y\right)\right),$$

where terms independent of x that contribute only to the normalization are left out. Therefore,

$$X \mid Y = y \sim \mathcal{N}\left(\mathsf{L}_{11}^{-1}\mathsf{L}_{12}y, \mathsf{L}_{11}^{-1}\right).$$

Finally, we need to express the matrix blocks L_{ij} in terms of the matrices of the model. The expressions follow from the classical matrix theory of Schur complements [24]: We have

$$\mathsf{L}_{11}^{-1} = \Gamma - \left(\Gamma\mathsf{A}^{\mathsf{T}} + \mathsf{C}\right)\left(\mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}} + \Sigma\right)^{-1}\left(\mathsf{A}\Gamma + \mathsf{C}^{\mathsf{T}}\right), \tag{17}$$

and

$$\mathsf{L}_{11}^{-1}\mathsf{L}_{12}y = \left(\Gamma\mathsf{A}^{\mathsf{T}} + \mathsf{C}\right)\left(\mathsf{A}\Gamma\mathsf{A}^{\mathsf{T}} + \Sigma\right)^{-1}y. \tag{18}$$

Although a closed form solution, to evaluate the expression (18) for the posterior mean may require iterative solvers.

When the image and the noise are mutually independent, implying that $\mathsf{C} = 0$, we find a frequently encountered form of the MAP estimate arising from writing the Gaussian posterior density directly by using Bayes' formula, that is,

$$\begin{aligned}\pi(x \mid y) &\propto \pi_{\text{prior}}(x)\pi(y \mid x)\\ &\propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\Gamma^{-1}x - \frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x)\right),\end{aligned}$$

and so the MAP estimate, and simultaneously the posterior mean estimate, is the maximizer of the above expression, or, equivalently, the minimizer of the quadratic functional

$$H(x) = (y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x) + x^{\mathsf{T}}\Gamma^{-1}x.$$

By substituting the factorizations

$$\Sigma^{-1} = \mathsf{D}^{\mathsf{T}}\mathsf{D}, \quad \Gamma^{-1} = \mathsf{R}^{\mathsf{T}}\mathsf{R},$$

the minimization problem becomes the previously discussed standard least squares problem of minimizing

$$H(x) = \|\mathsf{D}(y - \mathsf{A}x)\|^2 + \|\mathsf{R}x\|^2, \tag{19}$$

leading to the least squares problem (16). Whether one should use this formula or (18) depends on the application and, in particular, on the sparsity properties of the covariance matrices and their inverses.

Iterative Linear Least Squares Solvers

The computation of the ML or MAP estimate under the Gaussian additive linear noise model and, in the latter case, with a Gaussian prior, amounts to the solution of system of linear equations (14), (16), or (18) in the least squares sense. Since the dimensions of the problem are proportional to the number of pixels in the image except when the observation model has a particular structure or sparsity properties which can be exploited to reduce the memory allocation, solution by direct methods is unfeasible, hence making in general the iterative solvers the methods of choice.

Among the iterative methods specifically designed for the solution of least squares problems, the LSQR version with shifts [55, 56] of the Conjugate Gradient for Least Squares (CGLS) method originally proposed in [37] combines robustness and numerical efficiency. CGLS-type iterative methods have been designed to solve the system $\mathsf{A}x = y$, minimize $\|\mathsf{A}x - y\|^2$, or minimize $\|\mathsf{A}x - y\|^2 + \delta\|x\|^2$, where the matrix A may be square or rectangular – either overdetermined or underdetermined – and may have any rank. The matrix A does not need to be stored, but instead its action is represented by a routine for computing matrix–vector products of the forms $v \mapsto \mathsf{A}v$ and $u \mapsto \mathsf{A}^{\mathsf{T}}u$.

Minimizing the expression (19) may be transformed in a standard form by writing it as

$$\min\left\{\|\mathsf{D}\left(y - \mathsf{A}\mathsf{R}^{-1}w\right)\|^2 + \|w\|^2\right\}, \quad w = \mathsf{R}x$$

In practice, the matrix R^{-1} should not be computed, unless it is trivial to obtain. Rather, R^{-1} acts as a preconditioner, and its action should be implemented together with the action of the matrix A as a routine called from the iterative linear solver. The interpretation of the action of the prior as a preconditioner has led to the concept of prior conditioner; see [12, 14] for details.

Nonlinear Maximization

In the more general case where either the observation model is nonlinear or the likelihood and prior are non-Gaussian, the computation of the ML and MAP estimates requires the solution of a maximization problem. Maximizers of nonlinear functions can be found by quasi-Newton methods with global convergence strategy. Since Newton-type methods proceed by solving a sequence of linearized problems whose dimensions are proportional to the size of the image, iterative linear solvers are typically used for the solution of the linear subproblem [20, 43]. In imaging applications, it is not uncommon that the a priori information includes nonnegativity constraints on the pixel values or bounds on their range. In these cases, the computation of the MAP estimate amounts to a constrained maximization problem and may be very challenging. Algorithms for maximization problems with nonnegativity constraints arising in imaging applications based on the projected

gradient have been proposed in the literature; see [2] and references therein. We shall not review Newton-based methods here, since usually the fine points are related to the particular applications at hand and not so much to the statistical description of the problem. Instead, we review some algorithms that stem directly from the statistical setting of the problem and are therefore different from the methods used in regularized deterministic literature.

EM Algorithm

The MAP estimator is the maximizer of the posterior density $\pi(x \mid y)$, or, equivalently, the maximizer the logarithm of it,

$$L(x \mid y) = \log \pi(x \mid y) = \log \pi(y \mid x) + \log \pi_{\text{prior}}(x) + \text{constant},$$

where the simplest form of Bayes' rule was used to represent the posterior density as a product of the likelihood and the prior. However, note that above, the vector x may represent the unknown of primary interest, or if hierarchical models are used, the model parameters related to the likelihood and/or prior may be included in it.

The *expectation–maximization* algorithm is a method developed originally for maximizing the likelihood function and later extended to the Bayesian setting to maximize the posterior density, in a situation where part of the data is "missing." While in many statistical application the concept of missing data appears natural, for example, when incomplete census data or patient data are discussed, in imaging applications, this concept is a rather arbitrary and to some extent artificial. However, during the years, EM has found its way to numerous imaging applications, partly because it often leads to algorithms that are easy to implement. Early versions of the imaging algorithms with counting data such as the Richardson–Lucy iteration [49, 57], popular in astronomical imaging, were independently derived. Later, similar EM-based algorithms were rederived in the context of medical imaging [29, 36, 60]. Although EM algorithms are discussed in more detail elsewhere in this book, we include a brief discussion here in order to put EM in the context of general statistical imaging formalism.

As pointed out above, in imaging problems, data is not missing: Data, *per definitionem*, is what one is able to observe and register. Therefore, the starting point of the EM algorithm in image applications is to augment the actual data y by *fictitious*, nonexistent data z that would make the problem significantly easier to handle.

Consider the statistical inference problem of estimating a random variable X based on an observed realization of Y, denoted by $Y = y = y_{\text{obs}}$. We assume the existence of a third random variable Z and postulate that the joint probability density of these three variables is available and is denoted by $\pi(x, y, z)$. The EM algorithm consists of the following steps:

1. Initialize $x = x^0$ and set $k = 0$.
2. *E-step*: Define the probability distribution, or a fictitious likelihood density,

$$\pi^k(z) = \pi\left(z \mid x^k, y\right) \propto \pi\left(x^k, y, z\right), \quad y = y_{\text{obs}},$$

and calculate the integral

$$Q^k(x) = \int L(x \mid y, z)\pi^k(z)dz, \quad L(x \mid y, z) = \log(\pi(x \mid y, z)). \tag{20}$$

3. *M-step*: Update x^k by defining

$$x^{k+1} = \text{argmax}\, Q^k(x). \tag{21}$$

4. If a given convergence criterion is satisfied, exit; otherwise, increase k by one and repeat from step 2 until convergence.

The E-step above can be interpreted as computing the expectation of the real-valued random variable $\log(\pi(x, y, Z))$, x and y fixed, with respect to a conditional measure of Z conditioned on $X = x^j$ and $Y = y = y_{\text{obs}}$, hence the name expectation step.

The use of the EM algorithm is often advocated on the basis of the convergence proof given in [19]. Unfortunately, the result is often erroneously quoted as an automatic guarantee of convergence, without verifying the required hypotheses. The validity of the convergence is further obfuscated by the error in the proof (see [70]), and in fact, counterexamples of lack of convergence are well known [10, 69]. We point out that as far as convergence is concerned, global convergence of quasi-Newton algorithm is well established, and compared to the EM algorithm, the algorithm is often more effective [20].

As the concept of missing data is not well defined in general, we outline the use of the EM algorithm in an example that is meaningful in imaging applications.

SPECT imaging: The example discussed here follows the article [29]. Consider the SPECT image formation problem, where the two-dimensional object is divided in N pixels, each one emitting photons that are recorded through collimators by M photon counting devices. If x_j is the expected number of photons emitted by the jth pixel, the photon count at ith photon counter, denoted by Y_i, is an integer-valued random variable and can be modeled by a Poisson process,

$$Y_i \sim \text{Poisson}\left(\sum_{j=1}^{M} a_{ij}x_j\right) = \text{Poisson}((\mathsf{A}x)_i),$$

the variables Y_i being mutually independent and the matrix elements a_{ij} of $\mathsf{A} \in \mathbb{R}^{M\times N}$ being known. We assume that X, the stochastic extension of the unknown vector $x \in \mathbb{R}^N$, is a priori distributed according to a certain probability distribution,

$$X \sim \pi_{\text{prior}}(x) \propto \exp(-V(x)).$$

To apply the EM algorithm, we need to decide how to define the "missing data." Photon counter devices detect the emitted photons added over the line of sight; evidently, the problem would be more tractable if we knew the number of emitted photons from each pixel separately. Therefore, we define a fictitious measurement,

$$Z_{ij} \sim \text{Poisson}(a_{ij} x_j),$$

and posit that these variables are mutually independent. Obviously, after the measurement $Y = y$, we have

$$\sum_{j=1}^{N} Z_{ij} = y_i. \tag{22}$$

To perform the E-step, assuming that x^k is given, consider first the conditional density $\pi^k(z) = \pi\left(z \mid x^k, y\right)$.

A basic result from probability theory states that if N independent random variables Λ_j are a priori Poisson distributed with respective means μ_j, and in addition

$$\sum_{j=1}^{N} \Lambda_j = K,$$

then, a posteriori, the variables Λ_j conditioned on the above data are binomially distributed,

$$\Lambda_j \mid \left(\sum_{j=1}^{N} \Lambda_j = K \right) \sim \text{Binom}\left(K, \frac{\mu_j}{\sum_{j=1}^{N} \mu_j} \right).$$

In particular, the conditional expectation of Λ_j is

$$\mathsf{E}\left\{ \Lambda_j \mid \sum_{j=1}^{N} \Lambda_j = K \right\} = K \frac{\mu_j}{\sum_{j=1}^{N} \mu_j}.$$

We therefore conclude that the conditional density $\pi^k(z)$ is a product of binomial distributions of Z_{ij} with a priori means $\mu_j = a_{ij} x_j^k$, $\sum_{j=1}^{N} \mu_j = (\mathsf{A}x^k)_i$, and $K = y_i$, so in particular,

$$\mathsf{E}\left\{ Z_{ij} \mid \sum_{j=1}^{N} Z_{ij} = y_i \right\} = \int z_{ij} \pi^k(z) dz = y_i \frac{a_{ij} x_j^k}{(\mathsf{A}x^k)_i} \stackrel{\text{def}}{=} z_{ij}^k. \tag{23}$$

Furthermore, by Bayes' theorem,

$$\pi(x \mid y, z) = \pi(x \mid z) = \pi(z \mid x)\pi_{\text{prior}}(x),$$

where we used the fact that the true observations y add no information on x that would not be included in z, we have, by definition of the Poisson likelihood and the prior,

$$L(x \mid y, z) = \sum_{ij} \big(z_{ij} \log(a_{ij}x_j) - a_{ij}x_j\big) - V(x) + \text{constant},$$

and therefore, up to an additive constant, we have

$$Q^k(x) = \sum_{ij} \Big(z_{ij}^k \log(a_{ij}x_j) - a_{ij}x_j\Big) - V(x),$$

where z_{ij}^k is defined in (23). This completes the E-step.

The M-step requires the minimization of $Q^k(x)$ given above. Assuming that V is differentiable, the minimizer should satisfy

$$\frac{1}{x_\ell}\sum_{i=1}^m z_{i\ell}^k - \sum_{i=1}^m a_{i\ell} - \frac{\partial V}{\partial x_\ell}(x) = 0.$$

How complicated it is to find a solution to this condition depends on the prior contribution V and may require an internal Newton iteration. In [29], an approximate "one-step late" (OSL) algorithm was suggested, which is tantamount to a fixed-point iteration: Initiating with $\tilde{x}^0 = x^k$, an update scheme $\tilde{x}^t \to \tilde{x}^{t+1}$ is given by

$$\tilde{x}_\ell^{t+1} = \frac{\sum_{i=1}^m z_{i\ell}^k}{\sum_{i=1}^m a_{i\ell} + \frac{\partial V}{\partial x_\ell}(\tilde{x}^t)},$$

and this step is repeated until a convergence criterion is satisfied at some $t = t*$. Finally, the M-step is completed by updating $x^{k+1} = \tilde{x}^{t*}$.

The EM algorithm has been applied to other imaging problems such as blind deconvolution problem [44] and PET imaging [36, 71].

Markov Chain Monte Carlo Sampling

In Bayesian statistical imaging, the real solution of the imaging problem is the posterior density of the image interpreted as a multivariate random variable. If a closed form of the posterior is either unavailable or not suitable for the tasks at hand, the alternative is to resort to exploring the density by generating a representative sample from it. Markov chain Monte Carlo (MCMC) samplers yield samples from a target distribution by moving from a point in a chain to the next by the transition

rule which characterizes the specific algorithm. MCMC sampling algorithms are usually subdivided into those which are variants of the Metropolis–Hastings (MH) algorithm or the Gibbs sampler. While the foundations of the MH algorithm were laid first [25, 35, 51], Gibbs samplers have sometimes the appeal of being more straightforward to implement.

The basic idea of Monte Carlo integration is rather simple. Assume that $\pi(x)$ is a probability density in $\mathbb{R}^N$, and let $\{X^1, X^2, X^3, \ldots\}$ denote a stochastic process, where the random variables X^i are independent and identically distributed, $X^i \sim \pi(x)$. The central limit theorem asserts that for any measurable $f : \mathbb{R}^N \to \mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X^i) \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}^N} f(x)\pi(x)dx \quad \text{almost certainly,} \tag{24}$$

and moreover, the convergence takes place asymptotically with the rate $1/\sqrt{n}$, independently of the dimension N. The difficulty is to find a computationally efficient way of drawing independently from a given distribution π. Indeed, when N is large, it may be even difficult to decide where the numerical support of the density is. In MCMC methods, instead of producing an independent chain, the idea is to produce a Markov process $\{X^i\}$ with the property that π is the equilibrium distribution. It can be shown (see [53, 61, 65]) that with rather mild assumptions (irreducibility, aperiodicity), the limit (24) holds, due to the law of large numbers.

In applications to imaging, the computational burden associated with MCMC methods has become proverbial and is often presented as the main obstacle to the use of Bayesian method in imaging. It is easy to imagine that sampling random variable with hundreds of thousands of components will require a large amount of computer resources and that collecting and storing a large number of images will require much more time than estimating a single one. On the other hand, since an ensemble of images from a distribution carries a lot of additional information which cannot be included in single-point estimates, it seems unreasonable to rate methods simply according to computational speed. That said, since collecting a well-mixed, representative sample poses several challenges, in the description of the Gibbs sampling and Metropolis–Hastings algorithms, we will point out references to variants which can improve the independence and mixing of the ensemble; see [30–32].

In its first prominent appearance in the imaging arena [26], the Gibbs sampler was presented as part of a stochastic relaxation algorithm to efficiently compute MAP estimates. The systematic or fully conditional Gibbs sampler algorithm proceeds as follows [61].

Let $\pi(x)$ be a probability density defined on $\mathbb{R}^N$, denoted by $\pi(x) = \pi(x_1, \ldots, x_N)$, $x \in \mathbb{R}^N$ to underline that it is the joint density of the components of X. Furthermore, denote by $\pi(x_j \mid x_{-j})$ the conditional density of the jth component x_j given all the other components, collected in the vector $x_{-j} \in \mathbb{R}^{N-1}$. Let x^1 be the initial element of the Markov chain. Assuming that we are at a point x^i in the chain, we need a rule stating how to proceed to the next point x^{i+1}, i.e.,

we need to describe the updating method of proceeding from the current element x^i to x^{i+1}. This is done by updating sequentially each component as follows.

Fully conditional Gibbs sampling update: Given x^i, compute the next element x^{i+1} by the following algorithm:

$$\begin{aligned}
&\text{draw} \quad x_1^{i+1} \quad \text{from} \quad \pi\left(x_1 \mid x_{-1}^i\right);\\
&\text{draw} \quad x_2^{i+1} \quad \text{from} \quad \pi\left(x_2 \mid x_1^{i+1}, x_3^i, \ldots, x_N^i\right);\\
&\text{draw} \quad x_3^{i+1} \quad \text{from} \quad \pi\left(x_3 \mid x_1^{i+1}, x_2^{i+1}, x_4^i \ldots, x_N^i\right);\\
&\qquad \vdots\\
&\text{draw} \quad x_N^{i+1} \quad \text{from} \quad \pi\left(x_N \mid x_{-N}^{i+1}\right).
\end{aligned}$$

In imaging applications, this Gibbs sampler may be impractical because of the large number of components of the random variable to be updated to generate a new element of the chain. In addition, if some of the components are correlated, updating them independently may slow down the chain to explore the full support of the distribution, due to slow movement at each step. The correlation among components can be addressed by updating blocks of correlated components together, although this will imply that the draws must be from multivariate instead of univariate conditional densities.

It follows naturally from the updating scheme that the speed at which the chain will reach equilibrium is strongly dependent on how the system of coordinate axes relates to the most prominent correlation directions. A modification of the Gibbs sampler that can ameliorate the problems caused by correlated components performs a linear transformation of the random variable using correlation information. Without going into details, we refer to [48, 58, 61] for different variants of Gibbs sampler.

The strategy behind the Metropolis–Hastings samplers is to generate a chain with the target density as equilibrium distribution by constructing at each step the transition probability function from the current $X^i = x$ to next realization of X^{i+1} in the chain in the following way. Given an initial transition probability function $q(x, x')$ with $X^i = x$, x' drawn from $q(x, x')$ is a *proposal* for the value of X^{i+1}. Upon acceptance of $X^{i+1} = x'$, which occurs with probability $\alpha(x, x')$, defined by

$$\alpha(x, x') = \min\left\{\frac{\pi(x')q(x', x)}{\pi(x, x')}, 1\right\}, \quad \pi(x)q(x, x') > 0.$$

We add it to the chain; otherwise, we reject the proposed value and we set $X^{i+1} = x$. In the latter case, the chain did not move and the value x is replicated in the chain. The transition probability $p(x, x')$ of the Markov chain thus defined is

$$p(x, x') = q(x, x')\alpha(x, x'),$$

while the probability to stay put is

$$1 - \int_{\mathbb{R}^N} q(x, y)\alpha(x, y)dy.$$

This construction guarantees that the transition probability satisfies the detailed balance equation $\pi(x)p(x, x') = \pi(x')p(x', x)$, from which it follows that, for reasonable choices of the function q, $\pi(x)$ is the equilibrium distribution of the chain.

This algorithm is particularly convenient when the target distribution $\pi(x)$ is a posterior. In fact, since the only way in which π enters is via the ratio of its values at two points, it is sufficient to compute the density modulo a proportionality constant, which is how we usually define the posterior. Specific variants of the MH algorithm correspond to different choices of $q(x, x')$; in the original formulation [51], a symmetric proposal, for example, a random walk, was used, so that $q(x, x') = q(x', x)$, implying that

$$\alpha(x, x') = \min\{\pi(x')/\pi(x), 1\},$$

while the general formulation above is due to Hastings [35]. An overview of the different possible choices for $q(x, x')$ can be found in [65].

A number of hybrid sampling schemes which combine different chains or use MH variants to draw from the conditional densities inside Gibbs samplers have been proposed in the literature; see [48, 61] and references therein. Since the design of efficient MCMC samplers must address the specific characteristics of the target distribution, it is to be expected that as the use of densities becomes more pervasive in imaging, new hybrid MCMC scheme will be proposed.

The convergence of Monte Carlo integration based on MCMC methods is a key factor in deciding when to stop sampling. This is particularly pertinent in imaging applications, where the calculations needed for additions of a point to the chain may be quite time consuming. Due to the lack of a systematic way of translating theoretical convergence results of MCMC chains [7, 65] into pragmatic stopping rules, in practice, the issue is reduced to monitoring the behavior of the already collected sample.

As already pointed out, MCMC algorithms are not sampling independently from the posterior. When computing sample-based estimates for the posterior mean and covariance,

$$\hat{x}_{\mathrm{CM}} = \frac{1}{n}\sum_{j=1}^{n} x^j, \quad \hat{\Gamma}_{\mathrm{CM}} = \frac{1}{n}\sum_{j=1}^{n} \left(x^j - \hat{x}_{\mathrm{CM}}\right)\left(x^j - \hat{x}_{\mathrm{CM}}\right)^{\mathsf{T}}.$$

A crucial question is how accurately these estimates approximate the posterior mean and covariance. The answer depends on the sample size n and the sampling strategy itself. Ideally, if the sample vectors x^j are realizations of independent identically

distributed random variables, the approximations converge with the asymptotic rate $1/\sqrt{n}$, in agreement with the central limit theorem. In practice, however, the MCMC sampling produces sample points that are mutually correlated, and the convergence is slower.

The convergence of the chain can be investigated using the *autocovariance function* (ACF) of the sample [27, 64]. Assume that we are primarily interested in estimating a real-valued function $f : \mathbb{R}^N \to \mathbb{R}$ of the unknown, and we have generated an MCMC sample or a realization $\{x^1, \ldots, x^n\}$ of a stationary stochastic process $\{X^1, \ldots, X^n\}$. The random variables X^j are equally distributed, their distribution being the posterior distribution $\pi(x)$ of a random variable X. The estimation of the mean quantity $f(X)$ can be done by calculating

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} f(x^j),$$

while the theoretical mean of $f(X)$ is

$$\mu = \mathsf{E}\{f(X)\} = \int f(x)\pi(x)dx.$$

Each sample yields a slightly different value for $\hat{\mu}$, which is itself a realization of the random variable F defined as

$$F = \frac{1}{n} \sum_{j=1}^{n} f(X^j).$$

The problem is now how to estimate the variance of F, which gives us an indication of how well the computed realization approximates the mean. The identical distribution of the random variables X^j implies that

$$\mathsf{E}\{F\} = \frac{1}{n} \sum_{j=1}^{n} \underbrace{\mathsf{E}\{f(X^j)\}}_{=\mu} = \mu,$$

while the variance of F, which we want to estimate starting from the available realization of the by stochastic process, is

$$\mathrm{var}(F) = \mathsf{E}\{F^2\} - \mu^2.$$

To this end, we need to introduce some definitions and notations.

We define the autocovariance function of the stochastic process $f(X^j)$ with lag $k \geq 0$ to be

$$C(k) = \mathsf{E}\{f(X^j) f(X^{j+k})\} - \mu^2$$

which, if the process is stationary, is independent of j. The normalized ACF is defined as

$$c(k) = \frac{C(k)}{C(0)}.$$

The ACF can be estimated from an available realization as follows

$$\hat{C}(k) = \frac{1}{n-k} \sum_{j=1}^{n-k} f(x^j) f(x^{j+k}) - \hat{\mu}^2. \tag{25}$$

It follows from the definition of F that

$$\mathsf{E}\left\{F^2\right\} = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathsf{E}\left\{f(X^i) f(X^j)\right\}.$$

Let us now focus on the random matrix $\left[f(X^i) f(X^j)\right]_{i,j=1}^{n}$. The formula above takes its expectation and subsequently computes the average of its entries. By stationarity, the expectation is a symmetric Toeplitz matrix; hence, its diagonal entries are all equal to

$$\mathsf{E}\left\{f(X^i) f(X^i)\right\} = C(0) + \mu^2,$$

while the kth subdiagonal entries are all equal to

$$\mathsf{E}\left\{f(X^i) f(X^{i+k})\right\} = C(k) + \mu^2.$$

This observation provides us with a simple way to perform the summation by accounting for the elements along the diagonals, leading to the formula

$$\mathsf{E}\left\{F^2\right\} = \frac{1}{n^2} \left(nC(0) + 2 \sum_{k=1}^{n-1} (n-k) C(k) \right) + \mu^2,$$

from which it follows that the variance of F is

$$\operatorname{var}(F) = \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) C(k) \right).$$

If we assume that the ACF is negligible when $k > n_0$, for some n_0 significantly smaller than the sample size n, we may use the approximation

$$\operatorname{var}(F) \approx \frac{1}{n} \left(C(0) + 2 \sum_{k=1}^{n_0} C(k) \right) = \frac{C(0)}{n} \tau,$$

where

$$\tau = 1 + 2\sum_{k=1}^{n_0} c(k). \tag{26}$$

If we account fully for all contributions,

$$\tau = 1 + 2\sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k), \tag{27}$$

which is the Cesàro mean of the normalized ACFs or low-pass filtered mean with the triangular filter. The quantity τ is called the *integrated autocorrelation time* (IACT) and can be interpreted as the time that it takes for our MCMC to produce an independent sample. If the convergence rate for independence samplers is $1/\sqrt{n}$, the convergence rate for the MCMC sampler is $1/\sqrt{n/\tau}$. If the variables X_j are independent, then $\tau = 1$, and the result is exactly what we would expect from the central limit theorem, because in this case, $C(0) = n\,\mathrm{var}(f(X))$.

The estimate of τ requires an estimate for the normalized ACF, which can be obtained with the formula (25), and a value for n_0 to use in formula (26). In the choice of n_0, it is important to remember that $\hat{C}(k)$ is a realization of a random sequence $C(k)$, which in practice contains noise. Some practical rules for choosing n_0 are suggested in [27].

In [27], it is shown that since the sequence

$$\gamma(k) = c(2k) + c(2k+1), \quad k = 0, 1, 2, \ldots$$

is *strictly positive*, *strictly decreasing*, and *strictly convex*, that is,

$$\gamma(k) > 0, \quad \gamma(k+1) < \gamma(k), \quad \gamma(k+1) < \frac{1}{2}(\gamma(k) + \gamma(k+2)),$$

when the sample-based estimated sequence,

$$\hat{\gamma}(k) = \hat{c}(2k) + \hat{c}(2k+1), \quad k = 0, 1, 2, \ldots$$

fails to be so, this is an indication that the contribution is predominantly coming from noise; hence, it is wise to stop summing the terms to estimate τ. Geyer proposes three initial sequence estimators, in the following order:

1. Initial positive sequence estimator (IPSE): Choose n_0 to be the largest integer for which the sequence remains positive,

$$n_0 = n_{\mathrm{IPSE}} = \max\{k \mid \gamma(k) > 0\}.$$

2. Initial monotone sequence estimator (IMSE): Choose n_0 to be the largest integer for which the sequence remains positive and monotone,

$$n_0 = n_{\text{IMSE}} = \max\{k \mid \gamma(k) > 0,\ \gamma(k) < \gamma(k-1)\}.$$

3. Initial convex sequence estimator (ICSE): Choose n_0 to be the largest integer for which the sequence remains positive, monotone, and convex,

$$n_0 = n_{\text{ICSE}} = \max\left\{k \mid \gamma(k) > 0,\ \gamma(k) < \gamma(k-1),\ \gamma(k-1) < \frac{1}{2}(\gamma(k) + \gamma(k-2))\right\}.$$

From the proof in [27], it is obvious that also the sequence $\{c(k)\}$ itself must be positive and decreasing. Therefore, to find n_0 for IPSE or IMSE, there is no need for passing to the sequence $\{\gamma(k)\}$. As for ICSE, again from the proof in the cited article, it is also clear that the sequence

$$\eta(k) = c(2k+1) + c(2k+2), \quad k = 0, 1, 2, \ldots$$

too, is positive, monotonous, and convex. Therefore, to check the condition for ICSE, it might be advisable to form both sequences $\{\gamma(k)\}$ and $\{\eta(k)\}$ and set n_{ICSE} equal to the maximum index for which both $\gamma(k)$ and $\eta(k)$ remain strictly convex.

Summarizing a practical rule, using for instance, the IMSE, to compute τ is:

1. Estimate the ACF sequence $\hat{C}(k)$ from the sample by formula (25) and normalize it by $\hat{C}(0)$ to obtain $\hat{c}(k)$.
2. Find n_0 equal to the largest integer for which the sequence $\hat{c}(0), \hat{c}(1), \ldots, \hat{c}(n_0)$ remains positive and strictly decreasing. Notice that the computation of ACFs can be stopped when such an n_0 is reached.
3. Calculate the estimate for the IACT τ,

$$\tau = 1 + 2\sum_{k=1}^{n_0}\left(1 - \frac{k}{n}\right)c(k) \approx 1 + 2\sum_{k=1}^{n_0} c(k). \tag{28}$$

Notice that if n is not much larger than n_0, the sample is too small.

The accuracy of the approximation of μ by $\hat{\mu}$ is often expressed, with some degree of imprecision, by writing an estimate

$$\mu = \hat{\mu} \pm 2\left(\frac{C(0)}{n}\tau\right)^{1/2}$$

with the 95 % belief. This interpretation is based on the fact that, with a probability of about 95 %, the values of a Gaussian random variable are within ± 2 STD from

the mean. Such an approximate claim is justified when n is large, in which case the random variable F is asymptotically Gaussian by the central limit theorem.

Statistical Approach: What Is the Gain?

Statistical methods are often pitted against deterministic ones, and the true gain of the approach is sometimes lost, especially if the statistical methods are used only to produce single estimates. Indeed, it is not uncommon that the statistical framework is seen simply as an alternative way of explaining regularization. Another criticism of statistical methods concerns the computation times. While there is no doubt that computing a posterior mean using MCMC methods is more computationally intensive than resorting to optimization-based estimators, it is also obvious that a comparison in these terms does not make much sense, since a sample contains enormously more information of the underlying distribution than an estimate of its mode.

To emphasize what there is to be gained when using the statistical approach, we consider some algorithms that have been found useful and are based on the interpretation images as random variables.

Beyond the Traditional Concept of Noise

The range of interpretation of the concept of noise in imaging is usually very restricted, almost exclusively referring to uncertainties in observed data due to exogenous sources. In the context of deterministic regularization, the noise model is almost always additive, in agreement with the paradigm that only acknowledges noise as the difference between a "true" and "noisy" data, giving no consideration to its statistical properties. Already the proper noise modeling of counting data clearly demonstrates the shortcomings of such models. The Bayesian – or subjective – use of probability as an expression of uncertainty allows to extend the concept of noise to encompass a much richer terrain of phenomena, including shortcomings in the forward model, prior, or noise statistics itself.

To demonstrate the possibilities of the Bayesian modeling, consider an example where it is assumed that a forward model with additive noise,

$$y = F(x) + e. \tag{29}$$

which describes, to the best of our knowledge, as completely as possible, the interdependency of the data y and the unknown. We refer to it as the *detailed model*. Here, the noise e is thought to be exogenous, and its statistical properties are known.

Assume further that the detailed model is computationally too complex to be used with the imaging algorithms and the application at hand for one or several of the following reasons. The dimensionality of the image x may be too high for the model to be practical; the model may contain details such as boundary conditions that need to be simplified in practice; the deblurring kernel may be non-separable, while in practice, a fast algorithm for separable kernels may exist. To

accommodate these difficulties, a simpler model is constructed. Let z be possibly a simpler representation of x, obtained, for example, via a projection to a coarser grid, and let f denote the corresponding forward map. It is a common procedure to write a simplified model of the form

$$y = f(z) + e, \tag{30}$$

which, however, may not explain the data as well as the detailed model (29). To properly account for the errors added by the model reduction, we should write instead

$$\begin{aligned} y &= F(x) + e = f(z) + [F(x) - f(z)] + e \\ &= f(z) + \varepsilon(x, z) + e, \quad \varepsilon(x, z) = F(x) - f(z), \end{aligned} \tag{31}$$

where the term $\varepsilon(x, z)$ is referred to as *modeling error*.

In the framework of deterministic imaging, modeling errors pose unsurmountable problems because they depend on both the unknown image x and its reduced counterpart z. A common way to address errors coming from model reduction is to artificially increase the variance of the noise included in the reduced model until it masks the modeling error. Such an approach introduces a statistical structure in the noise that does not correspond to the modeling error and may easily waste several orders of magnitude of the accuracy of the data. On the other hand, neglecting the error introduced by model reduction may lead to overly optimistic estimates of the performance of algorithms. The very questionable procedure of testing algorithms with data simulated with the same forward map used for the inversion is referred to as *inverse crime* [42]. Inverse criminals, who tacitly assume that $\varepsilon(x, z) = 0$, should not be surprised if the unrealistically good results obtained from simulated data are not robust when using real data.

While modeling error often is neglected also in the statistical framework, its statistical properties can be described in terms of the prior. Consider the stochastic extension of $\varepsilon(x, z)$,

$$\tilde{E} = \varepsilon(X, Z),$$

where X and Z are the stochastic extensions of x and z, respectively. Since, unlike an exogenous noise term, the modeling error is not independent of the unknowns Z and X, the likelihood and the prior cannot be described separately, but instead must be specified together.

To illustrate how ubiquitous modeling error is, consider the following example.

Boundary clutter and image truncation: Consider a denoising/deblurring example of the type encountered in astronomy, microscopy, and image processing. Let $u : \mathbb{R}^2 \to \mathbb{R}$ be a continuous two-dimensional model of a scenery that is recorded through an out-of-focus device. The noiseless model for the continuous problem is a convolution integral,

$$v(r) = \int_{\mathbb{R}^2} a(r-s)u(s)ds, \quad r \in \mathbb{R}^2,$$

the convolution kernel $a(r-s)$ describing the point spread of the device. We assume that $r \mapsto a(r)$ decays rapidly enough to justify an approximation as a compactly supported function.

Let $Q \subset \mathbb{R}^2$ define a bounded *field of view*. We consider the following imaging problem: *Given a noisy version of the blurred image v over the field of view Q, estimate the underlying image u over the field of view Q.*

Assume that a sufficiently fine discretization of Q into N pixels is given, and denote by $r_i \in Q$ the center of the ith pixel. Assume further that the point spread function a is negligibly small outside a disc D of radius $\delta > 0$. By selecting an *extended field of view* Q' such that

$$Q + D = \{s \in \mathbb{R}^2 \mid s = r + r',\ r \in Q,\ r' \in D\} \subset Q',$$

we may restrict the domain of integration in the definition of the convolution integral

$$v(r_i) = \int_{\mathbb{R}^2} a(r_i - s)u(s)ds \approx \int_{Q'} a(r_i - s)u(s)ds.$$

After discretizing Q' into N' pixels p_j with center points s_j, N of which are within the field of view, coinciding with R^J we can restate the problem in the form

$$\begin{aligned} v(s_i) &\approx \int_{Q'} a(s_i - s)u(s)ds \approx \sum_{j=1}^{N'} |p_j| a(s_i - s_j)u(s_j) \\ &= a_{ij}u(s_j), \quad a_{ij} = |p_j| a(s_i - s_j), \quad 1 \le i \le N. \end{aligned}$$

After accounting for the contribution of exogenous noise at each recorded pixel, we arrive at the complete discrete model

$$y = \mathsf{A}'x + e, \quad \mathsf{A}' \in \mathbb{R}^{N \times N'}, \tag{32}$$

where $x_j = u(s_j)$ and y_i represent the noisy observation of $v(s_i)$. If the pixelization is fine enough, we may consider this model to be a good approximation of the continuous problem.

A word of caution is in order when using this model, because the right-hand side depends not only on pixels within the field of view, where we want to estimate the underlying image, but also on pixels in the frame $C = Q' \setminus Q$ around it. The vector x is therefore partitioned into two vectors, where the first one, denoted by $z \in \mathbb{R}^N$, contains values in the pixels within the field of view, and the second one, $\zeta \in \mathbb{R}^K$, $K = N' - N$, consists of values of pixels in the frame. After suitably rearranging the indices, we may write x in the form

$$x = \begin{bmatrix} z \\ \zeta \end{bmatrix} \begin{matrix} \mathbb{R}^N \\ \in \times \\ \mathbb{R}^K \end{matrix},$$

and, after partitioning the matrix A' accordingly,

$$\mathsf{A}' = \begin{bmatrix} \mathsf{A} & \mathsf{B} \end{bmatrix} \in \mathbb{R}^{N\times N} \times \mathbb{R}^{N\times K},$$

we can rewrite the model (32) in the form

$$y = \mathsf{A}z + \mathsf{B}\zeta + e = \mathsf{A}z + \varepsilon + e,$$

where the modeling errors are collected in second term ε, which we will refer to as *boundary clutter*. It is well known that ignoring the contribution to the recorded image coming from and beyond the boundary may cause severe artifacts in the estimation of the image x within the field of view. In a determinist framework, the boundary clutter term is often compensated for by extending the image outside the field of view in a manner believed to be closest to the actual image behavior. Periodic extension or extensions obtained by reflecting the image symmetrically or antisymmetrically are quite popular in the literature, because they will significantly simplify the computations; details on such an approach can be found, for example, in [21].

Consider a Gaussian prior and a Gaussian likelihood,

$$X \sim \mathcal{N}(0, \Gamma), \quad E \sim \mathcal{N}(0, \Sigma_{\text{noise}}),$$

and partition the prior covariance matrix according to the partitioning of x,

$$\Gamma \in \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}, \quad \Gamma_{11} \in \mathbb{R}^{N\times N},\ \Gamma_{12} = \Gamma_{21}^{\mathsf{T}} \in \mathbb{R}^{N\times K},\ \Gamma_{22} \in \mathbb{R}^{K\times K}.$$

The covariance matrix of the total noise term, which also includes the boundary clutter $\tilde{E}$, is

$$\mathsf{E}\left\{\left(\tilde{E} + E\right)\left(\tilde{E} + E\right)^{\mathsf{T}}\right\} = \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\text{noise}} = \Sigma$$

and the cross covariance of the image within the field of view and the noise is

$$\mathsf{C} = \mathsf{E}\left\{Z\left(\tilde{E} + E\right)^{\mathsf{T}}\right\} = \Gamma_{12}\mathsf{B}^{\mathsf{T}}.$$

The posterior distribution of the vector Z conditioned on $Y = y$ now follows from (17) and (18). The posterior mean is

$$z_{\mathrm{CM}} = \left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)\left(\mathsf{A}\Gamma_{11}\mathsf{A}^{\mathsf{T}} + \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\mathrm{noise}}\right)^{-1} y,$$

and the posterior covariance is

$$\Gamma_{\mathrm{post}} = \Gamma_{11} - \left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)\left(\mathsf{A}\Gamma_{11}\mathsf{A}^{\mathsf{T}} + \mathsf{B}\Gamma_{22}\mathsf{B}^{\mathsf{T}} + \Sigma_{\mathrm{noise}}\right)^{-1}\left(\Gamma_{11}\mathsf{A} + \Gamma_{12}\mathsf{B}^{\mathsf{T}}\right)^{\mathsf{T}}.$$

A computationally efficient and robust algorithm for computing the conditional mean is proposed in [13]. For further applications of the modeling error approach in imaging, see [1, 38, 47].

Sparsity and Hypermodels

The problem of reconstructing sparse images or more generally images that can be represented as sparse linear combinations of prescribed basis images using data consisting of few measurements has recently received a lot of attention and has become a central issue in compressed sensing [11]. Bayesian hypermodels provide a very natural framework for deriving algorithms for sparse reconstruction.

Consider a linear model with additive Gaussian noise, the likelihood being given by (12) and a conditionally Gaussian prior (13) with hyperparameter θ. As explained in section "Adding Layers: Hierarchical Models," if we select the hyperprior $\pi_{\mathrm{hyper}}(\theta)$ in such a way that it favors solutions with variances Θ_j close to zero except for only few outliers, the overall prior for (X, Θ) will be biased toward sparse solutions. Two hyperpriors well suited for sparse solutions are the gamma and the inverse gamma hyperpriors. For the sake of definiteness, consider the inverse gamma hyperprior with mutually independent components,

$$\pi_{\mathrm{hyper}}(\theta_j) = \theta_j^{-k-1}\exp\left(-\frac{\theta_0}{\theta_j}\right) = \exp\left(-\frac{\theta_0}{\theta_j} - (k+1)\log\theta_j\right).$$

Then the posterior distribution for the pair (X, Θ) is of the form

$$\pi(x, \theta \mid y) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(y - \mathsf{A}x) - \frac{1}{2}x^{\mathsf{T}}\mathsf{D}_\theta^{-1}x - \sum_{j=1}^{N} V(\theta_j)\right)$$

where

$$V(\theta_j) = \frac{\theta_0}{\theta_j} + \left(k + \frac{3}{2}\right)\log\theta_j, \quad \mathsf{D}_\theta = \mathrm{diag}(\theta) \in \mathbb{R}^{N\times N}.$$

An estimate for (X, Θ) can be found by maximizing $\pi(x, \theta \mid y)$ with respect to the pair (x, θ) using, for example, a quasi-Newton optimization scheme. Alternatively, the following two algorithms that make use of the special form of the expression above can also be used.

In the articles [66, 67] on Bayesian machine learning, the starting point is the observation that the posterior density $x \mapsto \pi(x, \theta \mid y)$ is Gaussian and therefore it is possible to integrate it explicitly with respect to x. It can be shown, after some tedious but straightforward algebraic manipulations, that the marginal posterior distribution is

$$\begin{aligned}\pi(\theta \mid y) &= \int_{\mathbb{R}^N} \pi(x, \theta \mid y) dx \\ &\propto \left(\frac{1}{\det(\mathsf{M}_\theta)}\right)^{1/2} \exp\left(-\sum_{j=1}^{N} V(\theta_j) + \frac{1}{2}\tilde{y}^{\mathsf{T}} \mathsf{M}_\theta^{-1} \tilde{y}\right),\end{aligned}$$

where

$$\mathsf{M}_\theta = \mathsf{A}^{\mathsf{T}} \Sigma^{-1} \mathsf{A} + \mathsf{D}_\theta^{-1}, \quad \tilde{y} = \mathsf{A}^{\mathsf{T}} \Sigma^{-1} y.$$

The *most probable* estimate or the *maximum evidence* estimator $\hat{\theta}$ of Θ is, by definition, the maximizer of the above marginal, or equivalently, the maximizer of its logarithm,

$$L(\theta) = -\frac{1}{2} \log\left(\det(\mathsf{M}_\theta)\right) - \sum_{j=1}^{N} V(\theta_j) + \frac{1}{2}\tilde{y}^{\mathsf{T}} \mathsf{M}_\theta^{-1} \tilde{y}$$

which must satisfy

$$\frac{\partial L}{\partial \theta_j} = 0, \quad 1 \leq j \leq N.$$

It turns out that, although the computation of the determinant may in general be a challenge, its derivatives can be expressed in a formally simple form. To this end separate the element depending on θ_j from D_θ^{-1}, writing

$$\mathsf{D}_\theta^{-1} = \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} + \mathsf{D}_{\theta'}^{\dagger},$$

where e_j is the jth coordinate unit vector, θ' is the vector θ with the jth element replaced by a zero and "†" denotes the pseudo-inverse. Then

$$\begin{aligned}\mathsf{M}_\theta &= \mathsf{A}^{\mathsf{T}} \Sigma^{-1} \mathsf{A} + \mathsf{D}_{\theta'}^{\dagger} + \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} = \mathsf{M}_{\theta'} + \frac{1}{\theta_j} e_j e_j^{\mathsf{T}} \\ &= \mathsf{M}_{\theta'} \left(\mathsf{I} + \frac{1}{\theta_j} q e_j^{\mathsf{T}}\right), \quad q = \mathsf{M}_{\theta'}^{-1} e_j.\end{aligned} \tag{33}$$

It follows from the properties of the determinant that

$$\det(\mathsf{M}_\theta) = \det\left(\mathsf{I} + \frac{1}{\theta_j} q e_j^{\mathsf{T}}\right) \det(\mathsf{M}_{\theta'}) = \left(1 + \frac{q_j}{\theta_j}\right) \det(\mathsf{M}_{\theta'}),$$

where $q_j = e_j^{\mathsf{T}} q$. After expressing the inverse of M_θ in the expression of $L(\theta)$ via the Sherman–Morrison–Woodbury formula [28] as

$$\mathsf{M}_\theta^{-1} = \mathsf{M}_{\theta'}^{-1} - \frac{1}{\theta_j + q_j} q q^{\mathsf{T}},$$

we find that the function $L(\theta)$ can be written as

$$L(\theta) = \frac{1}{2} \log\left(1 + \frac{q_j}{\theta_j}\right) - V(\theta_j) + \frac{1}{2} \frac{(q^{\mathsf{T}} \tilde{y})^2}{\theta_j + q_j} + \text{terms that are independent of } \theta_j.$$

The computation of the derivative of $L(\theta)$ with respect to θ_j and its zeros is now straightforward, although not without challenges because reevaluation of the vector q may potentially be expensive. For details, we refer to the article [67].

After having found an estimate $\hat{\theta}$, an estimate for X can be obtained by observing that the conditional density $\pi(x \mid y, \hat{\theta})$ is Gaussian,

$$\pi(x \mid y, \hat{\theta}) \propto \exp\left(-\frac{1}{2}(y - \mathsf{A}x)^{\mathsf{T}} \Sigma^{-1} (y - \mathsf{A}x) - \frac{1}{2} x^{\mathsf{T}} \hat{\theta} x\right),$$

and an estimate for x is obtained by solving in the least squares sense the linear system

$$\begin{bmatrix} \Sigma^{-1/2}\mathsf{A} \\ \mathsf{D}_{\hat{\theta}}^{-1/2} \end{bmatrix} x = \begin{bmatrix} \Sigma^{-1/2} y \\ 0 \end{bmatrix}. \tag{34}$$

In imaging applications, this is a large-scale linear problem and typically, iterative solvers need to be employed [59].

A different approach leading to a fast algorithm of estimating the MAP estimate $(x, \theta)_{\text{MAP}}$ was suggested in [15]. The idea is to maximize the posterior distribution using an alternating iteration: Starting with an initial value $\theta = \theta^1$, $\ell = 1$, the iteration proceeds as follows:

1. Find $x^{\ell+1}$ that maximizes $x \mapsto L(x, \theta^\ell) = \log\big(\pi(x, \theta^\ell \mid y)\big)$.
2. Update $\theta^{\ell+1}$ by maximizing $\theta \mapsto L(x^{\ell+1}, \theta) = \log\big(\pi(x^{\ell+1}, \theta \mid y)\big)$.

The efficiency of this algorithm is based on the fact that for $\theta = \theta^\ell$ fixed, the maximization of $L(x, \theta^\ell)$ in the first step is tantamount to minimizing the quadratic expression

$$\frac{1}{2}\|\Sigma^{-1/2}(y - \mathsf{A}x)\|^2 + \frac{1}{2}\left\|\mathsf{D}_{\theta^\ell}^{-1/2}x\right\|^2,$$

the non-quadratic part being independent of x. Thus, step 1 only requires an (approximate) linear least squares solution of the system similar to (34). On the other hand, when $x = x^{\ell+1}$ is fixed, the minimizer of the second step is found as a zero of the gradient of the function $L(x^{\ell+1}, \theta)$ with respect to θ. This step, too, is straightforward, since the component equations are mutually independent,

$$\frac{\partial}{\partial\theta_j}L(x^{\ell+1}, \theta) = -\left(\frac{1}{2}\left(x_j^{\ell+1}\right)^2 + \theta_0\right)\frac{1}{\theta_j^2} + \left(k + \frac{3}{2}\right)\frac{1}{\theta_j} = 0,$$

leading to the explicit updating formula

$$\theta_j^{\ell+1} = \frac{1}{2k+3}\left(\left(x_j^{\ell+1}\right)^2 + 2\theta_0\right).$$

For details and performance of the method in image applications, we refer to [15].

5 Conclusion

This chapter gives an overview of statistical methods in imaging. Acknowledging that it would be impossible to give a comprehensive review of all statistical methods in imaging in a chapter, we have put the emphasis on the Bayesian approach, while making repeated forays in the frequentists' field. These editorial choices are reflected in the list of references, which only covers a portion of the large body of literature published on the topic. The use of statistical methods in subproblems of imaging science is much wider than presented here, extending, for example, from image segmentation to feature extraction, interpretation of functional MRI signals, and radar imaging.

Cross-References

▸ EM Algorithms
▸ Iterative Solution Methods
▸ Linear Inverse Problems
▸ Total Variation in Imaging

References

1. Arridge, S.R., Kaipio, J.P., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T., Vauhkonen, M.: Approximation errors and model reduction with an application in optical diffusion tomography. Inverse. Probl. **22**, 175–195 (2006)
2. Bardsley, J., Vogel, C.R.: A nonnegatively constrained convex programming method for image reconstruction. SIAM J. Sci. Comput. **25**, 1326–1343 (2004)
3. Bernardo, J.: Bayesian Theory. Wiley, Chichester (2000)
4. Bertero, M., Boccacci, P.: Introduction to Inverse Problems in Imaging. IOP, Bristol (1998)
5. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. J. Stat. R. Soc. **36**, 192–236 (1974)
6. Besag, J.: On the statistical analysis of dirty pictures. J. R. Stat. Soc. B **48**, 259–302 (1986)
7. Besag, J., Green, P.: Spatial statistics and Bayesian computation. J. R. Stat. Soc. B **55**, 25–37 (1993)
8. Billingsley, P.: Probability and Measure, 3rd edn. Wiley, New York (1995)
9. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
10. Boyles, R.A.: On the convergence of the EM algorithm. J. R. Stat. Soc. B **45**, 47–50 (1983)
11. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34–81 (2009)
12. Calvetti, D.: Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. J. Comput. Appl. Math. **198**, 378–395 (2007)
13. Calvetti, D., Somersalo, E.: Statistical compensation of boundary clutter in image deblurring. Inverse Probl. **21**, 1697–1714 (2005)
14. Calvetti, D., Somersalo, E.: Introduction to Bayesian Scientific Computing – Ten Lectures on Subjective Probability. Springer, Berlin (2007)
15. Calvetti, D., Somersalo, E.: Hypermodels in the Bayesian imaging framework. Inverse Probl. **24**, 034013 (2008)
16. Calvetti, D., Hakula, H., Pursiainen, S., Somersalo, E.: Conditionally Gaussian hypermodels for cerebral source localization. SIAM J. Imaging Sci. **2**, 879–909 (2009)
17. Cramér, H.: Mathematical Methods in Statistics. Princeton University Press, Princeton (1946)
18. De Finetti, B.: Theory of Probability, vol 1. Wiley, New York (1974)
19. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)
20. Dennis, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. SIAM, Philadelphia (1996)
21. Donatelli, M., Martinelli, A., Serra-Capizzano, S.: Improved image deblurring with anti-reflective boundary conditions. Inverse Probl. **22**, 2035–2053 (2006)
22. Franklin, J.N.: Well-posed stochastic extension of ill-posed linear problem. J. Math. Anal. Appl. **31**, 682–856 (1970)
23. Fox, C., Nicholls, G.: Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. AIP Conf. Proc. ISSU **568**, 252–263 (2001)
24. Gantmacher, F.R.: Matrix Theory. AMS, New York (1990)
25. Gelfand, A.E., Smith, A.F.M.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)
26. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and Bayesian rerstoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
27. Geyer, C.: Practical Markov chain Monte Carlo. Stat. Sci. **7**, 473–511 (1992)
28. Golub, G., VanLoan, C.F.: Matrix Computations. Johns Hopkins University Press, London (1996)
29. Green, P.J.: Bayesian reconstructions from emission tomography data using modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
30. Green, P.J., Mira, A.: Delayed rejection in reversible jump Metropolis-Hastings. Biometrika **88**, 1035–1053 (2001)

31. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**, 223–242 (2001)
32. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: efficient adaptive MCMC. Stat. Comput. **16**, 339–354 (2006)
33. Hansen, P.C.: Rank-Deficient and Ill-Posed Inverse Problems. SIAM, Philadelphia (1998)
34. Hansen, P.C.: Discrete Inverse Problems. Insights and Algorithms. SIAM, Philadelphia (2010)
35. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
36. Herbert, T., Leahy, R.: A generalized EM algorithm for 3D Bayesian reconstruction from Poisson data using Gibbs priors. IEEE Trans. Med. Imaging **8**, 194–202 (1989)
37. Hestenes, M.R., Stiefel, E.: Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. **49**, 409–436 (1952)
38. Huttunen, J.M.J., Kaipio, J.P.: Model reduction in state identification problems with an application to determination of thermal parameters. Appl. Numer. Math. **59**, 877–890 (2009)
39. Jeffreys, H.: An invariant form for the prior probability in estimation problem. Proc. R. Soc. Lond. A **186**, 453–461 (1946)
40. Ji, S., Carin, L.: Bayesian compressive sensing and projection optimization. In: Proceedings of 24th International Conference on Machine Learning, Cornvallis (2007)
41. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, Berlin (2004)
42. Kaipio, J.P., Somersalo, E.: Statistical inverse problems: discretization, model reduction and inverse crimes. J. Comput. Appl. Math. **198**, 493–504 (2007)
43. Kelley, T.: Iterative Methods for Optimization. SIAM, Philadelphia (1999)
44. Lagendijk, R.L., Biemond, J.: Iterative Identification and Restoration of Images. Kluwer, Boston (1991)
45. Laksameethanasan, D., Brandt, S.S., Engelhardt, P., Renaud, O., Shorte, S.L.: A Bayesian reconstruction method for micro-rotation imaging in light microscopy. Microsc. Res. Tech. **71**, 158–167 (2007)
46. LeCam, L.: Asymptotic Methods in Statistical Decision Theory. Springer, New York (1986)
47. Lehikoinen, A., Finsterle, S., Voutilainen, A., Heikkinen, L.M., Vauhkonen, M., Kaipio, J.P.: Approximation errors and truncation of computational domains with application to geophysical tomography. Inverse Probl. Imaging **1**, 371–389 (2007)
48. Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer, Berlin (2003)
49. Lucy, L.B.: An iterative technique for the rectification of observed distributions. Astron. J. **79**, 745–754 (1974)
50. Melsa, J.L., Cohn, D.L.: Decision and Estimation Theory. McGraw-Hill, New York (1978)
51. Metropolis, N., Rosenbluth, A.W., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
52. Mugnier, L.M., Fusco, T., Conan, J.-L.: Mistral: a myopic edge-preserving image restoration method, with application to astronomical adptive-optics-corrected long-exposure images. J. Opt. Soc. Am. A **21**, 1841–1854 (2004)
53. Nummelin, E.: MC's for MCMC'ists. Int. Stat. Rev. **70**, 215–240 (2002)
54. Ollinger, J.M., Fessler, J.A.: Positron-emission tomography. IEEE Signal Proc. Mag. **14**, 43–55 (1997)
55. Paige, C.C., Saunders, M.A.: LSQR: an algorithm for sparse linear equations and sparse least squares. TOMS **8**, 43–71 (1982)
56. Paige, C.C., Saunders, M.A.: Algorithm 583; LSQR: sparse linear equations and least-squares problems. TOMS **8**, 195–209 (1982)
57. Richardson, H.W.: Bayesian-based iterative method of image restoration. J. Opt. Soc. Am. **62**, 55–59 (1972)
58. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004)
59. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia (2003)
60. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction in positron emission tomography. IEEE Trans. Med. Imaging **MI-1**, 113–122 (1982)

61. Smith, A.F.M., Roberts, R.O.: Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. B **55**, 3–23 (1993)
62. Snyder, D.L.: Random Point Processes. Wiley, New York (1975)
63. Starck, J.L., Pantin, E., Murtagh, F.: Deconvolution in astronomy: a review. Publ. Astron. Soc. Pac. **114**, 1051–1069 (2002)
64. Tan, S.M., Fox, C., Nicholls, G.K.: Lecture notes (unpublished), Chap 9. http://www.math.auckland.ac.nz/
65. Tierney, L.: Markov chains for exploring posterior distributions. Ann. Stat. **22**, 1701–1762 (1994)
66. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. **1**, 211–244 (2001)
67. Tipping, M.E., Faul, A.C.: Fast marginal likelihood maximisation for sparse Bayesian models. In: Proceedings of the 19th International Workshop on Artificial Intelligence and Statistics, Key West, 3–6 Jan 2003
68. Van Kempen, G.M.P., Van Vliet, L.J., Verveer, P.J.: A quantitative comparison of image restoration methods in confocal microscopy. J. Microsc. **185**, 354–365 (1997)
69. Wei, G.C.G., Tanner, M.A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algoritms. J. Am. Stat. Assoc. **85**, 699–704 (1990)
70. Wu, J.: On the convergence properties of the EM algorithm. Ann. Stat. **11**, 95–103 (1983)
71. Zhou, J., Coatrieux, J.-L., Bousse, A., Shu, H., Luo, L.: A Bayesian MAP-EM algorithm for PET image reconstruction using wavelet transform. Trans. Nucl. Sci. **54**, 1660–1669 (2007)