CWI Quarterly Volume 11 (4) 1998, pp. 413 – 437

Markov Random Fields and Images

Patrick Pérez
Irisa/Inria, Campus de Beaulieu,
35042 Rennes Cedex, France
e-mail: perez@irisa.fr

At the intersection of statistical physics and probability theory, Markov random fields and Gibbs distributions have emerged in the early eighties as powerful tools for modeling images and coping with high-dimensional inverse problems from low-level vision. Since then, they have been used in many studies from the image processing and computer vision community. A brief and simple introduction to the basics of the domain is proposed.

1. Introduction and general framework

With a seminal paper by Geman and Geman in 1984 [18], powerful tools known for long by physicists [2] and statisticians [3] were brought in a comprehensive and stimulating way to the knowledge of the image processing and computer vision community. Since then, their theoretical richness, their practical versatility, and a number of fruitful connections with other domains, have resulted in a profusion of studies. These studies deal either with the modeling of images (for synthesis, recognition or compression purposes) or with the resolution of various high-dimensional *inverse* problems from early vision (e.g., restoration, deblurring, classification, segmentation, data fusion, surface reconstruction, optical flow estimation, stereo matching, etc. See collections of examples in [11, 30, 40]).

The implicit assumption behind probabilistic approaches to image analysis is that, for a given problem, there exists a probability distribution that can capture to some extent the variability and the interactions of the different sets of relevant image attributes. Consequently, one considers the variables of the problem as *random variables* forming a set (or random vector) $X = (X_i)_{i=1}^n$ with joint probability distribution P_X [1].

[1] P_X is actually a probability mass in the case of discrete variables, and a probability density function when the X_i's are continuously valued. In the latter case, all summations over states or configurations should be replaced by integrals.

The first critical step toward probabilistic modeling thus obviously relies on the choice of the multivariate distribution P_X. Since there is so far no really generic theory for selecting a model, a tailor-made *parameterized* function P_X^θ is generally chosen among standard ones, based on intuition of the desirable properties [2].

The basic characteristic of chosen distributions is their decomposition as a product of factors depending on just a few variables (one or two in most cases). Also, a given distribution involves only a few types of factors.

One has simply to specify these *local* "interaction" factors (which might be complex, and might involve variables of different nature) to define, up to some multiplicative constant, the joint distribution $\mathsf{P}_X(x_1, \ldots, x_n)$: one ends up with a *global* model.

With such a setup, each variable only *directly* depends on a few other "neighboring" variables. From a more global point of view, all variables are mutually dependent, but only through the combination of successive local interactions. This key notion can be formalized considering the graph for which i and j are neighbors if x_i and x_j appear within a same local component of the chosen factorization. This graph turns out to be a powerful tool to account for local and global structural properties of the model, and to predict changes in these properties through various manipulations. From a probabilistic point of view, this graph neatly captures Markov-type conditional independencies among the random variables attached to its vertices.

After the specification of the model, one deals with its actual use for modeling a class of problems and for solving them. At that point, as we shall see, one of the two following things will have to be done: (1) drawing samples from the joint distribution, or from some conditional distribution deduced from the joint law when part of the variables are observed and thus fixed; (2) maximizing some distribution (P_X itself, or some conditional or marginal distribution deduced from it).

The very high dimensionality of image problems under concern usually excludes any direct method for performing both tasks. However the local decomposition of P_X fortunately allows to devise suitable deterministic or stochastic iterative algorithms, based on a common principle: at each step, just a few variables (often a single one) are considered, all the others being "frozen". Markovian properties then imply that the computations to be done remain local, that is, they only involve neighboring variables.

This paper is intended to give a brief (and definitely incomplete) overview of how Markovian models can be defined and manipulated in the prospect of modeling and analyzing images. Starting from the formalization of Markov random fields (MRFs) on graphs through the specification of a Gibbs distribution (§2), the standard issues of interest are then grossly reviewed: sampling from a high-dimensional Gibbs distribution (§3); learning models (at least parameters) from observed images (§4); using the Bayesian machinery to cope with

[2] Superscript θ denotes a parameter vector. Unless necessary, it will be dropped for notational convenience.

inverse problems, based on learned models (§5); estimating parameters with partial observations, especially in the case of inverse problems (§6). Finally two modeling issues (namely the introduction of so-called auxiliary variables, and the definition of hierarchical models), which are receiving a great deal of attention from the community at the moment, are evoked (§7).

2. Gibbs distribution and graphical Markov properties

Let us now make more formal acquaintance with Gibbs distributions and their Markov properties. Let X_i, $i = 1, \ldots, n$, be random variables taking values in some discrete or continuous *state space* Λ, and form the random vector $X = (X_1, \ldots, X_n)^T$ with *configuration* set $\Omega = \Lambda^n$. All sorts of state spaces are used in practice. More common examples are: $\Lambda = \{0, \ldots, 255\}$ for 8-bit quantized luminances; $\Lambda = \{1, \ldots, M\}$ for semantic "labelings" involving M classes; $\Lambda = \mathbb{R}$ for continuously-valued variables like luminance, depth, etc.; $\Lambda = \{-u_{\max}, \ldots, u_{\max}\} \times \{-v_{\max}, \ldots, v_{\max}\}$ in matching problems involving displacement vectors or stereo disparities for instance; $\Lambda = \mathbb{R}^2$ in vector field-based problems like optical flow estimation or shape-from-shading.

As said in the introduction, P_X exhibits a factorized form:

$$\mathsf{P}_X(x) \propto \prod_{c \in \mathcal{C}} f_c(x_c), \tag{1}$$

where $\mathcal{C}$ consists of small index subsets c, the factor f_c depends only on the variable subset $x_c = \{x_i, i \in c\}$, and $\prod_c f_c$ is summable over Ω. If, in addition, the product is positive ($\forall x \in \Omega, \mathsf{P}_X(x) > 0$), then it can be written in exponential form (letting $V_c = -\ln f_c$):

$$\mathsf{P}_X(x) = \frac{1}{Z} \exp\{-\sum_c V_c(x_c)\}. \tag{2}$$

Well known from physicists, this is the *Gibbs* (or Boltzman) *distribution* with *interaction potential* $\{V_c, c \in \mathcal{C}\}$, *energy* $U = \sum_c V_c$, and *partition function* (of parameters) $Z = \sum_{x \in \Omega} \exp\{-U(x)\}$ [3]. Configurations of lower energies are the more likely, whereas high energies correspond to low probabilities.

The interaction structure induced by the factorized form is conveniently captured by a graph that statisticians refer to as the *independence graph*: the independence graph associated with the factorization $\prod_{c \in \mathcal{C}} f_c$ is the simple undirected graph $\mathbb{G} = [S, E]$ with vertex set $S = \{1, \ldots, n\}$, and edge set E defined as: $\{i, j\} \in E \iff \exists c \in \mathcal{C} : \{i, j\} \subset c$, i.e., i and j are neighbors if x_i and x_j appear simultaneously within a same factor f_c. The *neighborhood* $\mathsf{n}(i)$ of site i is then defined as $\mathsf{n}(i) = \{j \in S : \{i, j\} \in E\}$ [4]. As a consequence

[3] Formal expression of the normalizing constant Z must not veil the fact that it is unknown and beyond reach in general, due to the intractable summation over Ω.

[4] $\mathsf{n} = \{\mathsf{n}(i), i \in S\}$ is called a *neighborhood system*, and the neighborhood of some subset $a \subset S$ is defined as $\mathsf{n}(a) = \{j \in S - a : \mathsf{n}(j) \cap a \neq \emptyset\}$.

of the definition, any subset c is either a singleton or composed of mutually neighboring sites: $\mathcal{C}$ is a set of *cliques* for $\mathbb{G}$.

When variables are attached to the pixels of an image, the most common neighborhood systems are the regular ones where a site away from the border of the lattice has four or eight neighbors. In the first case (first-order neighborhood system, like in Figure 1.a) subsets c have at most two elements, whereas, in the second-order neighborhood system cliques can exhibit up to 4 sites. However, other graph structures are also used: in segmentation applications where the image plane is partitioned, $\mathbb{G}$ might be the planar graph associated to the partition (Figure 1.b); and hierarchical image models often live on (quad)-trees (Figure 1.c).

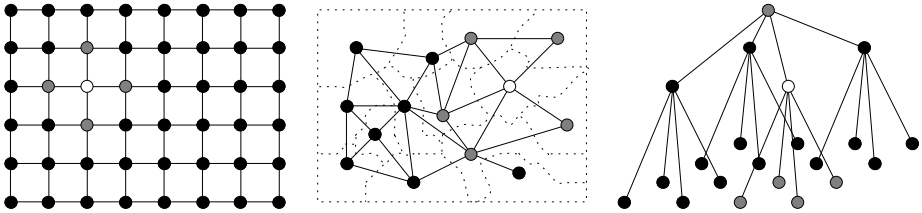

FIGURE 1. Three typical graphs supporting MRF-based models for image analysis: (a) rectangular lattice with first-order neighborhood system; (b) non-regular planar graph associated to an image partition; (c) quad-tree. For each graph, the grey nodes are the neighbors of the white one.

The independence graph conveys the key probabilistic information by *absent edges*: if i and j are not neighbors, $\mathsf{P}_X(x)$ can obviously be split into two parts respectively independent from x_i and from x_j. This suffices to conclude that the random variables X_i and X_j are independent given the others. It is the *pair-wise Markov property* [29, 39].

In the same fashion, given a set $a \subset S$ of vertices, P_X splits into $\prod_{c:c\cap a\neq\emptyset} f_c \times \prod_{c:c\cap a=\emptyset} f_c$ where the second factor does not depend on x_a. As a consequence $\mathsf{P}_{X_a|X_{S-a}}$ reduces to $\mathsf{P}_{X_a|X_{\mathfrak{n}(a)}}$, with:

$$\mathsf{P}_{X_a|X_{\mathfrak{n}(a)}}(x_a|x_{\mathfrak{n}(a)}) \propto \prod_{c:c\cap a\neq\emptyset} f_c(x_c) = \exp\{-\sum_{c:c\cap a\neq\emptyset} V_c(x_c)\}, \qquad (3)$$

with some normalizing constant $Z_a(x_{\mathfrak{n}(a)})$, whose computation by summing over all possible x_a is usually tractable. This is the *local Markov property*. The conditional distributions (3) constitute the key ingredients of iterative procedures to be presented, where a small site set a (often a singleton) is considered at a time.

It is possible (but more involved) to prove the *global Markov property* according to which, if a vertex subset A separates two other disjoint subsets B and C in $\mathbb{G}$ (i.e., all chains from $i \in B$ to $j \in C$ intersect A) then the random

vectors X_B and X_C are independent given X_A [29, 39]. It can also be shown that the three Markov properties are equivalent for strictly positive distributions. Conversely, if a positive distribution P_X fulfills one of these Markov properties w.r.t. graph $\mathbb{G}$ (X is then said to be a MRF on $\mathbb{G}$), then P_X is a Gibbs distribution of the form (2) relative to the same graph. This equivalence constitutes the Hammersley-Clifford theorem [3].

To conclude this section, we introduce two very standard Markov random fields which have been extensively used for image analysis purposes.

EXAMPLE 1: *Ising* MRF. Introduced in the twenties in statistical physics of condensed matter, and studied since then with great attention, this model deals with binary variables ($\Lambda = \{-1, 1\}$) that interact locally. Its simpler formulation is given by energy

$$U(x) = -\beta \sum_{\langle i,j \rangle} x_i x_j,$$

where summation is taken over all edges $\langle i, j \rangle \in E$ of the chosen graph. The "attractive" version ($\beta > 0$) statistically favors identity of neighbors. The conditional distribution at site i is readily deduced:

$$\mathsf{P}_{X_i|X_{\mathsf{n}(i)}}(x_i|x_{\mathsf{n}(i)}) = \frac{\exp\{\beta x_i \sum_{j \in \mathsf{n}(i)} x_j\}}{2\cosh(\beta \sum_{j \in \mathsf{n}(i)} x_j)}.$$

The Ising model is useful in detection-type problems where binary variables are to be recovered. Some other problems (essentially segmentation and classification) require the use of symbolic (discrete) variables with more than two possible states. For these cases, the Potts model also stemming from statistical physics [2], provides an immediate extension of the Ising model, with energy $U(x) = -\beta \sum_{\langle i,j \rangle} [2\delta(x_i, x_j) - 1]$, where δ is the Kronecker delta. □

EXAMPLE 2: *Gauss-Markov field.* It is a continuously-valued random vector ($\Lambda = \mathbb{R}$) ruled by a multivariate Gaussian distribution

$$\mathsf{P}_X(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\},$$

with expectation vector $\mathbb{E}(X) = \mu$ and covariance matrix Σ. The "Markovianity" shows up when each variable only interacts with a few others through the quadratic energy, that is, when the matrix $A = \Sigma^{-1}$ is *sparse*. Sites i and j are then neighbors in $\mathbb{G}$ iff the corresponding entry $a_{ij} = a_{ji}$ is non-null. Graph $\mathbb{G}$ is exactly the so-called *structure* of sparse matrix A [24]. In practice a Gauss-Markov field is often defined simply by its quadratic energy function $U(x) = \frac{1}{2} x^T A x - x^T b = \sum_{\langle i,j \rangle} a_{ij} x_i x_j + \sum_i (\frac{a_{ii}}{2} x_i - b_i) x_i$, with $b \in \mathbb{R}^n$ and A a sparse symmetric definite positive matrix. In that case, the expectation which corresponds to the most probable configuration, is the solution of the large sparse linear system $A\mu = b$.

Note that in the Gaussian case, any conditional or marginal distribution taken from P_X is Gaussian and can be explicitly written down by using adequate block partitioning of A and b [29, 37, 39]. All Markovian properties can then be directly deduced from this. Site-wise conditional distributions in particular turn out to be Gaussian laws

$$(X_i|X_{\mathsf{n}(i)} = x_{\mathsf{n}(i)}) \sim \mathcal{N}(\frac{1}{a_{ii}}(b_i - \sum_{j\in\mathsf{n}(i)} a_{ij}x_j), a_{ii}^{-1}) .$$

At this point, it is worth emphasizing the difference with so-called *simultaneous autoregressive models* [3] defined by:

$$\forall i \; a_{ii}X_i = b_i - \sum_{j\in\mathsf{n}(i)} a_{ij}X_j + W_i ,$$

with W_i's being i.i.d. reduced zero-mean Gaussian variables. As a fact, the resulting inverse covariance matrix of X in this case is A^TA, whose filling structure is larger than the one of A.

The nice properties of Gaussian MRFs, inherited from the quadratic form of their energy, make them the more popular models in case of continuous or "almost continuous" (i.e., $|\Lambda|$ very large) variables. □

3. Sampling from Gibbs distributions

In order to visually evaluate the statistical properties of the specified model, or simply to get synthetic images, one might want to draw the samples from the distribution P_X. A more important and technical reason for which this sampling can be of much help is that for all issues requiring an exhaustive visit of configuration set Ω (intractable in practice), e.g., summation or maximization over all possible occurrences, an approximate way to proceed consists in randomly visiting Ω for long enough according to distribution P_X. These approximation methods belong to the class of Monte Carlo methods [25, 35].

The dimensionality of the model makes sampling delicate (starting with the fact that normalizing constant Z is beyond reach). One has to resort to a Markov chain procedure on Ω which allows to sample successively from random fields whose distributions get closer and closer to the target distribution P_X. The locality of the model indeed permits to design a chain of random vectors $X^1, \ldots, X^m, \ldots$, without knowing Z, and such that P_{X^m} tends to P_X (for some distance) as m goes to infinity, whatever initial distribution P_{X^0} is. The crux of the method lies in the design of transition probabilities $\mathsf{P}_{X^{m+1}|X^m}$ based only on local conditional distributions stemming from P_X, and such that the resulting Markov chain is irreducible [5], aperiodic [6], and preserves the target

[5] There is a non-null probability to get from any x to any x' within a finite number of transitions: $\exists m : \; \mathsf{P}_{X^m|X^0}(x'|x) > 0$.

[6] Configuration set cannot be split in subsets that would be visited in a periodic way: $\nexists \; d > 1 : \Omega = \cup_{k=0}^{d-1}\Omega^k$ with $X^0 \in \Omega^0 \Rightarrow X^m \in \Omega^{m-d\lfloor m/d \rfloor}, \forall m$

distribution, i.e., $\sum_{x' \in \Omega} \mathsf{P}_{X^{m+1}|X^m}(x|x')\mathsf{P}_X(x') = \mathsf{P}_X(x)$ for any configuration $x \in \Omega$. The latter condition is especially met when the so-called "detailed balance" holds: $\mathsf{P}_{X^{m+1}|X^m}(x|x')\mathsf{P}_X(x') = \mathsf{P}_{X^{m+1}|X^m}(x'|x)\mathsf{P}_X(x)$.

Various suitable dynamics can be designed [35]. A quite popular one for image models is the *Gibbs sampler* [18] which directly uses conditional distributions from P_X to generate the transitions. More precisely S is split into small pieces (often singletons), which are "visited" either at random, or according to some pre-defined schedule. If a is the concerned set of sites at step m, and x the realization of X^m, a realization of X^{m+1} is obtained by replacing x_a in $x = (x_a, x_{S-a})$ by x'_a, according to the conditional law

$$\mathsf{P}_{X_a|X_{S-a}}(.|x_{S-a}) = \mathsf{P}_{X_a|X_{\mathsf{n}(a)}}(.|x_{\mathsf{n}(a)})$$

(this law of a few variables can be exactly derived, unlike P_X). The chain thus sampled is clearly irreducible and aperiodic (the null transition $x \to x$ is always possible), and detailed balance is verified since

$$\begin{aligned}
&\mathsf{P}_{X^{m+1}|X^m}(x'_a, x_{S-a}|x_a, x_{S-a})\mathsf{P}_X(x_a, x_{S-a}) = \\
&\quad = \mathsf{P}_{X_a|X_{S-a}}(x'_a|x_{S-a})\mathsf{P}_X(x_a, x_{S-a}) \\
&\quad = \frac{\mathsf{P}_X(x'_a, x_{S-a})\mathsf{P}_X(x_a, x_{S-a})}{\mathsf{P}_{X_{S-a}}(x_{S-a})}
\end{aligned}$$

is symmetric in x_a and x'_a.

From a practical point of view the chain thus designed is started from any configuration x_0, and run for a large number of steps. For m large enough, it has almost reached an equilibrium around P_X, and following configurations can be considered as (non-independent) samples from P_X. The decision whether equilibrium is reached is intricate, though. The design of criteria and indicators to answer this question is an active area of research in MCMC studies.

If the expectation of some function f of X has to be evaluated, ergodicity properties yield $\mathbb{E}[f(X)] = \lim_{m \to +\infty} \frac{f(x^0)+\ldots+f(x^{m-1})}{m}$. Practically, given a large but finite realization of the chain, one gets rid of the first m_0 out-of-equilibrium samples, and computes an average over the remainder:

$$\mathbb{E}[f(X)] \approx \frac{1}{m_1 - m_0} \sum_{m=m_0+1}^{m_1} f(x^m)\,.$$

EXAMPLE 3: *sampling Ising model.* Consider the Ising model from example 1 on a rectangular lattice with first-order neighborhood system. S is visited site-wise (sets a are singletons). If i is the current site and x the current configuration, x_i is updated according to the sampling of associated local conditional distribution: it is set to -1 with probability $\propto \exp\{-\beta \sum_{j \in \mathsf{n}(i)} x_j\}$, and to 1 with probability $\propto \exp\{\beta \sum_{j \in \mathsf{n}(i)} x_j\}$. Small size samples are shown in Figure 2, illustrating typical behavior of the model for increasing interaction parameter β.

This basic model can be refined. It can, for instance, be made anisotropic by weighting in a different way "horizontal" and "vertical" pairs of neighbors:

$$U(x) = -\beta_1 \sum_{\boxed{i}\boxed{j}} x_i x_j - \beta_2 \sum_{\substack{\boxed{i}\\ \boxed{j}}} x_i x_j.$$

Some typical patterns drawn from this version are shown in Figure 3 for different combinations of β_1 and β_2. □

FIGURE 2. Samples from an Ising model on a lattice with first-order neighborhood system and increasing $\beta = 0$ (in that case X_is are i.i.d., and sampling is direct), 0.7, 0.9, 1.1, 1.5, and 2.

FIGURE 3. Samples from an anisotropic Ising model on a lattice with first-order neighborhood system and $(\beta_1, \beta_2) = (5, 0.5),\ (5, 0.1),\ (1, -1),\ (-1, -1)$, respectively.

EXAMPLE 4: *sampling from Gauss-Markov random fields.* S being a rectangular lattice equipped with the second-order neighborhood system, consider the quadratic energy

$$\begin{aligned} U(x) \quad &= \beta_1 \sum_{\boxed{i}\boxed{j}} (x_i - x_j)^2 + \beta_2 \sum_{\substack{\boxed{i}\\ \boxed{j}}} (x_i - x_j)^2 + \\ &+ \beta_3 \sum_{\substack{\boxed{i}\\ \quad\boxed{j}}} (x_i - x_j)^2 + \beta_4 \sum_{\substack{\quad\boxed{i}\\ \boxed{j}}} (x_i - x_j)^2 + \varepsilon \sum_i x_i^2, \end{aligned}$$

with $\beta_1, \beta_2, \beta_3, \beta_4, \varepsilon$ some positive parameters. The quadratic form is obviously positive definite (since the minimum is reached for the unique configuration $x \equiv 0$) and thus defines a Gauss-Markov random field whose A matrix can be readily assembled using $a_{ij} = \frac{\partial^2 U}{\partial x_i \partial x_j}$. For current site i (away from the border) and configuration x, the Gibbs sampler resorts to sampling from the Gaussian law

$(X_i|X_{\mathsf{n}(i)} = x_{\mathsf{n}(i)}) \sim \mathcal{N}(\frac{\sum_{j\in\mathsf{n}(i)}\beta_{ij}x_j}{\varepsilon+8\bar{\beta}}, (2\varepsilon+16\bar{\beta})^{-1})$, where $\bar{\beta} = \frac{\beta_1+\beta_2+\beta_3+\beta_4}{4}$ and $\beta_{ij} = \beta_1, \beta_2, \beta_3$, or β_4, depending on $\langle i,j\rangle$ orientation. Some typical "textures" can be obtained this way for different parameter values (Figure 4). □

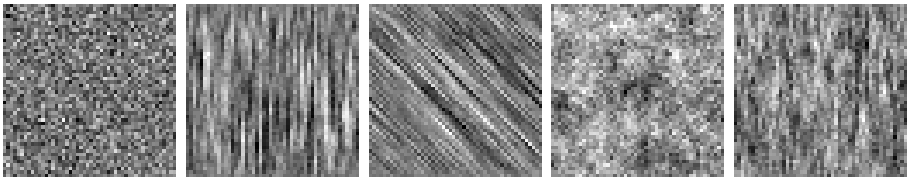

FIGURE 4. Samples from an anisotropic Gaussian model on a lattice with second-order neighborhood system and various values of $(\beta_1, \beta_2, \beta_3, \beta_4, \varepsilon)$.

4. ESTIMATING PARAMETERS

When it is possible to make available various occurrences of X, like images in Figure 2, 3 or 4, one can try to *learn* parameters of the assumed underlying Gibbs distribution $\mathsf{P}_X^\theta = Z(\theta)^{-1}\exp\{-\sum_c V_c^\theta\}$. The distribution thus adjusted might then be used either to generate "similar" realizations through sampling or simply as a compact information characterizing a class of patterns. Also, if different models are learned, they then might be used in competition to analyze and identify the content of an image, in the context of statistical pattern recognition.

Learning distribution parameters based on observed samples is a standard issue from statistics. This estimation is often conducted based on the *likelihood* of observed data. However, in the case of the Gibbs distributions under concern in image problems, the high dimensionality rises once again technical difficulties which have to be specifically addressed or circumvented.

Given a realization x^o (it could also be a set of realizations) supposed to arise from one Gibbs distribution among the family $\{\mathsf{P}_X^\theta, \theta \in \Theta\}$, the question is to estimate at best the "real" θ. Maximum likelihood estimation (MLE) seeks parameters that maximize the occurrence probability of x^0: $\hat{\theta} = \arg\max_\theta \mathcal{L}(\theta)$ with (log-)likelihood function $\mathcal{L}(\theta) = \ln \mathsf{P}_X^\theta(x^o)$. Unfortunately, the partition function $Z(\theta)$ which is here required, cannot be derived in general (apart from causal cases where it factorizes as a product of local partition functions).

A simple and popular way to tackle this problem, is to look instead at site-wise conditional distributions associated with data x^o: the global likelihood as a goodness-of-fit indicator is replaced by a sum of local indicators through the so-called *pseudo-likelihood* function [3], $\mathbb{L}(\theta) = \sum_i \ln \mathsf{P}^\theta_{X_i|X_{\mathsf{n}(i)}}(x_i^o|x^o_{\mathsf{n}(i)})$. Maximum pseudo-likelihood estimate (MPLE) is then:

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i\in S} [\ln Z_i(x^o_{\mathsf{n}(i)}, \theta) + \sum_{c\ni i} V_c^\theta(x_c^o)],$$

where local partition functions $Z_i(x^o_{\mathsf{n}(i)}, \theta)$ are usually accessible.

EXAMPLE 5: MPLE *of Ising model.* Consider the Ising model from example 1. Denoting $\mathtt{n}(x_a) = \sum_{i \in a} x_i$, the MPLE is obtained by maximizing

$$\mathbb{L}(\beta) = \sum_i [-\ln \cosh(\beta \mathtt{n}(x^o_{\mathtt{n}(i)})) + \beta x^o_i \mathtt{n}(x^o_{\mathtt{n}(i)})].$$

Gradient ascent can be conducted using $\frac{\mathrm{d}\mathbb{L}(\beta)}{\mathrm{d}\beta} = \sum_i \mathtt{n}(x^o_{\mathtt{n}(i)})[x^o_i - \tanh(\beta \mathtt{n}(x^o_{\mathtt{n}(i)}))]$. □

EXAMPLE 6: MPLE *of Gaussian model.* Consider the Gauss-Markov model from example 4, with all β_k's equal to β. Denoting $\sigma^2 = (2\varepsilon + 16\beta)^{-1}$ and $\gamma = 2\beta\sigma^2$, the site-wise conditional laws are $\mathcal{N}(\gamma \mathtt{n}(x_{\mathtt{n}(i)}), \sigma^2)$. Setting the pseudo-likelihood gradient to zero yields MPLE $\hat{\gamma} = \frac{\sum_i x^o_i \mathtt{n}(x^o_{\mathtt{n}(i)})}{\sum_i \mathtt{n}(x^o_{\mathtt{n}(i)})^2}$, $\hat{\sigma}^2 = \frac{\sum_i (x^o_i - \hat{\gamma}\mathtt{n}(x^o_{\mathtt{n}(i)}))^2}{n}$. □

In widely encountered cases where the energy U^θ depends linearly on the parameters (family $\{\mathsf{P}^\theta_X, \theta \in \Theta\}$ is then said to be *exponential*), i.e., $U^\theta = \sum_k \theta_k \sum_{c \in \mathcal{C}_k} V_k(x_c)$, for some partition of $\mathcal{C} = \cup_k \mathcal{C}_k$, the gradients of the likelihood and pseudo-likelihood take special forms:

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta)}{\partial \theta_k} &= [U^k(x^o) - \mathbb{E}_\theta(U^k(X))], \\ \frac{\partial \mathbb{L}(\theta)}{\partial \theta_k} &= \sum_i [U^k_i(x^o_i, x^o_{\mathtt{n}(i)}) - \mathbb{E}_\theta(U^k_i(X_i, x^o_{\mathtt{n}(i)}) | x^0_{\mathtt{n}(i)})], \end{aligned}$$

with $U^k(x) = \sum_{c \in \mathcal{C}_k} V_k(x_c)$ and $U^k_i(x_i, x_{\mathtt{n}(i)}) = \sum_{c \in \mathcal{C}_k : i \in c} V_k(x_c)$. In that context, the MLE can be directly sought by (stochastic) gradient ascent techniques where the expectation $\mathbb{E}_\theta(U^k(X))$ is approximated by sampling. Resulting MCMC MLE methods [16, 19, 41] that recent progress in MCMC techniques (like the use of importance sampling) makes more and more tractable, thus offer an alternative to MPLE .

Also, (good) theoretical properties of MLE and MPLE can be thoroughly investigated in the exponential case. In particular, the *asymptotic consistency*, that is the desirable property that $\hat{\theta}$ tends to the real parameter θ^* when data are actually drawn from some stationary [7] distribution $\mathsf{P}^{\theta^*}_X$, and the "size" of data increases to infinity, has been established under rather general conditions (see [12] for instance).

In the case of stationary Gibbs distributions on regular lattices, with small finite state space Λ, another useful approach to parameter estimation has been developed in a slightly different perspective. The idea of this alternative technique is to tune parameters such that site-wise conditional distributions arising from P^θ_X fit at best those *empirically* estimated [15]. One has first to determine on which information of neighborhood configuration $x_{\mathtt{n}(i)}$, the local distribution

[7] i.e., graph structure is regular, and off-border variables have the same conditional distributions.

$\mathsf{P}^{\theta}_{X_i|X_{\mathtt{n}(i)}}(.|x_{\mathtt{n}(i)})$ really depends. For instance, in an anisotropic Ising model, only $\mathtt{n}(x_{\mathtt{n}(i)})$ is relevant. The neighborhood configuration set $\Upsilon = \Lambda^{|\mathtt{n}(i)|}$ is partitioned accordingly:

$$\Upsilon = \bigcup_{\alpha} \Upsilon^{\alpha}, \text{ with } \mathsf{P}^{\theta}_{X_i|X_{\mathtt{n}(i)}}(.|x_{\mathtt{n}(i)}) = \mathsf{p}^{\theta}_{\alpha}(.) \text{ if } x_{\mathtt{n}(i)} \in \Upsilon^{\alpha}.$$

For each type α of neighborhood configuration, the conditional distribution is empirically estimated based on x^o [8] as $\hat{\mathsf{p}}_{\alpha}(\lambda) = \frac{\#\{i \in S:\ x^o_i = \lambda \text{ and } x^o_{\mathtt{n}(i)} \in \Upsilon^{\alpha}\}}{\#\{i \in S:\ x^o_{\mathtt{n}(i)} \in \Upsilon^{\alpha}\}}$. Then one tries to make $\mathsf{p}^{\theta}_{\alpha}$ and $\hat{\mathsf{p}}_{\alpha}$ as close as possible, for all α. One way to proceed consists in solving the simultaneous equations

$$\ln \frac{\mathsf{p}^{\theta}_{\alpha}(\lambda)}{\mathsf{p}^{\theta}_{\alpha}(\nu)} = \ln \frac{\hat{\mathsf{p}}_{\alpha}(\lambda)}{\hat{\mathsf{p}}_{\alpha}(\nu)},\ \forall \{\lambda, \nu\} \subset \Lambda,\ \forall \alpha.$$

Where energy is linear w.r.t. parameters, one ends up with an over-determined linear system of equations which is solved in the least-square sense. The asymptotic consistency of the resulting estimator has been established under certain conditions [22].

EXAMPLE 7: *empirical parameter estimation of Ising model.* For each possible value n_{α} of $\mathtt{n}(x_{\mathtt{n}(i)})$, the local conditional distribution is $\mathsf{p}^{\beta}_{\alpha}(\lambda) = \frac{\exp\{\beta \lambda n_{\alpha}\}}{2 \cosh \beta n_{\alpha}}$. The system to be solved is then:

$$\ln \frac{\mathsf{p}^{\beta}_{\alpha}(+1)}{\mathsf{p}^{\beta}_{\alpha}(-1)} = 2\beta n_{\alpha} = \ln \frac{\#\{i \in S:\ x^o_i = +1 \text{ and } \mathtt{n}(x^o_{\mathtt{n}(i)}) = n_{\alpha}\}}{\#\{i \in S:\ x^o_i = -1 \text{ and } \mathtt{n}(x^o_{\mathtt{n}(i)}) = n_{\alpha}\}},\ \forall \alpha,$$

yielding least-square solution $\hat{\beta} = \frac{\sum_{\alpha} n_{\alpha} g_{\alpha}}{2 \sum_{\alpha} n^2_{\alpha}}$, where g_{α} denotes the left hand side of the linear equation above. □

5. Inverse problems and Bayesian estimation

A learned Gibbs distribution might be used to identify pieces of texture present in an image, provided they are pointed out in some way. However, if one wants to address the joint issue of separating and recognizing these different pieces of texture, another question has to be dealt with at the same time as recognition, namely "where are the pieces of texture to be recognized?". This *classification* problem is typical of so-called *inverse problems*: based on a luminance image, an hidden underlying partition is searched.

In inverse problems from early vision, one tries to recover a large number of unknown or *hidden* variables based on the knowledge of another bunch of variables: given a set of data $y = (y_j)_{j=1}^{q}$, which are either plain luminance image(s) or quantities extracted beforehand from images(s) such as discontinuities, hidden variables of interest $x = (x)_{i=1}^{n}$ are sought (Figure 5).

[8] The model being stationary over the lattice, empirical estimation makes sense for large enough configuration x^o.

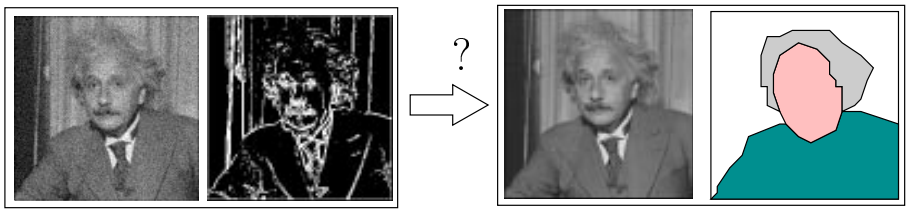

FIGURE 5. Typical inverse problem: data might be a degraded image and/or discontinuities of this image; variables to be estimated might constitute a restored version of the original image, or a partition of the image into meaningful regions.

Vectors x and y can be of the same nature (two images in restoration problems) or of completely different nature (a labeling in terms of region numbers and an image in segmentation and classification problems; a vector field and a couple of images in disparity or displacement estimation). Also, components of one of the vectors can be of various natures, like restored luminance and binary indicators of discontinuities on the edge lattice in case of image restoration with preservation of abrupt changes. For sake of concision, this latter possibility is not integrated in the following notations: Λ (resp. Λ^o) will indistinctly denote state spaces of all x_i's (resp. y_j's).

Stated in probabilistic terms, the problem consists in inferring the "best" occurrence of the random vector X in $\Omega = \Lambda^n$ given a realization $y \in (\Lambda^o)^q$ of the random vector Y. The key ingredient will be the conditional distribution $\mathsf{P}_{X|Y}(.|y)$, referred to as the *posterior distribution*. This distribution can be specified at once, or according to a two-step Bayesian modeling combining knowledge about how data could be explained from hidden variables, and expected (or at least desirable) statistical characteristics of the variables of interest [37].

Let us make more precise this standard Bayesian construction process. The first step amounts to modeling Y as a random function of X, $Y = F(X, W)$ where W is a non-observed random vector. F captures the process yielding observed data from underlying attributes. The simplest case is met in the restoration of an image corrupted with additive noise: $Y = X + W$.

Equivalently, one specifies in some way, the conditional distribution of data given X. This is the *conditional data likelihood*, $\mathsf{P}^{\theta_\ell}_{Y|X}$, which depends on some parameter vector θ_ℓ. With the goal of getting back to x from given y, one may simply try to invert this modeling as $X = f(Y, W)$. The maximum likelihood estimate, which corresponds to the configuration equipping observed data with highest probability, is then obtained by setting W to its most probable occurrence (null in general). Unfortunately such a method is, in general, either intractable ($F(x, w)$ is a many-to-one mapping for given w) or simply not sensi-

ble (e.g., for additive white noise, maximum likelihood restoration provides the observed image itself, as a result!). The inverted model might also happen to be far too sensitive (non-differentiable), yielding completely different estimates for slightly different input data.

A Bayesian approach allows to fix these problems through the specification of a *prior* distribution $\mathsf{P}_X^{\theta_p}(x)$ for X. Often, prior knowledge captured by this distribution is loose and generic, merely dealing with the regularity of desired estimates (it is then related to *Tikhonov regularization* [38]). The prior might however be far more specific about the class of acceptable configurations (in that case Ω might even be restricted to some parametric subspace) and their respective probabilities of occurrence. Except in extreme cases where all prior has been put into the definition of a reduced configuration set equipped with uniform prior, the prior distribution is chosen as an interacting Gibbs distribution over Ω.

Modeling is then completed by forming the joint and posterior distributions from previous ingredients. Bayes' rule provides:

$$\mathsf{P}_{X|Y}^{\theta} = \frac{\mathsf{P}_{Y|X}^{\theta_\ell}\mathsf{P}_X^{\theta_p}}{\mathsf{P}_Y} \propto \mathsf{P}_{Y|X}^{\theta_\ell}\mathsf{P}_X^{\theta_p} = \mathsf{P}_{XY}^{\theta},$$

with $\theta = (\theta_p, \theta_\ell)$. The estimation problem can now be defined in terms of the posterior distribution.

A natural way to proceed is to look for the most probable configuration x given the data:

$$\hat{x} = \arg\max_{x\in\Omega} \mathsf{P}_{X|Y}(x|y).$$

This constitutes the maximum *a posteriori* (MAP) estimator. Although it is often considered as a "brutal" estimator [32], it remains the most popular for it is simply connected to the posterior distribution and the associated energy function. One has simply to devise the energy function (as a sum of local functions) of x and y, and the estimation goal is fixed at once, as a global minimizer in x of this energy. This means in particular that here, no probabilistic point of view is finally required. Numerous energy-based approaches to inverse problems have thus been proposed as an alternative to the statistical framework. A very active class of such deterministic approaches is based on *continuous* functionals, variational methods, and PDEs [9].

[9] As a consequence, one might legitimately wonder why one should bother with a probabilistic formulation. Elements in favor of probabilistic point of view rely upon the variety of tools it offers. As partly explained in this paper, statistical tools allow to learn parameters, to generate "typical" instances, to infer unknown variables in different ways (not only the one using energy minimization), to assess estimation uncertainty, or to capture and combine all sorts of priors within the Bayesian machinery. On the other hand, continuous (deterministic) approaches allow to derive theoretical properties of models under concern, in a fashion that is beyond reach of most discrete approaches, whether they are stochastic or not. Both points of view are therefore complementary. Besides, it is common that they eventually yield similar discrete implementations.

To circumvent the crude globality of the MAP estimator, a more local estimator is also widely employed. It associates to each site the value of X_i that is the most probable given all the data:

$$\hat{x}_i = \arg\max_{\lambda \in \Lambda} \mathsf{P}_{X_i|Y}(\lambda|y).$$

It is referred to as the "marginal posterior mode" (MPM) estimator [32]. It relies on site-wise posterior marginals $\mathsf{P}_{X_i|Y}$ which have to be derived, or approximated, from the global posterior distribution $\mathsf{P}_{X|Y}$. Note that for Gaussian posterior distribution, the MAP and MPM estimators coincide with the posterior expectation whose determination amounts to solving a linear system of equations.

Both Bayesian estimators have now to be examined in the special case of factorized distributions under concern. The prior model is captured by a Gibbs distribution $\mathsf{P}_X^{\theta_p}(x) \propto \exp\{-\sum_{c\in\mathcal{C}^p} V_c^{\theta_p}(x_c)\}$ specified on some prior graph $[S, E^p]$ through potential $\{V_c^{\theta_p}, c \in \mathcal{C}^p\}$, and a data model is often chosen of the following form:

$$\mathsf{P}_{Y|X}^{\theta_\ell}(y|x) \propto \exp\{-\sum_{j\in R} V_j^{\theta_\ell}(x_{d_j}, y_j)\},$$

where $R = \{1, \ldots, q\}$ is the data site set and $\{d_j, j \in R\}$ is a set of small site subsets of S (Figure 6.a). Note that this data model specification should be such that the normalizing constant, if unknown, is independent from x (otherwise the posterior distribution of x will be incompletely defined; see footnote 10). The resulting posterior distribution is a Gibbs distribution parameterized by θ and y: [10]

$$\mathsf{P}_{X|Y}^{\theta}(x|y) = \frac{1}{Z(\theta, y)} \exp\{\underbrace{-\sum_{c\in\mathcal{C}^p} V_c^{\theta_p}(x_c) - \sum_{j\in R} V_j^{\theta_\ell}(x_{d_j}, y_j)}_{-U^\theta(x,y)}\}.$$

In the associated posterior independence graph $[S, E]$, any two sites $\{i, k\}$ can be neighbors either through a V_c function (i.e., $\{i, k\} \in E^p$), or through a function V_j (i.e., $\exists j \in R : \{i, k\} \subset d_j$). This second type of neighborhood relationship thus occurs between components of X that both participate to the "formation" of a same component of Y (Figure 6.a). The neighborhood system of the posterior model is then at least as big as the one of the prior model (see example in Figure 6.b). In site-wise measurement cases (i.e., $R \equiv S$, and, $\forall i$, $d_i = \{i\}$), the two graphs are identical.

EXAMPLE 8: *luminance-based classification.* The luminance of the observed image being seen as continuous ($\Lambda^o = \mathbb{R}$), one tries to partition the pixel set S

[10] Many studies actually start by the specification of some energy function $U(x, y) = \sum_c V_c(x_c) + \sum_j V_j(x_{d_j}, y_j)$. Then, unless $\sum_y \exp\{-\sum_j V_j\}$ is independent from x, the prior deduced from $\mathsf{P}_{XY} \propto \exp\{-U\}$ is *not* $\mathsf{P}_X \propto \exp\{-\sum_c V_c\}$.

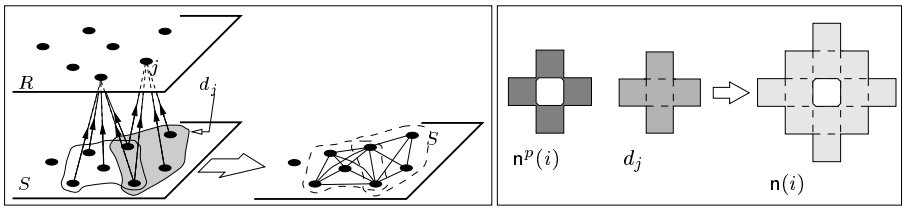

FIGURE 6. (a) Ingredients of data model and neighboring relationship they induce on x components; (b) posterior neighborhood on lattice induced by first-order prior neighborhood system and symmetric five-site subsets d_j centered at j.

into M classes ($\Lambda = \{1, \ldots, M\}$ and $S \equiv R$) associated to previously learned means and variances $\theta_\ell = (\mu_\lambda, \sigma_\lambda^2)_{\lambda=1}^M$. Simple point-wise measurment model assuming $(Y_i|X_i = \lambda) \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2)$ results in

$$\mathsf{P}_{Y|X}(y|x) = \exp\{-\sum_i \underbrace{[\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} + \frac{1}{2}\ln(2\pi\sigma_{x_i}^2)]}_{V_i(x_i, y_i)}\}.$$

A standard prior is furnished by the Potts model

$$\mathsf{P}_X^{\theta_p}(x) \propto \exp\{\sum_{\langle i,j\rangle} \beta[2\delta(x_i, x_j) - 1]\}$$

with respect to some prior graph structure ($\theta_p = \beta$). □

EXAMPLE 9: *Gaussian deblurring.* The aim is to recover an image from a filtered and noisy observed version, within a continuous setting ($R \equiv S$, $\Lambda = \Lambda^o = \mathbb{R}$). The data model is $Y = BX + W$ with B a sparse matrix associated with a known convolution kernel, and W a Gaussian white noise with known variance σ_W^2. Using the stationary isotropic Gaussian smoothing prior from example 6, one gets the posterior model:

$$\mathsf{P}_{X|Y}(x|y) \propto \exp\{-\sum_{\langle i,j\rangle} \beta(x_i - x_j)^2 - \sum_i \varepsilon x_i^2 - \sum_j \underbrace{\frac{(y_j - \sum_{i \in d_j} b_{ji} x_i)^2}{2\sigma_W^2}}_{V_j(x_{d_j}, y_j)}\},$$

where $d_j = \{i \in S \ : \ b_{ji} \neq 0\}$ corresponds to the filter support window centered at j. The graph structure of the posterior model is then usually larger than the prior structure (Figure 6.b). The posterior energy in matrix form is $U(x, y) = \beta\|Dx\|^2 + \varepsilon\|x\|^2 + \frac{1}{2\sigma_W^2}\|y - Bx\|^2$ where $Dx = (x_i - x_j)_{\langle i,j\rangle \in E^p}$ defines a many-to-one operator. The MAP estimate is then the solution of the linear system

$$(\beta D^T D + \varepsilon \mathrm{Id} + \frac{1}{2\sigma_W^2} B^T B)x = \frac{1}{2\sigma_W^2} B^T y \qquad \square.$$

The dimensionality of models under concern makes in general intractable the direct and exact derivation of both Bayesian estimators: indeed, MAP requires a global minimization over Ω, while MPM is based on marginal computations (i.e., summing out x_{S-i} in $\mathsf{P}_{X|Y}$, for each i). Again, iterative procedures based on local moves have to be devised.

The marginal computation required by MPM is precisely the kind of task that can be approximately performed using sampling. As described in §3, a long sequence of configurations $x^0, \ldots, x^{m_1}$ can be iteratively generated such that beyond some rank m_0 they can be considered as sampled from the Gibbs distribution $\mathsf{P}_{X|Y}(.|y)$. Ergodicity relation applied to the function $f(X) = \delta(X_i, \lambda)$ for some $i \in S$ and $\lambda \in \Lambda$ yields

$$\mathsf{P}_{X_i|Y}(\lambda|y) = \mathbb{E}[\delta(X_i, \lambda)|Y = y] \approx \frac{\#\{m \in [m_0+1, m_1] : x_i^m = \lambda\}}{m_1 - m_0},$$

that is the posterior marginal is approximated by the empirical frequency of appearance. The MPM estimator is then replaced by $\hat{x}_i = \arg\max_\lambda \#\{m \in [m_0+1, m_1] : x_i^m = \lambda\}$.

Iterative research of MAP estimate can be either stochastic or deterministic. In the first case, one makes use of so-called *simulated annealing* which relies on a clever use of MCMC sampling techniques [18]. The idea consists in sampling from Gibbs distribution with energy $\frac{U(x,y)}{T}$, where T is a "temperature" parameter which slowly decreases to 0. The cooling schedules insuring theoretical convergence to a *global* minimizer [11] unfortunately result in impractically long procedures. Deterministic counterparts are therefore often preferred, although they usually require a sensible initialization not to get stuck in too poor *local* minima. With a continuous state space, all gradient descent techniques or iterative system solving methods can be used. With both discrete and continuous state spaces, the simple "iterated conditional modes" (ICM) method [4] can be used: the component at current site i is updated so as to minimize the energy. In the point-wise measurement case, this yields:

$$x_i^{m+1} = \arg\min_\lambda \sum_{c \in \mathcal{C}^p : i \in c} V_c[(\lambda, x_{c-i}^m)] + V_i(\lambda, y_i),$$

where x^m is the current configuration.

EXAMPLE 10: *using model from example 8.* The posterior distribution at site i is

$$\mathsf{P}_{X_i|X_{\mathsf{n}(i)},Y_i}(\lambda|x_{\mathsf{n}(i)}, y_i) \propto \exp\{\beta \sum_{j \in \mathsf{n}(i)} [2\delta(\lambda, x_j) - 1] - \frac{(y_i - \mu_\lambda)^2}{2\sigma_\lambda^2} - \ln \sigma_\lambda\}.$$

The MPM estimate is approximated using a Gibbs sampler based on these distributions, simulated annealing also iteratively samples from these distributions

[11] If $T_m \geq \frac{C}{\ln m}$ for large enough constant C then $\lim_{m \to +\infty} \Pr\{X^m \in \arg\max U\} = 1$ [18].

but with energy scaled by temperature T_m, and the ICM updates the current x by setting x_i to $\arg\min_\lambda[\ln\sigma_\lambda + \frac{(y_i-\mu_\lambda)^2}{2\sigma_\lambda^2} - 2\beta\sum_{j\in\mathfrak{n}(i)}\delta(\lambda,x_j)]$. □

EXAMPLE 11: *using Gaussian model from example 9.* Letting $\varepsilon = 0$, ICM update at current site i picks the mean of the site-wise posterior conditional distribution:

$$x_i^{m+1} = (16\sigma_W^2\beta + \sum_j b_{ji}^2)^{-1}[2\sigma_W^2\beta\mathfrak{n}(x_{\mathfrak{n}(i)}^m) + \sum_j b_{ji}(y_j - \sum_{k\neq i} b_{jk}x_k^m)].$$

This exactly corresponds to the Gauss-Seidel iteration applied to the linear system of which the MAP estimate is the solution. □

6. Parameter estimation with incomplete data

At last, we deal with the estimation of parameters within inverse problem modeling. Compared with the setting from §4, the issue is dramatically more complex since only part of the variables (namely y) of the distribution P_{XY}^θ to be tuned are available. One talks about an *incomplete data* case, or *partially observed model*. This complicated issue arises when it comes to designing systems able to automatically adapt their underlying posterior distribution to various input data. It is for instance at the heart of *unsupervised* classification problems.

A pragmatic technique is to cope simultaneously with estimation of x and estimation of θ within an alternate procedure: for a given θ, infer x according to the chosen Bayesian estimator; given current estimate of x, learn parameters from the pair (x,y), as in the complete data case.

A more satisfactory and sounder idea consists in extending likelihood ideas from the complete data case [12]. The aim is then to maximize the data likelihood $\mathcal{L}(\theta) = \ln \mathsf{P}_Y^\theta(y) = \ln\sum_x \mathsf{P}_{XY}^\theta(x,y)$, whose derivation is intractable. However, its gradient $\nabla\mathcal{L}(\theta) = -\mathbb{E}_\theta[\nabla_\theta \ln \mathsf{P}_{XY}^\theta(X,y)|Y=y]$ suggests an iterative process where expectation is taken with respect to the current fit $\theta^{(k)}$ and the resulting expression is set to zero. This amounts to maximizing the conditional expectation $\mathbb{E}_{\theta^{(k)}}\left[\ln \mathsf{P}_{XY}^\theta(X,y)|Y=y\right]$, with respect to θ. The maximizer is the new parameter fit $\theta^{(k+1)}$. The resulting procedure is the *Expectation-Maximization algorithm* (EM), which has been introduced in the different context of mixtures of laws [14]. It can be shown that the associated sequence of likelihoods $\{\mathcal{L}(\theta^{(k)})\}$ is well increasing. However, the application of the plain EM algorithm involves twofold difficulties with high-dimensional Gibbs distributions: (1) the joint distribution P_{XY}^θ is usually known up to its partition function $Z(\theta)$ (the indetermination usually comes from the one of prior partition function); (2) computation of the conditional expectation is usually intractable. As in the complete data case, the first problem can be circumvented either by

[12] For exponential families, gradient ascent techniques have in particular been extended to partially observed case [1, 42].

MCMC techniques [42] or by considering pseudo-likelihood, i.e., replacing P_{XY} by $\mathsf{P}_{Y|X} \times \prod_i \mathsf{P}_{X_i|X_{\mathsf{n}(i)}}$. Concerning the second point, MCMC averaging allows to approximate expectations based on samples $x^{m_0}, \ldots, x^{m_1}$ drawn from the conditional distribution $\mathsf{P}_{X|Y}^{\theta^{(k)}}$ [10]. Note that in the incomplete data case, both MLE and MPLE remain asymptotically consistent in general [13].

Distinguishing prior parameters from data model parameters, MPLE with EM procedure and MCMC expectation approximations yields the following update:

$$\begin{cases} \theta_p^{(k+1)} = \arg\min_{\theta_p} \frac{1}{m_1 - m_0} \sum_i \sum_m [\ln Z_i(\theta_p, x^m_{\mathsf{n}(i)}) + \sum_{c \ni i} V_c^{\theta_p}(x_c^m)] \\ \theta_\ell^{(k+1)} = \arg\min_{\theta_\ell} \frac{1}{m_1 - m_0} \sum_j \sum_m V_j^{\theta_\ell}(x^m_{d_j}, y_j) \end{cases}$$

when $V_j = -\ln \mathsf{P}_{Y_j|X_{d_j}}$. It has also been suggested to compute first maximum (pseudo-)likelihood estimators on each sample, and then to average the results. The resulting update scheme is similar to the previous one, with maximization and summation w.r.t. samples being switched. With this method of averaging complete data-based estimators computed on samples drawn from $\mathsf{P}_{X|Y}^{\theta^{(k)}}$, any other estimator might be used as well. In particular, in case of reduced finite state space Λ, the empirical estimator presented in §4 can replace the MPLE estimator on the prior parameters. It is the *iterative conditional estimation* (ICE) method [34].

It must be said that parameter estimation of partially observed models remains in a Markovian context a very tricky issue due to the huge computational load of the techniques sketched here, and to convergence problems toward local minima of low quality. Besides, some theoretical aspects (e.g., asymptotic normality) still constitute open questions.

EXAMPLE 12: ICE *for classification model from example 8.* For a given parameter fit $\theta^{(k)}$, Gibbs sampling using conditional distributions

$$\mathsf{P}^{\theta^{(k)}}_{X_i|X_{\mathsf{n}(i)},Y_i}(\lambda|x_{\mathsf{n}(i)}, y_i) \propto \exp\{ \sum_{j \in \mathsf{n}(i)} \beta^{(k)}[2\delta(\lambda, x_j) - 1] - \frac{1}{2{\sigma_\lambda^{(k)}}^2}(y_i - \mu_\lambda^{(k)})^2 \}$$

provides samples $x^0, \ldots, x^m$. MLEs of data model parameters are computed from (x^m, y)'s and averaged:

$$\begin{aligned} \mu_\lambda^{(k+1)} &= \frac{1}{m_1 - m_0} \sum_{m=m_0+1}^{m_1} \frac{\sum_{i:x_i^m=\lambda} y_i}{\#\{i : x_i^m = \lambda\}}, \\ {\sigma_\lambda^{(k+1)}}^2 &= \frac{1}{m_1 - m_0} \sum_{m=m_0+1}^{m_1} \frac{\sum_{i:x_i^m=\lambda} (y_i - \mu_\lambda^{(k+1)})^2}{\#\{i : x_i^m = \lambda\}}. \end{aligned}$$

As for the prior parameter β, empiric estimators can be used for small M. It will exploit the fact that the prior local conditional distribution $\mathsf{P}_{X_i|X_{\mathsf{n}(i)}}(.|x_{\mathsf{n}(i)})$ depends only on composition $(n_1^\alpha \ldots n_M^\alpha)$ of neighborhood:

$$\forall x_{\mathsf{n}(i)} \in \Upsilon^\alpha,\ \#\{j \in \mathsf{n}(i) : x_j = \lambda\} = n_\lambda^\alpha$$

and

$$\mathsf{P}_{X_i|X_{\mathsf{n}(i)}}(\lambda|x_{\mathsf{n}(i)}) \propto \exp\{\beta(2n_\lambda^\alpha - |\mathsf{n}(i)|)\}. \qquad \square$$

7. Some research directions

Within the domain of MRFs applied to image analysis, two particular areas have been exhibiting a remarkable vitality for the past few years. Both tend to expand the modeling capabilities and the inference facilities of standard MRF-based models. The first one concerns the adding of a new set of so-called *auxiliary variables* to a given model P_X, whereas the second one deals with the definition of *hierarchical models*.

7.1. Auxiliary variables and augmented models

Auxiliary variable-based methods first appeared in statistical physics as a way to accelerate MCMC sampling. To this end, an augmented model P_{XD} is considered, where D is the set of auxiliary variables, such that (i) the marginal of X arising from P_{XD} coincides with the pre-defined P_X: $\sum_d \mathsf{P}_{XD}(.,d) = \mathsf{P}_X(.)$; (ii) sampling from P_{XD} can be done in a more efficient way than the one from P_X, by alternatively sampling from $\mathsf{P}_{D|X}$ and $\mathsf{P}_{X|D}$. The augmented model is usually specified through $\mathsf{P}_{D|X}$. The resulting joint model $\mathsf{P}_{XD} = \mathsf{P}_{D|X}\mathsf{P}_X$ then obviously verifies (i). Alternate sampling then first requires to derive $\mathsf{P}_{D|X}$ from the joint distribution.

For Ising and Potts models, the introduction of binary bond variables $D = \{D_{ij}, \langle i,j\rangle \in \mathcal{C}\}$ within the Swendsen-Wang algorithm [36] has been extremely fruitful. The original prior model is augmented through specification

$$\mathsf{P}_{D|X} = \prod_{\langle i,j\rangle} \mathsf{P}_{D_{ij}|X_i,X_j}, \text{ and } \mathsf{P}_{D_{ij}|X_i,X_j}(0|x_i,x_j) = \begin{cases} 1 & \text{if } x_i \neq x_j, \\ \exp -\beta & \text{otherwise.} \end{cases}$$

$\mathsf{P}_{D|X}$ is trivial to sample from. The nice fact is that the resulting distribution $\mathsf{P}_{X|D}$ specifies that sites of connected components defined on S by 1-valued bonds must have the same label, and that each of these *clusters* can get one of the possible states from Λ with equal probability. As a consequence sampling from $\mathsf{P}_{X|D}$ is very simple and introduces *long range interactions* by simultaneously updating all x_i's of a possibly large cluster.

In a different prospect, auxiliary variables have been used in order to introduce meaningful non-linearities within quadratic models. Despite their practical convenience, quadratic potentials like in Gaussian MRFs, are often far too "rigid": they discourage too drastically the deviations from mean configurations. Geman and Geman thus introduced a *binary line-process* to allow an adaptive detection and preservation of discontinuities within their Markovian restoration model [18]. In the simplest version, the augmented smoothing prior is $\mathsf{P}_{XD}(x,d) \propto \exp\{-\beta \sum_{\langle i,j\rangle} d_{ij}[(x_i - x_j)^2 - \tau]\}$ which favors discontinuity appearing $d_{ij} = 0$ (and thus a suspension of smoothing between x_i and x_j) as

soon as the difference $(x_i - x_j)^2$ exceeds some threshold τ. More precisely, in case of MAP estimation (with data model such that $\mathsf{P}_{Y|XD} = \mathsf{P}_{Y|X}$), it is readily established that

$$\begin{aligned}\min_{x,d}\{\beta \textstyle\sum_{\langle i,j\rangle} d_{ij} \quad & [(x_i - x_j)^2 - \tau] - \ln \mathsf{P}_{Y|X}(y|x)\} \\ & = \min_x\{\beta \textstyle\sum_{\langle i,j\rangle} \min[(x_i - x_j)^2 - \tau, 0] - \ln \mathsf{P}_{Y|X}(y|x)\},\end{aligned}$$

where the truncated quadratic potentials appear after optimal elimination of the binary d_{ij}'s [7]. In the view of that, the model associated with the energy in the right hand side could as well be used, without talking about auxiliary variables. However, the binary variables capture in that case the notion of discontinuity, and the Bayesian machinery allows to add a prior layer on them about likely and unlikely spatial contour configurations [18].

Starting from the idea of a bounded smoothing potential that appeared in the minimization rewriting above, a differentiable potentials of the same type, but with improved flexibility and numerical properties, have been introduced. Stemming from statistics where they allow *robust* fitting of parametric models [27], such cost functions penalize less drastically residuals than quadratics do: they are even functions $\rho(u)$, increasing on $\mathbb{R}^+$ and with derivative negligible at infinity compared with the one of u^2 ($\lim_{u\to+\infty} \frac{\rho'(u)}{2u} = 0$). It turns out that if $\phi(u) = \rho(\sqrt{u})$ defined on $\mathbb{R}^+$ is concave, $\rho(u)$ is the inferior envelope of a family of parabolas $\{zu^2 + \psi(z), z \in [0, \xi]\}$ continuously indexed by variable z, where $\xi = \lim_{u\to 0^+} \phi'(u)$ [5, 17]:

$$\rho(u) = \min_{z\in[0,\xi]} zu^2 + \psi(z), \text{ with } \arg\min_{z\in[0,\xi]} zu^2 + \psi(z) = \frac{\rho'(u)}{2u}.$$

Function $\rho(u) = 1 - \exp(-u^2)$ is a common example of such a function. Note that the optimal auxiliary variable $\rho'(u)/2u = \phi'(u^2)$ decreases to zero as the residual u^2 goes to infinity. Defining a smoothing potential with such a function yields, from a MAP point of view, and only concentrating on the prior part:

$$\min_x \sum_{\langle i,j\rangle} \rho(x_i - x_j) = \min_{x,d} \sum_{\langle i,j\rangle} d_{ij}(x_i - x_j)^2 + \psi(d_{ij}).$$

A *continuous line process* D is thus introduced. MAP estimation can be performed on the augmented model according to an alternate procedure: given d, the model is quadratic with respect to x; given x, the optimal d is obtained in closed form as $\hat{d}_{ij} = \phi'[(x_i - x_j)^2]$. Where the model is Gaussian for fixed d, this minimization procedure amounts to *iteratively reweighted least squares*. Note that such non-quadratic cost function can as well be used within the data model [5], to permit larger departures from this (usually crude) model.

It is only recently that a global picture of the link between robust estimators from statistics, line processes for discontinuity preservation, mean field approximation from statistical physics, "graduate non-convexity" continuation

method, and adaptative filtering stemming from anisotropic diffusion, has been clearly established [5, 6]. This new domain of research is now actively investigated by people with various backgrounds.

7.2. Hierarchical model

While permitting tractable single-step computations, the locality which is at the heart of Markovian modeling, results in a very slow propagation of information. As a consequence, iterative sampling and inference procedures reviewed in previous sections may converge very slowly. This motivates the search either for improved algorithms, or for new models allowing non-iterative or more efficient manipulation.

So far, the more fruitful approaches in both cases have relied on some notion of *hierarchy* (see [21] for a recent review). Hierarchical models or algorithms allow to integrate in a progressive and efficient way the information (especially in the case of multiresolution data, when images come into a hierarchy of scales). They thus provide gains both in terms of computational efficiency and of result quality, along with new modeling capabilities.

Algorithm-based hierarchical approaches are usually related to well-known *multigrid* resolution techniques from numerical analysis [23] and statistical physics [20], where an increasing sequence of nested spaces $\Omega^L \subset \Omega^{L-1} \ldots \subset \Omega^0 = \Omega$ is explored in a number of possible ways. Configurations in subset Ω^l are described by a reduced number of variables (the coordinates in a basis of subspace Ω^l in the linear case) according to some proper mapping Φ^l, i.e., $\Omega^l = \mathrm{Im}\Phi^l$. Let Γ^l be the corresponding reduced configuration set which Φ^l is defined from. If, for instance, Ω^l is the set of configurations that are piecewise constant with respect to some partition of S, $x \in \Omega^l$ is associated to the reduced vector x^l made up of the values attached by x to each subset of the partition. As for MAP estimation, inference conducted in Ω^l yields:

$$\arg\max_{x \in \Omega^l} P_{X|Y}(x|y) = \Phi^l[\arg\min_{x^l \in \Gamma^l} U(\Phi^l(x^l), y)].$$

The exploration of the different subsets is usually done in a *coarse-to-fine* fashion, especially in the case of discrete models [8, 26]: approximate MAP estimate $\hat{x}^{l+1}$ reached in Γ^{l+1} provides initialization of iterations in Γ^l, through $(\Phi^l)^{-1} \circ \Phi^{l+1}$ (which exists due to inclusion $\mathrm{Im}(\Phi^{l+1}) \subset \mathrm{Im}(\Phi^l)$). This hierarchical technique has proven useful to accelerate deterministic minimization while improving the quality of results [26].

Model-based hierarchical approaches, instead, aim at defining a new global hierarchical model $X = (X^0, X^1, \ldots, X^K)$ on $S = S^0 \cup S^1 \cup \ldots \cup S^K$, which has nothing to do with any original (spatial) model. This is usually done in a Markov chain-type causal way, specifying the "coarsest" prior P_{X^0} and coarse-to-fine transition probabilities $\mathsf{P}_{X^{k+1}|X^k \ldots X^0} = \mathsf{P}_{X^{k+1}|X^k}$ as local factor products: $\mathsf{P}_{X^{k+1}|X^k} = \prod_{i \in S^{k+1}} \mathsf{P}_{X_i|X_{p(i)}}$, where each node $i \in S^{k+1}$ is bound to a *parent* node $p(i) \in S^k$. When S^0 reduces to a single site r, the independence

graph corresponding to the Gibbs distribution $\mathsf{P}_X = \mathsf{P}_{X_r} \prod_{i \neq r} \mathsf{P}_{X_i|X_{p(i)}}$ is a *tree* rooted at r. When X_i's at level K are related to pixels, this hierarchical model usually lies on the nodes of a *quad-tree* whose leaves fit the pixel lattice [9, 28, 31].

Even if one is only interested in last level variables X^K, the whole structure has to be manipulated. But, its peculiar tree nature allows, like in case of Markov chains, to design *non-iterative* MAP and MPM inference procedures made of two sweeps: a bottom-up pass propagating all information to the root, and a top-down one which, in turn, allows to get optimal estimate at each node given *all the data*. In case of Gaussian modeling, these procedures are identical to techniques from linear algebra for direct factorization and solving of large sparse linear systems with tree structure [24].

As for the sampling issue, drawing from a hierarchical prior is immediate, provided that one can sample P_X^0. Drawing from the posterior model $\mathsf{P}_{X|Y} \propto \mathsf{P}_{Y|X}\mathsf{P}_X$ requires first to recover its causal factorization, i.e., to derive $\mathsf{P}_{X_i|X_{p(i)},Y}$. This can be done in one single pass resorting to bottom-up marginalizations. Also note that the prior partition function is exactly known as a product of normalizations of the local transition probabilities. This makes EM-type procedures much lighter [9, 28].

The computational and modeling appeal of these tree-based hierarchical approaches is, so far, moderated by the fact that their structure might appear quite artificial for certain types of problems or data: it is not shift-invariant with respect to the pixel lattice for instance, often resulting in visible non-stationarity in estimates; also the relevance of the inferred variables at coarsest levels is not obvious (especially at the root). Much work remains ahead for extending the versatility of models based on this philosophy.

8. Last words

Within the limited space of this presentation, technical details and most recent developments of evoked issues were hardly touched, whereas some other issues were completely left out (e.g., approximation of multivariate distributions with causal models, advanced MCMC sampling techniques, learning of independence structure and potential form, etc.). A couple of recent books referenced hereafter propose more complete expositions with various emphasis and gather collections of state-of-the-art applications of Gibbsian modeling, along with complete reference lists.

As a final word, it might be worth noting that MRFs do not constitute any longer an isolated class of approaches to image problems. Instead they now exhibit a growing number of stimulating connections with other active areas of computer vision research (PDEs for image analysis, Bayesian network from AI, neural networks, parallel computations, stochastic geometry, mathematical morphology, etc.), as highlighted by a number of recent "transversal" publications (e.g., [6, 33]).

References

1. P.M. Almeida, B. Gidas (1993). A variational method for estimating the parameters of MRF from complete and noncomplete data. *Ann. Applied Prob.* **3**(1), 103–136.
2. R.J. Baxter (1992). *Exactly solved models in statistical mechanics.* Academic Press, London.
3. J. Besag (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Soc.* B **36**, 192–236.
4. J. Besag (1986). On the statistical analysis of dirty pictures. *J. Royal Statist. Soc.* B **48** (3), 259–302.
5. M. Black, A. Rangarajan (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int. J. Computer Vision* **19** (1), 75–104.
6. M. Black, G. Sapiro, D. Marimont, D. Heeger. Robust anisotropic diffusion. *IEEE Trans. Image Processing.* To appear.
7. A. Blake, A. Zisserman (1987). *Visual reconstruction.* The MIT Press, Cambridge.
8. C. Bouman, B. Liu (1991). Multiple resolution segmentation of textured images. *IEEE Trans. Pattern Anal. Machine Intell.* **13** (2), 99–113.
9. C. Bouman, M. Shapiro (1994). A multiscale image model for Bayesian image segmentation. *IEEE Trans. Image Processing* **3** (2), 162–177.
10. B. Chalmond (1989). An iterative Gibbsian technique for reconstruction of M-ary images. *Pattern Recognition* **22** (6), 747–761.
11. R. Chellappa, A.K. Jain, editors (1993). *Markov random fields, theory and applications.* Academic Press, Boston.
12. F. Comets (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20**, 455–486.
13. F. Comets, B. Gidas (1992). Parameter estimation for Gibbs distribution from partially observed data. *Ann. Appl. Probab.* **2**, 142–170.
14. A. Dempster, N. Laird, D. Rubin (1977) . Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Series B* **39**, 1–38. with discussion.
15. H. Derin, H. Elliot (1987). Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. Pattern Anal. Machine Intell.* **9** (1), 39–55.
16. X. Descombes, R. Morris, J. Zerubia, M. Berthod (1996). Estimation of Markov random field prior parameter using Markov chain Monte Carlo maximum likelihood. Technical Report 3015, Inria.
17. D. Geman, G. Reynolds (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Machine Intell.* **14** (3), 367–383.
18. S. Geman, D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Ma-*

chine Intell. **6** (6), 721–741.
19. C. Geyer, E. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Royal Statist. Soc.* B **54** (3), 657–699.
20. J. Goodman, A.D. Sokal (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D.* **40** (6), 2035–2071.
21. C. Graffigne, F. Heitz, P. Pérez, F. Prêteux, M. Sigelle, J. Zérubia. Hierarchical and statistical models applied to image analysis, a review. submitted to *IEEE Trans. Inf. Theory* available at `ftp://gdr-isis.enst.fr/pub/publications/Rapports/GDR/it96.ps`.
22. X. Guyon (1993). *Champs aléatoires sur un réseau.* Masson, Paris.
23. W. Hackbusch (1985). *Multi-grid methods and applications.* Springer-Verlag, Berlin.
24. W. Hackbusch (1994). *Iterative solution of large sparse systems of equations.* Springer-Verlag, New-York.
25. J. M. Hammersley and D. C. Handscomb (1965). *Monte Carlo methods.* Methuen, London.
26. F. Heitz, P. Pérez, P. Bouthemy (1994). Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP : Image Understanding* **59** (1), 125–134.
27. P. Huber (1981). *Robust Statistics.* John Wiley & Sons, New York.
28. J.-M. Laferté, F. Heitz, P. Pérez, E. Fabre (1995). Hierarchical statistical models for the fusion of multiresolution image data. In *Proc. Int. Conf. Computer Vision* Cambridge, June.
29. S. Lauritzen (1996). *Graphical models.* Oxford Science Publications.
30. S.Z. Li (1995). *Markov random field modeling in computer vision.* Springer-Verlag, Tokyo.
31. M. Luettgen, W. Karl, A. Willsky (1994). Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Trans. Image Processing* **3** (1), 41–64.
32. J.L. Marroquin, S. Mitter, T. Poggio (1987). Probabilistic solution of ill-posed problems in computational vision. *J. American Statis. Assoc.* **82**, 76–89.
33. D. Mumford (1995). Bayesian rationale for the variational formulation. B. ter Haar Romeny, editor, *Geometry-driven diffusion in computer vision.* Kluwer Academic Publishers, Dordrecht, 135–146.
34. W. Pieczynski (1992). Statistical image segmentation. *Mach. Graphics Vision* **1** (1/2), 261–268.
35. A.D. Sokal (1989). Monte Carlo methods in statistical mechanics : foundations and new algorithms. Cours de troisième cycle de la physique en Suisse Romande.
36. R.H. Swendsen, J.-S. Wang (1987). Non-universal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.* **58**, 86–88.
37. R. Szeliski (1989). *Bayesian modeling of uncertainty in low-level vision.* Kluwer, Boston.
38. A.N. Tikhonov, V.Y. Arsenin (1977). *Solution of ill-posed problems.*

Winston, New York.
39. J. WHITTAKER (1990). *Graphical models in applied multivariate statistics.* Wiley, Chichester.
40. G. WINKLER (1995). *Image analysis, random fields and dynamic Monte Carlo methods.* Springer, Berlin.
41. L. YOUNES (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincaré - Probabilités et Statistiques* **24** (2), 269–294.
42. L. YOUNES (1989). Parametric inference for imperfectly observed Gibbsian fields. *Prob. Th. Rel. Fields* **82**, 625–645.