Statistical mechanics of image restoration

1995 J. Phys. A: Math. Gen. 28 511

(http://iopscience.iop.org/0305-4470/28/3/009)

J. Phys. A: Math. Gen. 28 (1995) 511–532. Printed in the UK

Statistical mechanics of image restoration

J M Pryce and A D Bruce

Department of Physics, The University of Edinburgh, Edinburgh EH9 3JZ, UK

Received 4 July 1994, in final form 27 October 1994

Abstract. We develop the statistical mechanics formulation of the image restoration problem, pioneered by Geman and Geman. Using Bayesian methods we establish the posterior probability distribution for restored images, for given data (corrupted image) and prior (assumptions about source and corruption process). In the simplest cases, studied here, the posterior is controlled by a cost function analagous to the configurational energy of an Ising model with local fields whose sense is defined by the data. Through a combination of Monte Carlo simulation and mean-field theory we address three key issues. First, we explore the sensitivity of the posterior distribution to the choice of prior parameters: we find phase transitions separating regions in which the distribution is effective (data-dominated) from regions in which it is ineffective (prior-dominated). Second, we examine the question of how best to *use* the posterior distribution to prescribe a single 'optimal' restored image: we argue that the mean of the posterior is, in general, to be preferred over the mode, both in principle and in practice. Finally, borrowing from Monte Carlo techniques for free-energy calculations, we address the question of prior parameter estimation within the 'evidence' framework of Gull and MacKay: our results suggest that parameters identified by this framework provide effective priors, leading to optimal restoration, only to the extent that the *forms* of the priors are well matched to the processes they claim to represent.

1. Introduction

Data reconstruction—the inference of underlying structure from experimental data—is one of the key problems in modern science. In the context of image restoration the problem is to find an estimate of an original picture from a corrupted version of that picture. The essential principles of this task are readily identified: it requires the synthesis of information supplied in the corrupted image with the information available (or assumed) about the source of the image and about the corruption process. The problem lends itself naturally to a Bayesian formulation (see e.g. [1]), which allows one to construct a probability distribution (posterior) for the original picture, on the basis of model distributions (priors) for source and corruption processes, and the image actually observed. Interest in the Bayesian approach to the problem has a long history (see e.g. [2–5]), and a more recent resurgence with the introduction of prior models based on discrete Markov random fields (MRF) [6, 7, 9–12]. In particular, following earlier work by Hammersley and Clifford [8], Geman and Geman [9] (hereafter referred to as GG) developed the analogy between the posterior distribution of a simple but generic MRF model and the Boltzmann–Gibbs distribution of a lattice Ising-like model. Others have since built on this work, drawing on Monte Carlo algorithms, pioneered in the statistical physics context, to explore MRF models by numerical simulation [13–15]. There has, however, been little attempt to apply the analytic methods of statistical mechanics to explore the image restoration process, and, in the absence of any systematic development

0305-4470/95/030511+22$19.50 © 1995 IOP Publishing Ltd

of the statistical mechanics perspective, many issues remain poorly understood. This is the motivation for the present work [16].

In the next section we set out the basic theory of the restoration process. We derive the posterior distribution from first principles using a prior on the density of edges alone, and recover the source posterior proposed by GG.

Using both analytic and simulation methods we investigate the sensitivity of the posterior (and thus the effectiveness of the restoration scheme) to the parametrization of the prior. Using a mean-field approximation we construct the phase diagram of the model in the space of prior parameters, and discover distinct data-dominated and prior-dominated phases, thereby illuminating the successes and failures of such restoration schemes [17, 11].

We then proceed to investigate how one may best *use* the posterior distribution to identify a *single*, in some sense *optimal*, estimate of the source picture [6, 18, 19]. We make a comprehensive comparison between the MAP estimate (maximum *a posteriori*, the mode of the posterior distribution), determined by simulated annealing, and the TPM (thresholded posterior mean) estimate [18, 11]. We illuminate the differences between the methods, and their strengths and weaknesses [17], by appeal to the mean-field phase diagram.

Finally we consider a possible method for assigning prior parameters, in the absence of which one has to operate on an *ad hoc* basis (see e.g. [9, 19]). There has been much work on parameter estimation (see e.g. [10, 20–22]), but little that only uses the single corrupted image [23]. We explore a generalized maximum likelihood formulation of this problem, expressed in the 'evidence' framework developed by Gull [24] and MacKay [25]. Neal [26] has recognized that the task of comparing the evidence for different parameter choices is analogous to that of comparing free energies in a statistical mechanics problem. Building on this perspective, we use Monte Carlo methods developed for free-energy studies to explore the utility of the evidence method applied to image restoration.

Two final remarks are in order on the style of the paper, in the light of the cross-disciplinary character of the background. First, the paper is written—in language, notation and concept—for the physics (specifically statistical physics) community. Second, while providing a selection of references to the background, we have endeavoured to make the paper largely self-contained.

2. Formulation of the image restoration problem

2.1. *The general framework*

The elements of the image restoration problem are summarized in figure 1. We envisage a source image $\boldsymbol{S}$ defined by a set of N binary pixel elements $\{S_1 \ldots S_N\}$, each with possible values ± 1. The source is drawn from some ensemble described by a source distribution $P(\boldsymbol{S})$. The source image is subject to a corruption process yielding a noisy image, described by a set of N pixel elements $\boldsymbol{D} \equiv \{D_1 \ldots D_N\}$; a particular source image $\boldsymbol{S}$ yields a particular corrupted image $\boldsymbol{D}$ with probability $P(\boldsymbol{D}|\boldsymbol{S})$. The source distribution $P(\boldsymbol{S})$ and the likelihood function $P(\boldsymbol{D}|\boldsymbol{S})$ together specify, stochastically, the process by which the corrupted image is *generated*. A knowledge of their forms is not to be supposed as available to the process by which the image is *restored*. Rather the restoration process must proceed on the basis of *models* of the forms of these functions, synthesising the constraints imposed by the models with the information provided by the 'data' $\boldsymbol{D}$ to infer a distribution $\tilde{P}(\boldsymbol{R}|\boldsymbol{D})$ of possible restored images $\boldsymbol{R}$. We denote the model functions by $\tilde{P}(\boldsymbol{S})$ and $\tilde{P}(\boldsymbol{D}|\boldsymbol{S})$. Then

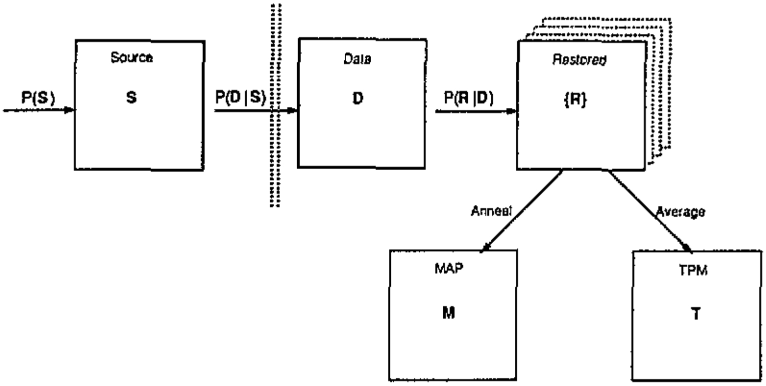

Figure 1. Elements of the image restoration problem. An image S appears on the source screen with a probability $P(S)$. A corrupted image D is generated with probability $P(D|S)$ and is displayed on the data screen. The restored screen displays an ensemble of images generated with probability $P(R|D)$. The thresholded mean of this ensemble provides the TPM estimator; its mode defines the MAP estimator.

Bayes theorem [1] identifies

$$\tilde{P}(S|D) = \frac{\tilde{P}(D|S)\tilde{P}(S)}{\tilde{P}(D)} \tag{2.1}$$

as the probability that a particular source image S underlies a specific given corrupted image D; the denominator of the right-hand side is defined in terms of the model functions by the normalization condition

$$\tilde{P}(D) = \sum_{\{S\}} \tilde{P}(D|S)\tilde{P}(S)\,. \tag{2.2}$$

The distribution of restored images may, indeed must, then be identified as

$$\tilde{P}(R|D) \equiv \tilde{P}(S|D)_{S\to R}\,. \tag{2.3}$$

A single (in some sense 'optimal') restored image may be identified from this distribution in any of a number of ways (figure 1).

Equations (2.1) and (2.3) provide a general framework for image restoration. In this work we explore their consequences in the context of explicit, simple, but non-trivial, assignments of the functions $P(S)$ and $P(D|S)$, and their model counterparts; these we now proceed to define.

2.2. *Specific forms*

The model functions $\tilde{P}(S)$ and $\tilde{P}(D|S)$ are supposed to express our hypotheses (what we know, or are prepared to believe) about the source of the image, and about the corruption process.

Consider first the model function $\tilde{P}(D|S)$. *We will adopt the hypothesis that the* corruption (image degradation) process is such that each pixel in the data differs from (is 'flipped' with respect to) its counterpart in the source with some probability $\tilde{q}$, which one may think of as expressing some 'noise level'. The model probability distribution for the data variable at site i, D_i, is then

$$\tilde{P}(D_i|S_i) = (1-\tilde{q})\delta_{D_i,S_i} + \tilde{q}\delta_{D_i,-S_i}\,. \tag{2.4}$$

Defining $\tilde{h}$ such that

$$\tilde{h} \equiv \tfrac{1}{2} \ln \left(\frac{1}{\tilde{q}} - 1 \right) \tag{2.5}$$

we have

$$\tilde{q} \equiv \frac{\mathrm{e}^{-\tilde{h}}}{\mathrm{e}^{+\tilde{h}} + \mathrm{e}^{-\tilde{h}}}$$

so that

$$\tilde{P}(D_i|S_i) = \frac{\mathrm{e}^{\tilde{h}D_iS_i}}{z_l(\tilde{h})}$$

where

$$z_l(\tilde{h}) = \mathrm{e}^{\tilde{h}} + \mathrm{e}^{-\tilde{h}} .$$

The corruption process is supposed to be site-independent, so the model likelihood function follows as

$$\tilde{P}(\boldsymbol{D}|\boldsymbol{S}) = \tilde{P}(\boldsymbol{D}|\boldsymbol{S}, \tilde{h}) \equiv \frac{1}{Z_l(\tilde{h})} \exp \left\{ \tilde{h} \sum_i D_i S_i \right\} \tag{2.6}$$

where $Z_l(\tilde{h})$ is determined by the normalization condition

$$Z_l(\tilde{h}) = \left[z_l(\tilde{h}) \right]^N = \sum_{\{\boldsymbol{D}\}} \exp \left\{ \tilde{h} \sum_i D_i S_i \right\} . \tag{2.7}$$

Now consider the model function $\tilde{P}(\boldsymbol{S})$. Suppose we have some prior knowledge about the frequency with which 'edges' tend to appear in the source, where by 'edge' we mean a pair of pixels, at neighbouring sites i and j, that are in opposite states (so that $S_i S_j = -1$). Specifically we adopt the hypothesis that edges appear with some density which we shall denote by $\tilde{\epsilon}_S$, a guess at some underlying 'true' value, ϵ_S. Then, appealing to standard information-theoretic arguments (see e.g. [27]), the model prior follows as

$$\tilde{P}(\boldsymbol{S}) = \tilde{P}(\boldsymbol{S}|\tilde{K}) \equiv \frac{1}{Z_p(\tilde{K})} \exp \left\{ \tilde{K} \sum_{\langle ij \rangle} S_i S_j \right\} \tag{2.8}$$

where $\sum_{\langle ij \rangle}$ denotes a sum extending over all pairs of neighbouring sites on the pixel lattice. The value of the coupling $\tilde{K}$ is specified implicitly in terms of the edge density $\tilde{\epsilon}_S$ by the constraint

$$\left\langle \sum_{\langle ij \rangle} S_i S_j \right\rangle_S = \frac{\partial \ln Z_p(\tilde{K})}{\partial \tilde{K}} = \tfrac{1}{2} \nu N (1 - 2\tilde{\epsilon}_S) \tag{2.9}$$

where ν is the number of nearest neighbours of each pixel, and $\langle \cdot \rangle_S$ denotes average with respect to the distribution (2.8). The normalization factor for this distribution is

$$Z_p(\tilde{K}) = \sum_{\{\boldsymbol{S}\}} \exp \left\{ \tilde{K} \sum_{\langle ij \rangle} S_i S_j \right\} . \tag{2.10}$$

Equations (2.6) and (2.8) together define the form to be assigned to the distribution of restored images. Appealing to (2.1) and (2.3) we find

$$\tilde{P}(\boldsymbol{R}|\boldsymbol{D}) = \frac{1}{Z(\tilde{K}, \tilde{h}; \boldsymbol{D})} \exp\{-\mathcal{H}\} \tag{2.11a}$$

where

$$\mathcal{H} = -\tilde{K} \sum_{\langle ij \rangle} R_i R_j - \tilde{h} \sum_i R_i D_i \tag{2.11b}$$

and

$$Z(\tilde{K}, \tilde{h}; \boldsymbol{D}) = \sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}\} \;. \tag{2.11c}$$

Equation (2.11*a*) defines an ensemble of images whose relative likelihood is controlled by the cost function (configurational energy) (2.11*b*), which is equivalent to the cost function used by GG in [9], except that we have neglected the line processes introduced there, and have chosen to restrict our analysis to the case of binary variables. The cost function is just that of a spin-$\frac{1}{2}$ Ising model with nearest-neighbour coupling, and a site-dependent field. The role of the two terms is clear. The term controlled by $\tilde{h}$ binds the restored configuration $\boldsymbol{R}$ to the data $\boldsymbol{D}$, the binding being stronger the smaller the value assigned to the noise-level $\tilde{q}$ (equation (2.5)). The term controlled by $\tilde{K}$ penalizes edges in the restored configuration, to an extent which reflects prior convictions about the edge density (equation (2.9)). The character of the restored ensemble is controlled by the competition between these two terms.

Now consider the functions $P(\boldsymbol{S})$ and $P(\boldsymbol{D}|\boldsymbol{S})$. Though not to be used in the *restoration* process itself, the forms of these functions have nevertheless to be prescribed in order to *generate* the data on which the restoration process is to be tested.

We choose the corruption process to be of the form assumed in the modelling process, but with a noise level q, which may differ from $\tilde{q}$. Thus (cf equation (2.6))

$$P(\boldsymbol{D}|\boldsymbol{S}) = \frac{1}{Z_l(h)} \exp\left\{h \sum_i D_i S_i\right\} \tag{2.12}$$

where, now, $h \equiv \frac{1}{2}\ln(1/q - 1)$. Varying the values assigned to h and $\tilde{h}$ (the most convenient parametrizations of the true and model noise levels) provides one way of exploring the sensitivity of the restoration procedure to the accuracy of the models.

We have chosen to explore two simple forms of source distribution.

To explore the situation in which $\tilde{P}(\boldsymbol{S})$ and $P(\boldsymbol{S})$ are at least potentially *well matched* we have examined the case in which source images are drawn from the distribution with the same (2D square lattice Ising) structure as that of the model function (2.8)

$$P(\boldsymbol{S}) = \frac{1}{Z_p(K)} \exp\left\{K \sum_{\langle ij \rangle} S_i S_j\right\} \tag{2.13}$$

but characterized by a coupling constant K, at our disposal.

To explore the case in which $\tilde{P}(\boldsymbol{S})$ and $P(\boldsymbol{S})$ are *ill matched* we have also studied the case where the source is some uniquely defined image $\boldsymbol{S}^{\mathrm{o}}$ so that

$$P(\boldsymbol{S}) = \delta_{\boldsymbol{S},\boldsymbol{S}^{\mathrm{o}}} \;. \tag{2.14}$$

Where a specific form is necessary, we have chosen the source image $\boldsymbol{S}^{\mathrm{o}}$ to be that of a $N \times N$ chequerboard with chequers of side m pixels, with m of the form $m = 2^c$. This choice provides a source image whose spatial structure and edge density are controlled, simply, by a single parameter.

3. Prelude: Monte Carlo studies

In the preceding section we have set out the elements of the Bayesian view of image restoration. The key result is (2.11*a*), which synthesizes the knowledge encoded in the data with that expressed in prior convictions about the source and the corruption process to give the probability distribution (the posterior) of possible restored images. In order to identify a *typical* or *optimal* restored image (i.e. in order to implement a *restoration procedure*) one must find a way of sampling from this distribution. As recognised by GG, Monte Carlo (MC) methods, widely used in exploring the statistical mechanics of condensed matter systems (see e.g. [28]) provide an easily realisable way of implementing this sampling, and indeed, the MC method has been widely used in this context (see e.g. [14, 9, 13]). In this section we use MC methods to explore the ensemble of restored images. Our aim is to expose a number of issues to which we shall subsequently give more systematic attention, notably

- how the quality of the posterior distribution depends upon the effectiveness of the underlying modelling process;
- how the posterior distribution should be used to estimate the true image;
- how the model parameters should be set.

To provide some qualitatively helpful initial insights consider figure 2 which shows two sets of images, associated with two model sources.

The upper row of figures, (*a*)–(*e*), are associated with a *source* ('true', uncorrupted) image, (*a*), drawn from an Ising ensemble (equation (2.13)) with chosen fixed coupling ($\tanh K = 0.42$). The lower row of figures is associated with a chequerboard source, (*a*), with chequers of size 16×16 pixels. All figures comprise a total of 128×128 pixels.

The second figure, (*b*), in each row shows the *data* (the input to the restoration procedure) generated from the associated source image by random flipping of pixels with chosen fixed probability ($q = 0.3$ for the Ising source; $q = 0.4$ for the chequerboard source).

The third figure, (*c*), in each row shows a *typical* restored image, drawn from the ensemble defined by the posterior distribution (2.11*a*), explored by MC methods. Specifically we have used the 'single spin-flip' Metropolis algorithm [29] in which a pixel is selected at random and flipped with probability min $[1, \mathrm{e}^{-\delta\mathcal{H}}]$, where $\delta\mathcal{H}$ is the change in the cost function (2.11*b*) entailed in the pixel flip. The starting configuration for the algorithm is, in each case, the associated corrupted image. In the case of the Ising source the parameters of the posterior are assigned to match the source and noise parameters ($\tilde{K} = K$, $\tilde{h} = h$). In the case of the chequerboard source, the posterior parameters are assigned to match the noise ($\tilde{h} = h$), and to match the density of edges in the source ($\tilde{K} = K_{\mathrm{eff}}$, where K_{eff} is the Ising coupling which would generate configurations with a density of edges equal to that of the chequerboard).

Finally, columns (*d*) and (*e*) show two ways in which the information inherent in the *ensemble* of restored images may be synthesised to yield a *single* estimator of the true image. The figures in column (*d*) show the thresholded posterior mean (TPM) estimate, the binary image closest to the average over the posterior distribution. Column (*e*) shows the maximum *a posteriori* (MAP) estimate, the single most probable image in the posterior distribution. These examples show that the two estimators *may* yield very different results. This is the motivation for the detailed comparison of the two presented in section 5.

Of course, irrespective of the estimator chosen, the quality of the final restored image is limited by the quality of the posterior distribution itself, which reflects both its *structure* and its *parametrization*. The parameter assignments underlying the examples shown in figure 2 presuppose knowledge other than that presented directly in the data itself. It is

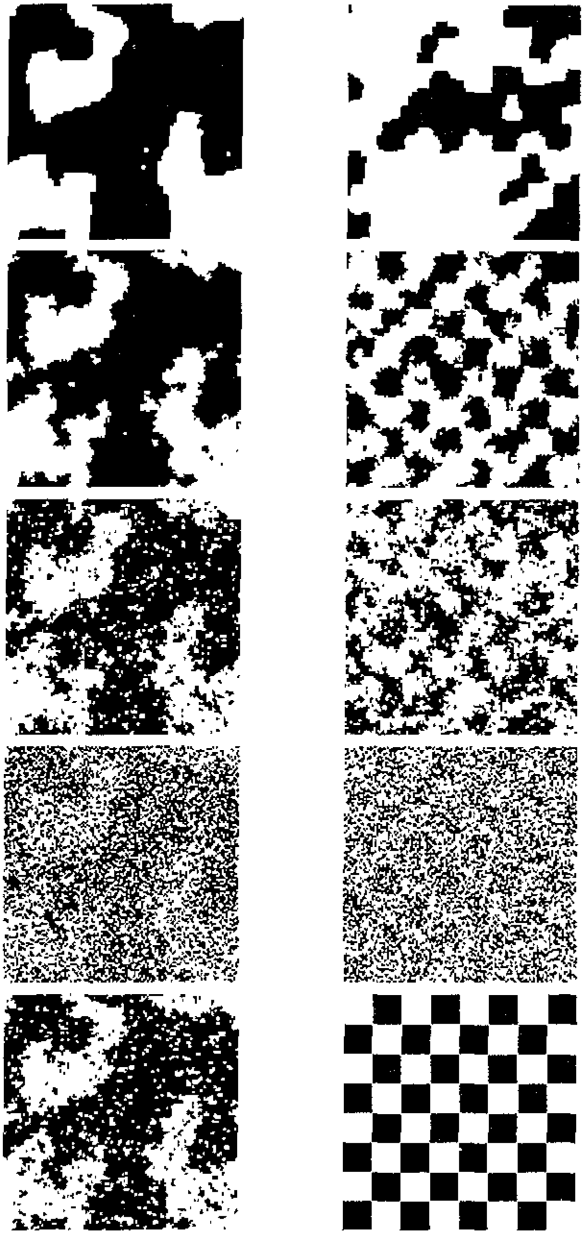

Figure 2. The stages of the restoration procedure as applied (upper) to an Ising source and (lower) to a chequerboard source: (*a*) the source image, (*b*) the corrupted image, (*c*) a representative of the restored ensemble, (*d*) the TPM estimator and (*e*) the MAP estimator.

clearly of interest to establish, quantitatively and systematically, how sensitive the posterior distribution is to these choices. To that end we introduce a *quality factor*, defined by

$$Q \overset{\text{def}}{=} \frac{\langle d[D, \langle S[D]\rangle_S] - d[\langle R[D]\rangle_R, \langle S[D]\rangle_S]\rangle_D}{\langle d[D, \langle S[D]\rangle_S]\rangle_D} \tag{3.1}$$

where

$$d\,[A, B] = \frac{1}{N}\sum_k [A_k - B_k]^2$$

while

$$\langle S[D]\rangle_S \stackrel{\mathrm{def}}{=} \sum_{\{S\}} P(S|D)\cdot S \qquad \text{and} \qquad \langle R[D]\rangle_R \stackrel{\mathrm{def}}{=} \sum_{\{R\}} \tilde{P}(R|D)\cdot R .$$

The quality factor is defined as a measure of how close the chosen posterior distribution $\tilde{P}(R|D)$ lies to the 'true posterior' $P(S|D)$ implied by full knowledge of the source function $P(S)$ and the corruption process $P(D|S)$. Its attributes are consistent with this role. Thus Q is normalized to unity when the model posterior coincides with the true posterior (in which case the moments $\langle S[D]\rangle_S$ and $\langle R[D]\rangle_R$ are equal); it is positive only if the first moment of the model posterior represents an improvement on the data, in the sense that it lies *closer* (than the data) to the first moment of the true posterior.

We have explored the sensitivity of the quality factor to the choice of model parameters through extensive Monte Carlo studies; the key results are summarized in figures 3 and 4.

Figure 3 shows the quality factor for the restoration of the data generated from the Ising source (for particular K and h), over the space of model parameters $\tilde{K}$ and $\tilde{h}$. Three features are noteworthy. First, as one would expect, the quality factor is optimized when the posterior parameters match those of the source ($\tilde{K} = K$) and noise ($\tilde{h} = h$), which is possible in this case where the space of model functions contains the underlying reality ($\tilde{P}(S)$ and $P(S)$ are 'well matched'). Second, one sees from the shape of Q-factor contours around the point of optimal Q that there is an interplay between the effects of the two model parameters: thus if $\tilde{K}$ is assigned a value other than K, the value of $\tilde{h}$ securing optimal Q is no longer $\tilde{h} = h$. Third, and most striking, one observes a boundary in the space of model parameters, marked by a steep variation of the quality factor. The boundary separates regions in which the posterior is effective (characterized by 'high' Q) from regions (of 'low' or negative Q) where it is ineffective.

Corresponding features are evident in the behaviour of the Q factor for chequerboard data. Figure 4 shows the results for a variety of chequerboards (of different chequer size, and thus different edge density) and noise levels. In this case the model parameter space

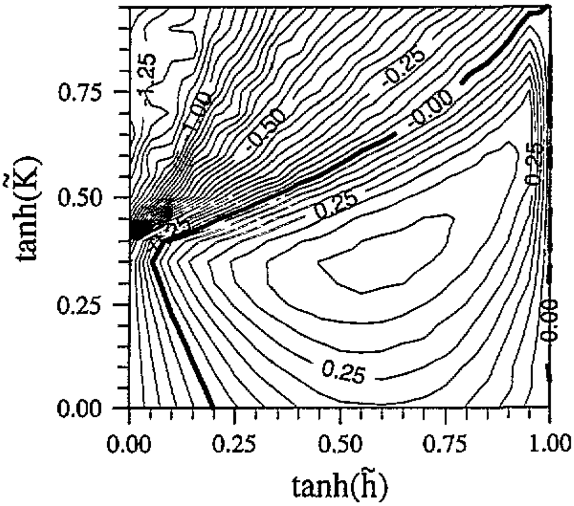

Figure 3. Simulation results for the well matched prior: contour plot of the quality factor, as a function of the restoration parameters, for an Ising source with density of edges $\varepsilon_S = 0.25$ and noise $q = 0.2$ ($\tanh(K) \simeq 0.36$, $\tanh(h) = 0.6$).

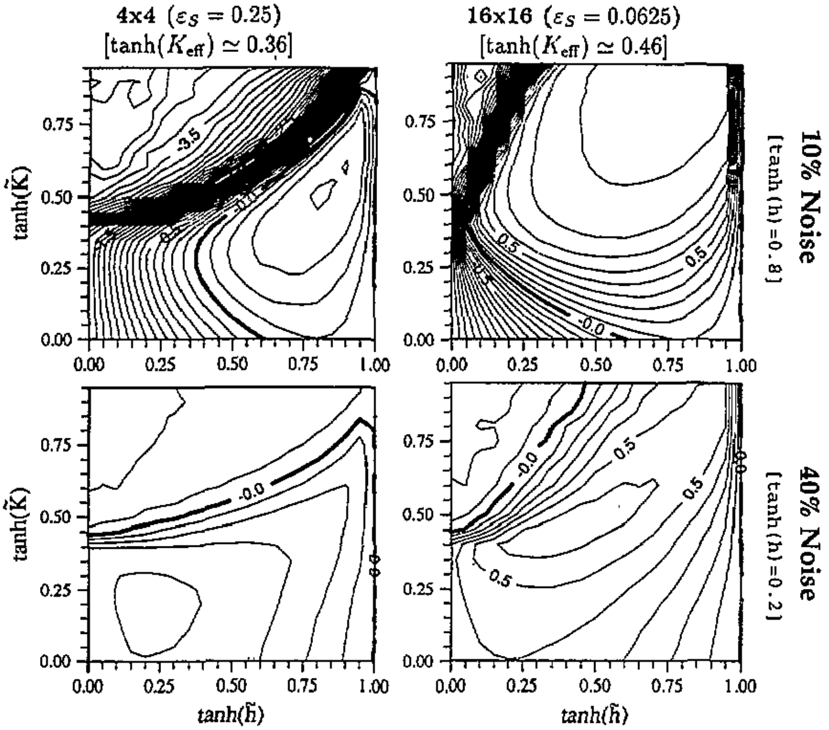

Figure 4. Simulation results for the ill matched prior: contour plot of the quality factor, as a function of the restoration parameters, for chequerboards with various chequer sizes (density of edges ϵ_S) and noise levels (q).

does not contain the data generator: it is impossible to choose posterior parameters that exactly 'match' the generation process ($\tilde{P}(S)$ and $P(S)$ are 'ill matched'). Moreover, the assignment (based on the density of edges in the source) $\tilde{K} = K_{\rm eff}$ turns out, in general, *not* to be Q-optimal: the results show that the Q-optimal values of *both* $\tilde{K}$ and $\tilde{h}$ depend upon the density of edges (the chequer size) *and* the noise level. Again the results suggest that there are distinct regions of effective and ineffective posteriors.

These observations provide the motivation and context for the other parts of the programme reported here. The interplay between the model parameters apparent in the Q-factor features in section 6 where we address the question of how model posterior parameters should be assigned *on the basis of the corrupted data alone*. (The Q-factor, of course, folds in information about the source!). The performance-boundaries displayed by the Q-factor are the focus of attention in the next section where we shall see that they are manifestations of phase transitions.

4. Assessing the posterior: prior-data competition

In this section we explore the phase structure ('thermodynamics') associated with the ensemble of restored variables $\boldsymbol{R}$, given a model posterior built from an edge density prior, and a Gaussian-noise likelihood. Our aim is to explore how the phase structure depends upon the parameters $\tilde{h}$ and $\tilde{K}$ modelling the corruption process and the source, which for the purposes of this calculation we shall take to be some fixed image S^0. The phase structure is controlled by the partition function (2.11*c*), and its dependence upon certain key macroscopic properties ('order parameters') to be identified below. The partition function is hard to evaluate exactly, because of the quenched disorder associated with the data $\boldsymbol{D}$. We proceed, rather, within a mean-field approximation which, we shall see, illuminates the essential features displayed by our Monte Carlo studies.

We shall utilize the variational formulation of the theory: the technique is well known (see e.g. [30]) and we shall describe the calculation in outline only.

We introduce a variational representation of the partition function of interest (equation (2.11*c*)):

$$Z_{\mathrm{V}} \overset{\mathrm{def}}{=} \sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}_{\mathrm{V}}\} \tag{4.1a}$$

where

$$\mathcal{H}_{\mathrm{V}} = \sum_i R_i\big[\tilde{h}D_i + \tfrac{1}{2}H^+(1+S_i) + \tfrac{1}{2}H^-(1-S_i)\big]\,. \tag{4.1b}$$

The fields H^+ and H^- are introduced with a view to the order parameters defined below; they are conjugate to the restored variables at sites at which the source variable has value $+1$ and -1, respectively. The true partition function may then be recast in the form

$$\begin{aligned} Z &= \sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}\} \\ &= \sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}_{\mathrm{V}}\}\, \frac{\sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}_{\mathrm{V}} + [\mathcal{H}_{\mathrm{V}} - \mathcal{H}]\}}{\sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}_{\mathrm{V}}\}} \\ &\equiv Z_{\mathrm{V}}\, \langle \exp\{\mathcal{H}_{\mathrm{V}} - \mathcal{H}\}\rangle_{\mathcal{H}_{\mathrm{V}}} \end{aligned} \tag{4.2}$$

where

$$\langle \cdot \rangle_{\mathcal{H}_{\mathrm{V}}} \overset{\mathrm{def}}{=} \frac{\sum_{\{\boldsymbol{R}\}} \cdot \exp\{-\mathcal{H}_{\mathrm{V}}\}}{\sum_{\{\boldsymbol{R}\}} \exp\{-\mathcal{H}_{\mathrm{V}}\}}\,.$$

The identity (4.2) motivates the mean-field approximation

$$Z \simeq Z_{\mathrm{MF}} \equiv Z_{\mathrm{V}} \exp\{\langle \mathcal{H}_{\mathrm{V}} - \mathcal{H}\rangle_{\mathcal{H}_{\mathrm{V}}}\}\,. \tag{4.3}$$

Since the variational cost function (4.1*b*) comprises a sum of independent (single-site) terms it is straightforward to implement the averages in (4.3). After some calculation we find

$$\begin{aligned} f_{\mathrm{MF}} \overset{\mathrm{def}}{=} -\frac{1}{N}\ln Z_{\mathrm{MF}} = &-\tfrac{1}{2}(1-q)\ln\cosh\{\nu\tilde{K}[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] + \tilde{h}\} \\ &-\tfrac{1}{2}q\ln\cosh\{\nu\tilde{K}[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] - \tilde{h}\} \\ &-\tfrac{1}{2}(1-q)\ln\cosh\{\nu\tilde{K}[(1-\varepsilon_S)R^- + \varepsilon_S R^+] - \tilde{h}\} \\ &-\tfrac{1}{2}q\ln\cosh\{\nu\tilde{K}[(1-\varepsilon_S)R^- + \varepsilon_S R^+] + \tilde{h}\} \\ &-\tfrac{1}{2}\tilde{K}\nu\big[\tfrac{1}{2}(1-\varepsilon_S)\big(R^{+2} + R^{-2}\big) + \varepsilon_S R^+R^-\big] \\ &+\tfrac{1}{2}\tilde{K}\nu R^+[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] + \tfrac{1}{2}\tilde{K}\nu R^-[(1-\varepsilon_S)R^- + \varepsilon_S R^+] - \ln 2\,. \end{aligned} \tag{4.4}$$

We have used the variational principle to eliminate the fields H^+ and H^- in favour of the order parameters

$$\begin{aligned} R^+ &\overset{\mathrm{def}}{=} \frac{1}{N}\sum_i (1+S_i)\,\langle R_i\rangle_{\mathrm{V}} \\ &= (1-q)\tanh(H^+ + \tilde{h}) + q\tanh(H^+ - \tilde{h}) \end{aligned} \tag{4.5a}$$

$$R^- \overset{\text{def}}{=} \frac{1}{N}\sum_i (1 - S_i)\,\langle R_i\rangle_{\mathrm{V}}$$
$$= (1-q)\tanh(H^- - \tilde{h}) + q\tanh(H^- + \tilde{h}) \qquad (4.5b)$$

while the source S is entirely parametrized by the (true) density of edges ϵ_S. The dependence upon the specific realization of the data D self-averages out in the course of the calculation, so that 4.4 serves as an approximation to the full quenched average:

$$\mathcal{F} \equiv -\frac{1}{N}\,\langle\langle \ln Z\rangle\rangle_D \simeq f_{\mathrm{MF}}\,.$$

The minima of this free energy in the space spanned by the order parameters R^+ and R^- are located by the solutions to the coupled equations

$$R^+ = (1-q)\tanh\{\tilde{K}\nu[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] + \tilde{h}\}$$
$$+\, q\tanh\{\tilde{K}\nu[(1-\varepsilon_S)R^+ + \varepsilon_S R^-] - \tilde{h}\} \qquad (4.6a)$$
$$R^- = (1-q)\tanh\{\tilde{K}\nu[(1-\varepsilon_S)R^- + \varepsilon_S R^+] - \tilde{h}\}$$
$$+\, q\tanh\{\tilde{K}\nu[(1-\varepsilon_S)R^- + \varepsilon_S R^+] + \tilde{h}\}\,. \qquad (4.6b)$$

The physical character of the minima is most naturally expressed through linear combinations of the two order parameters:

$$\mathcal{O} \overset{\text{def}}{=} \frac{1}{N}\sum_i \langle\langle R_i[D]\rangle_R\, S_i\rangle_D = \tfrac{1}{2}[R^+ - R^-] \qquad (4.7a)$$

and

$$\mathcal{M} \overset{\text{def}}{=} \frac{1}{N}\sum_i \langle\langle R_i[D]\rangle_R\rangle_D = \tfrac{1}{2}[R^+ + R^-]\,. \qquad (4.7b)$$

We call $\mathcal{O}$ the *overlap*: it measures the correlation between the mean restored image and the source, a correlation which is mediated entirely through the *data* (since the source itself does not appear explicitly in the posterior!) We call $\mathcal{M}$ the *bias*: a non-zero bias is attributable entirely to the influence of the prior (in particular the 'ferromagnetic' behaviour promoted by the edge-suppressing coupling) since our sources have no intrinsic bias. With these remarks in mind we shall call solutions with non-zero $\mathcal{M}$ *prior-like* and solutions with zero $\mathcal{M}$ *data-like*. Prior-like solutions typically have small (but not necessarily zero) overlap $\mathcal{O}$ (cf figure 7).

Figure 5 shows contours of the free energy surfaces defined by (4.4), in the space of the order parameters $R^\pm$, for a selection of couplings $\tilde{K}$ and $\tilde{h}$ (and with fixed but essentially arbitrary values of the data-generating parameters q and ϵ_S). The turning points in this surface correspond to the solutions of the coupled equations (4.6*a*) and (4.6*b*). The character of the surface changes qualitatively according to the values of $\tilde{K}$ and $\tilde{h}$. If we classify each point according to the number of local minima and the character of the global minimum, we generate the phase diagram shown in figure 6. Although its details reflect the values chosen for the data-generating parameters, the general structure of this phase diagram is typical of a wide range of parameters. Its essential character is intelligible on the basis of the limiting behaviour associated with its bounding axes.

Along the right-hand boundary ($\tilde{h} \to \infty$) the solution is trivially data-like: the restored image is fully bound to the data, implying zero $\mathcal{M}$ and non-zero $\mathcal{O} = 1 - 2q$. Along the lower boundary ($\tilde{K} = 0$) the solution is also data-like, but with overlap $(1-2q)\tanh(\tilde{h})$, which is *less* than that between data and source: *on average* the 'restoration' procedure actually *degrades* the image in this regime. Along the upper boundary ($\tilde{K} \to \infty$) the

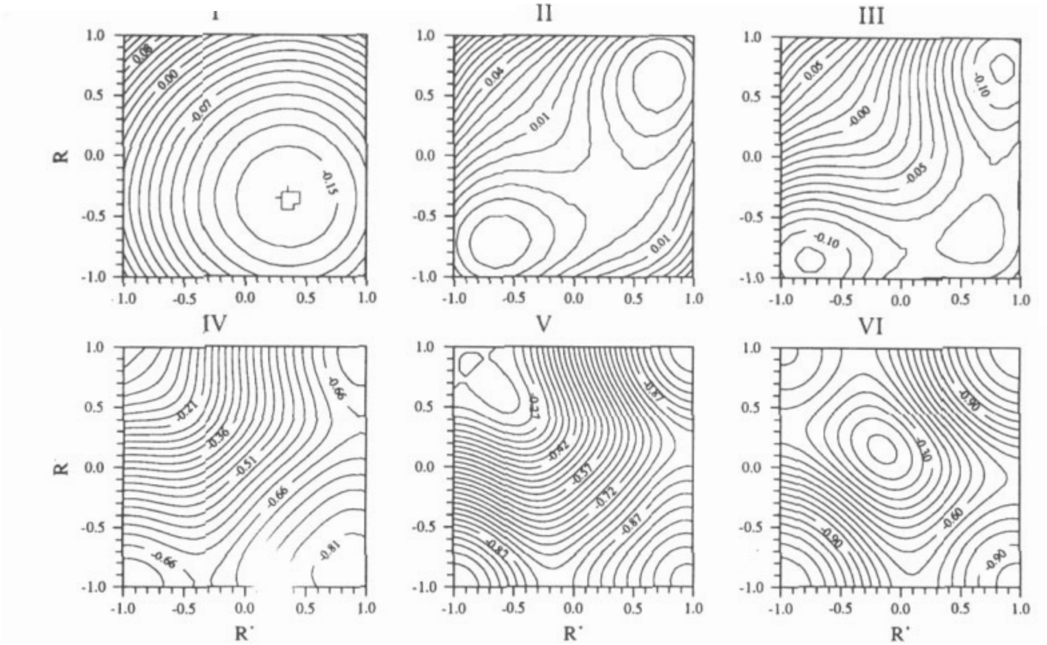

Figure 5. Contours of the mean-field free energy density $\tilde{f}_{\rm MF} \equiv f_{\rm MF} + \ln 2$, where $f_{\rm MF}$ is given by (4.4). The underlying data-generating parameters are $\epsilon_S = 0.125$ and $q = 0.3$. The six sets of data are each representative of one point in the space of restoration parameters $\tilde{K}, \tilde{h}$, drawn from each of the regions I–VI, depicted in the phase diagram (figure 6). In regions I, IV and V the global minimum is data-like; in the remaining regions the global minimum is prior-like.

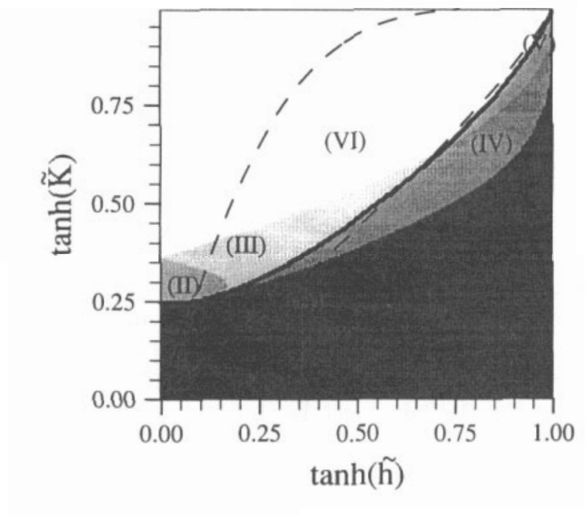

Figure 6. The mean-field phase diagram with $\epsilon_S = 0.125$ and $q = 0.3$. The full curve separates data-like and prior-like phases, each of which is subdivided into three regions, distinguished by the number and character of metastable states, as detailed in table 1. The broken curves identify possible paths followed in the course of MAP annealing schedules, as discussed in section 5.

solutions are prior-like, with zero overlap and $\mathcal{M} = \pm 1$, corresponding to the two edge-free single-colour 'ground-states' of the prior. Finally, the significant feature of the left-hand boundary ($\tilde{h} \to 0$) is a phase transition between zero bias and non-zero bias phases at $\tilde{K} = 1/\nu = 1/4$ ($\tanh \tilde{K} \simeq 0.25$), the mean-field critical coupling of the 2D Ising model. From this point there emerges a phase boundary (shown as a full curve in figure 6) which

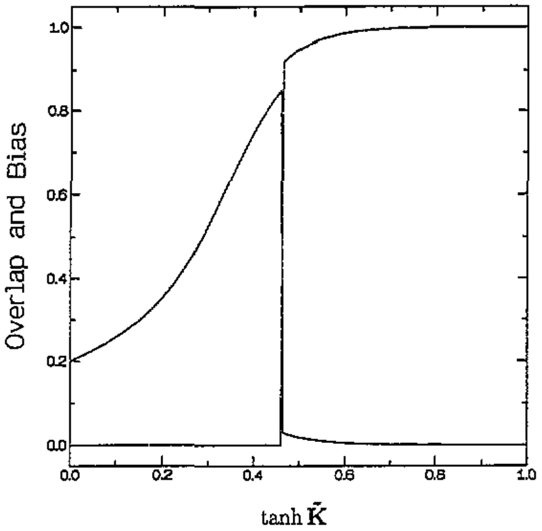

Figure 7. The mean-field predictions for the behaviour of the overlap $\mathcal{O}$ (full curve) and bias $\mathcal{M}$ (broken curve), for source parameters $\epsilon_S = 0.125$ and $q = 0.3$ (as in the phase diagram, 6), and with fixed restoration parameter $\tanh \tilde{h} = 0.5$.

Table 1. The numbers in the first column identify the six regions of the phase diagram shown in figure 6. The second column identifies the number and character of the associated free energy minima, which fall into three categories: *data-like* with $\mathcal{O} \neq 0$, $\mathcal{M} = 0$; *prior-like* with $\mathcal{O} = 0$, $\mathcal{M} \neq 0$; *anti-data-like* with $\mathcal{O}$ *negative*, $\mathcal{M} = 0$. In each case the global minimum is in italics. Note that region II (alone) has *no* data-like minima: a restoration scheme cannot work in this region. Our studies show that the size of this region grows with the degree of difficulty presented by the restoration problem, which increases with increasing q and ϵ_S.

Region number	Free energy minima
I	*1 data*
II	*2 prior*
III	*2 prior*, 1 data
IV	2 prior, *1 data*
V	2 prior, *1 data*, 1 anti-data
VI	*2 prior*, 1 data, 1 anti-data

divides the phase diagram into two phases, a data-like phase below the phase boundary, and a prior-like phase above it. Except in the $\tilde{h} \to 0$ limit, the data-prior phase transition resulting when this phase boundary is crossed is first-order, involving discontinuous changes in the order parameters. Figure 7 shows the behaviour for a typical value of $\tanh \tilde{h}$.

The data-like and prior-like phases are subdivided into a number of different regions, distinguished by the number and character of the metastable free energy minima they display in addition to the global minima which define their equilibrium behaviour (table 1). The boundaries of the three regions (I, IV and V) making up the data-like phase and the boundaries of the remaining three regions (II, III and VI) which form the prior-like phase do not signal phase transitions as such, but existence boundaries for metastable states. Thus, for example, the metastable prior-like solutions present in region IV disappear at the boundary with region I.

The existence of the phase boundary is reflected in the behaviour of the Q-factor which may also be calculated within the mean-field framework. When, as we envisage here, the source consists of a single specific image S^0 (the true source distribution is of the form

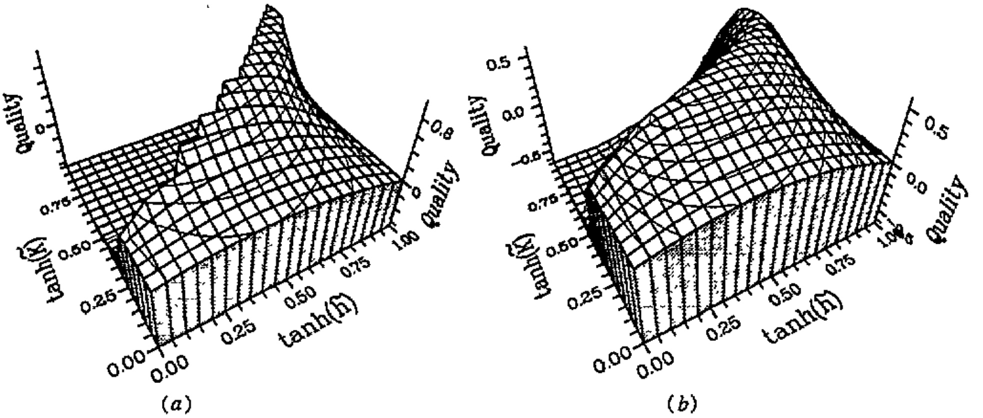

Figure 8. Comparison of mean field and simulation results for the quality factor Q. (a) Results for the mean-field calculation with parameters $\epsilon_S = 0.125$ and $q = 0.3$. (b) Simulation results for the corresponding 8×8 chequerboard.

(2.14)) the elements of the Q-factor (equation (3.1)) can be written as

$$\langle d[D, \langle S[D]\rangle_S]\rangle_D = \langle d[D, S^0]\rangle_D = 4q \tag{4.8a}$$

and

$$\langle d[\langle R[D]\rangle_R, \langle S[D]\rangle_S]\rangle_D = 1 - 2O + \frac{1}{N}\sum_i \left\langle \langle R_i[D]\rangle_R^2 \right\rangle_D . \tag{4.8b}$$

The final term in this equation can be expressed (through the auxiliary fields H^+ and H^-) in terms of R^+ and R^-, thence providing a complete specification of Q within the mean-field approximation. The results are displayed in figure 8 alongside the results of Monte Carlo studies. In the mean-field calculations (figure 8(a)), the parameters q and ϵ_S (which fully parameterize the source) are assigned values matching those of the source actually used in the Monte Carlo calculation (an 8×8 chequerboard). The mean-field calculation reproduces the observed structure rather well, qualitatively at least. Most significantly, perhaps, it shows that the catastrophic failure of the restoration scheme in some regions of parameter space is a manifestation of the underlying phase transition: if the restoration parameters are chosen to lie in the prior-dominated phase, the edge-suppressing coupling in the prior overwhelms the field binding the image to the data and the ensemble of restored images is largely edge-free and uncorrelated with the source. Closer inspection shows that, not unexpectedly, the mean-field analysis has its limitations. Of course, it misplaces the phase boundary somewhat: the mean-field approximation underestimates the critical coupling of the 2D Ising model by almost a factor of 2. More significantly, however, the discontinuity at the phase boundary in the mean-field calculation has as its counterpart in the simulations (figure 8(b)) a steep but continuous drop in the Q-factor. This difference can be understood, qualitatively, within the wider mean-field framework: we attribute it to the influence of long-lived states, bound to the data, the 'real' counterparts of the metastable data-like solutions which the mean-field calculation finds in the prior-like phase. Simulations initiated from the data configuration D may remain trapped in such states for the time span of the simulation. In fact, the mean-field calculations suggest that the Q-factor may actually continue to increase across the phase boundary if the metastable states are tracked; the simulations do not illuminate this issue, since they do not allow one to determine the precise location of the underlying phase boundary.

5. Utilizing the posterior: MAP versus TPM

Thus far we have focused on the quality of the model posterior distribution itself. We now turn to consider the two principal ways in which the posterior may be *used* to identify a single binary image that is in some sense optimal.

The *thresholded posterior mean* (TPM) estimator is defined by

$$T_k = \mathrm{sgn}\,\{\langle \mathrm{R}_k[D]\rangle_{\mathrm{R}}\}\ . \tag{5.1}$$

It is 'optimal' in the sense that (see e.g. [11]) it is the binary image whose overlap with the source has maximum expectation value (minimum mean-square bitwise error), given the data *and the modelling assumptions.*

The *maximum a posteriori* (MAP) estimator is defined by

$$M_k = R_k^{\mathrm{map}} \qquad \text{where} \quad \tilde{P}(\boldsymbol{R}^{\mathrm{map}}|\boldsymbol{D}) = \sup_{\{\boldsymbol{R}\}} \tilde{P}(\boldsymbol{R}|\boldsymbol{D})\,. \tag{5.2}$$

It is optimal in the sense that it identifies the single most likely binary image, given the data and the modelling assumptions.

The original GG paper [9] focused on the MAP estimator; much subsequent work has followed suit (see e.g. [31, 14, 10, 13, 32, 22]). There is only a small body of work that begins to recognize the utility of the TPM estimator [19, 33, 18, 11]. A systematic comparison of their merits seems overdue.

The first point to make is that the task of finding the MAP estimator is vastly more computationally intensive than is that of finding the TPM estimator. To determine the TPM (equation (5.1)) requires only† a direct sampling of the model posterior distribution. To determine the MAP estimator (equation (5.2)) requires that one finds the minimum of the cost function (ground state of the configurational energy) $\mathcal{H}$ (equation (2.11*b*)). *If fully implemented,* this is a computationally demanding task given the complexity of the landscape and the danger (although we shall see that it should not necessarily be seen as such) of the search process being trapped in local minima. This risk can be reduced by using simulated annealing [34]. Indeed, GG [9] present a proof that, with a suitable annealing schedule, simulated annealing is guaranteed to find the global minimum and hence the exact MAP estimator. However, such a schedule would take a prohibitive length of time to complete (the total number of site updates required is exponential in the system size N). They claim that acceptable results are obtained using a logarithmic schedule but even with this faster schedule the annealing process is still far more computationally intensive than the calculation of the TPM estimator.

Next let us consider the relative quality of the two estimators. We know *a priori* that the TPM estimator is *guaranteed* to give maximal overlap with the source in the idealized case in which the model posterior matches the true posterior‡.

This claim is quantified in the results presented in figure 9(*a*), which shows the overlap of TPM with the source, less the overlap of MAP with the source, for a range of Ising source data-generating parameters (K and h) with restoration parameters chosen to match ($\tilde{K} = K$ and $\tilde{h} = h$). The difference is non-negative: the MAP estimator never beats TPM for any choice of parameters, in accord with the *a priori* guarantee. Of rather more interest is the situation where the model posterior is less than perfect. Figure 9(*b*) makes the comparison for the case of a chequerboard source and Ising prior. Here we see that there *are* choices

† There is some small print here. In practice this means sampling from the *portion* of the posterior distribution explored in the quasi-equilibrium reached by simulations initiated from the data.

‡ The TPM image $\boldsymbol{T}$ has maximum overlap with $\langle \boldsymbol{R}[\boldsymbol{D}]\rangle_{\boldsymbol{R}}$ which coincides with $\langle \boldsymbol{S}[\boldsymbol{D}]\rangle_{\boldsymbol{S}}$ when the first moments of 'model' and 'true' posterior are equal, as signalled by a Q-factor (equation (3.1)) of unity.

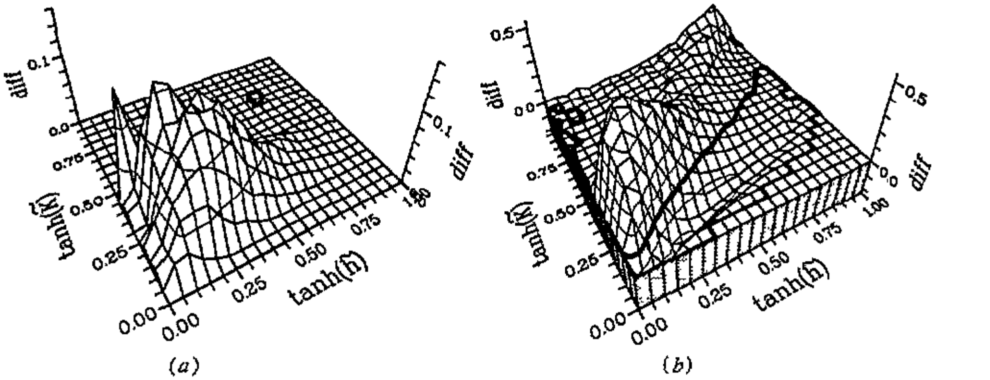

Figure 9. Comparison of MAP and TPM estimators. The vertical coordinate is the mean difference between the TPM and MAP overlap with the relevant source. (*a*) Result for an Ising source, with restoration parameters chosen to match the data-generating parameters ($\tilde{K} = K$; $\tilde{h} = h$). (*b*) Results for a 8×8 chequerboard source with $q = 0.3$. In the region lying within the heavy contour the MAP estimator does better than the TPM estimator.

of the restoration parameters for which MAP does better than TPM. Nevertheless, in wide ranging studies we have always found the optimal TPM estimator to be better than the best obtainable with MAP. Moreover, as figure 9(*b*) suggests, the region of parameter space where TPM does better is larger than the region where MAP does better.

To illuminate this point (and also some other features of MAP [17, 18, 11]) it is helpful to consider the 'path' followed by the system undergoing annealing to its ground state, in relation to the phase diagram shown in figure 6. The annealing process generates a family of restoration parameters $\tilde{K}(T)$ and $\tilde{h}(T)$, parametrized by an effective temperature T, and defined by

$$\mathcal{H}(\tilde{K}(T), \tilde{h}(T)) = \mathcal{H}(\tilde{K}, \tilde{h})/T\,. \tag{5.3}$$

In the $T \to 0$ limit the posterior distribution condenses on the MAP estimate. From the structure of the cost function (2.11*b*) it is apparent that the annealing process (lowering the effective temperature T) increases $\tilde{K}$ and $\tilde{h}$ simultaneously while maintaining a constant ratio $\tilde{K}/\tilde{h}$. Since the phase diagram has $\tanh(\tilde{K})$ and $\tanh(\tilde{h})$ as the axis variables, the lines of constant ratio are not straight (except for the trivial case of $\tilde{K} = \tilde{h}$); two such isolines are shown (figure 6), traversing two different regions (A) and (B) of the phase diagram, and terminating in its top right-hand corner.

First consider a system with restoration parameters lying at (or near) point (A), in the data-like phase. In this case the ensemble of restored images has non-zero overlap with the source, and the TPM estimator will produce reasonable results. If we anneal towards the MAP estimator, however, we follow the relevant isoline which crosses the phase boundary into the prior-like phase: the resulting MAP estimator thus has zero overlap with the source. MAP must fail in this way for any parameters that lie on an annealing path that crosses the phase boundary. This is why one finds regions of restoration parameter space where MAP fails catastrophically, while TPM is reasonable (cf figure 9(*b*), noting that the peak in the difference between MAP and TPM coincides with region (A) in figure 6).

This argument also explains why the region of parameter space in which TPM performs reasonably is always larger than the region in which MAP performs reasonably. Any point in parameter space where the TPM estimator is 'bad' (where 'bad' means worse than the data) lies in the prior-like phase. No annealing paths (isolines) cross back from the prior-like phase into the data-like phase; so the associated MAP estimator is also necessarily 'bad'.

The finer detail of the phase diagram also allows us to understand the discrepancies which others have noticed [17] between the *exact* MAP estimator and the MAP estimator found by simulated annealing. Consider a system with restoration parameters lying in region (B) of the phase diagram. Again the parameters lie in the data-like phase, and the TPM will be reasonable; again the annealing path crosses the boundary, and *if followed* leads to the exact MAP estimator which has no overlap with the source. However, in this case the annealing path crosses the phase boundary in a region where the parameters $\tilde{K}(T)$ and $\tilde{h}(T)$ are relatively large (i.e. at a low effective temperature T): the likelihood of the system being trapped in a metastable data-like state is substantial. Thus while the *exact* MAP estimator is prior-like and useless, the practically annealed MAP estimator is data-like and reasonable. Nevertheless, even this success for MAP is scarcely satisfying: it seems absurd to rely upon the metastable states for good restoration, when using a method specifically designed to avoid them!

6. Parameter estimation: evidence and free energy

6.1. Introduction

Irrespective of the way in which one chooses to use the posterior, the quality of any final reconstructed image is ultimately limited by the quality of the prior, both in regard to form and parametrization. The problem of (prior) *parameter estimation* is thus central to the image restoration task. A great deal of work has actually skirted the problem, assigning parameter values on an *ad hoc* basis (e.g. [9, 19]); this is clearly unsatisfactory. Most of the image restoration work that *does* address the issue of parameter estimation assumes (as does the evaluation of the quality factor Q, defined in (3.1)) that an ensemble of prototype uncorrupted pictures is available which can be analysed in an attempt to parameterize the source correctly [10, 20, 22, 35]. These papers build on the large body of work in the statistics literature on parameter estimation from complete or *fully observed data* [36–40]. There has been less work on parameter estimation from *incomplete data*. The iterative EM algorithm for parameter estimation [41] is now being applied to image restoration in the engineering literature [42]. A similar method found in both the engineering [43] and statistics [44] literature involves simultaneous image restoration and parameter estimation. However, there is no guarantee that such a re-estimation process will converge to even a *local* maximum of the parameters and the reconstruction simultaneously. Certainly the methods are unlikely to find the global maximum and, in general, the results depend upon the initial choice of parameters.

In this section we consider a generalized maximum likelihood formulation of the problem. Specifically we shall adopt the language of the 'evidence' framework developed by Gull [24] as a method for estimating the free parameter in conventional maximum entropy restoration, and subsequently applied to the Bayesian training problem for back-propagation neural networks [45, 46]. We note that the ideas involved have precursors in earlier work. (See e.g. [47]; the G-metric defined there corresponds to the negative of the 'average log-evidence' considered below.)

Let us first outline the key ideas in general terms. Appealing back to the general framework developed in section 2.1, we note that the normalizing factor in (2.1) must satisfy

$$\tilde{P}(D|\beta) = \sum_{\{S\}} \tilde{P}(D|S,\beta)\tilde{P}(S|\beta) \tag{6.1}$$

where we introduce $\{\beta\}$ to represent the set of parameters implicit in the model prior $\tilde{P}(\boldsymbol{S}|\beta)$ and likelihood $\tilde{P}(\boldsymbol{D}|\boldsymbol{S},\beta)$. The conditional probability $\tilde{P}(\boldsymbol{D}|\beta)$ is the *evidence* for the parameters $\{\beta\}$. The rationale for the name lies in the fact that, in the absence of biasing information on the parameters, the evidence provides a direct measure of the conditional probability $P(\beta|\boldsymbol{D})$. The parameters which maximize the evidence thus represent the single most likely (MAP) parameter values given the data (not to be confused with the MAP estimator of the *source image*, given the data *and* model parameters, considered in the previous section). We proceed to illustrate the idea with a simple explicit example before turning to consider the application of the idea to the parametrization of the non-trivial (edge-density) prior models we have focused on in this paper.

6.2. A toy example

For illustrative purposes we consider a simple problem in which the evidence can be evaluated analytically. Instead of our usual edge-density prior (equation (2.8)) we take a prior which assumes that the source is a chequerboard of chequer size $2^{\tilde{c}}$, denoted by $\boldsymbol{S}^{\tilde{c}}$. Thus we write

$$\tilde{P}(\boldsymbol{S}|\tilde{c}) = \tilde{P}(\boldsymbol{S}) = \delta_{\boldsymbol{S},\boldsymbol{S}^{\tilde{c}}}\,. \tag{6.2}$$

Retaining the form (2.6) for the model likelihood, the evidence (6.1) assumes the simple form

$$\tilde{P}(\boldsymbol{D}|\tilde{c},\tilde{h}) = \sum_{\{\boldsymbol{S}\}} \tilde{P}(\boldsymbol{D}|\boldsymbol{S},\tilde{h})\tilde{P}(\boldsymbol{S}|\tilde{c}) = \frac{1}{Z_l(\tilde{h})}\exp\left\{\tilde{h}\sum_i D_i S_i^{\tilde{c}}\right\}. \tag{6.3}$$

Now let us suppose that, in fact, the source is a chequerboard of chequer size 2^c, and that (as usual) the true noise q is parametrized by $h \equiv \frac{1}{2}\ln(1/q - 1)$. Then, invoking the thermodynamic limit (or, equivalently, a quenched average over the noise), we identify the log-evidence density

$$e \overset{\text{def}}{=} \lim_{N\to\infty}\frac{1}{N}\ln\tilde{P}(\boldsymbol{D}|\tilde{c},\tilde{h}) = \frac{1}{N}\langle\!\langle \ln\tilde{P}(\boldsymbol{D}|\tilde{c},\tilde{h})\rangle\!\rangle = \tilde{h}\delta_{c,\tilde{c}}\tanh h - \ln[2\cosh\tilde{h}] \tag{6.4}$$

where we have used the result that

$$\lim_{N\to\infty}\frac{1}{N}\sum_i S_i^c S_i^{\tilde{c}} = \delta_{c,\tilde{c}}\,.$$

Clearly, the evidence is maximized (the associated 'free energy' is minimized) when the chequer size is correctly assigned: $\tilde{c} = c$. Moreover, the turning point condition

$$0 = \frac{\partial e}{\partial\tilde{h}} = \delta_{c,\tilde{c}}\tanh h - \tanh\tilde{h}$$

correctly picks out $\tilde{h} = h$, *provided* the chequer size is assigned correctly. The failure of the method to identify the noise parameter when the chequer size is already assigned *incorrectly* is indicative of the unreliability of the method when elements of the prior are structurally incorrect.

6.3. Parameter estimation for the edge density prior

In general, evidence calculations pose a computationally intensive problem analagous to the calculation of the free energy of a system with quenched disorder. A number of novel and powerful techniques for Monte Carlo free energy evaluation have emerged in recent years (see e.g. [48, 49]). However, in the studies reported here we have found it adequate to appeal to standard 'thermodynamic' integration methods (see e.g. [50]).

For the edge density prior (2.8) the evidence is

$$\tilde{P}(\boldsymbol{D}|\tilde{K},\tilde{h}) = \sum_{\{\boldsymbol{S}\}} \tilde{P}(\boldsymbol{D}|\boldsymbol{S},\tilde{h})\tilde{P}(\boldsymbol{S}|\tilde{K}) = \frac{Z(\tilde{K},\tilde{h};\boldsymbol{D})}{Z_l(\tilde{h})Z_p(\tilde{K})} \tag{6.5}$$

from which one finds that

$$\ln \tilde{P}(\boldsymbol{D}|\tilde{K},\tilde{h}) = \int_0^{\tilde{K}} \mathrm{d}\tilde{K} \left\langle \sum_{\langle ij \rangle} R_i R_j \right\rangle_R - \ln Z_p(\tilde{K}) \tag{6.6}$$

where $\langle\cdot\rangle_R$ signifies averaging with respect to the probability $\tilde{P}(\boldsymbol{R}|\boldsymbol{D})$.

The evidence for a given coupling $\tilde{K}$ may thus be determined by numerical integration of the correlation function along the path to zero coupling.

Consider first the situation in which $\tilde{P}(\boldsymbol{S})$ and $P(\boldsymbol{S})$ are well matched: the source image is drawn from an Ising distribution of specific K, and the noise process is Gaussian with noise level q (corresponding to a field h). Figure 10 shows the log evidence $\ln \tilde{P}(\boldsymbol{D}|\tilde{K},\tilde{h})$, as a function of $\tilde{K}$, $\tilde{h}$, computed using representative corrupted images $\boldsymbol{D}$ derived from an Ising source with the same parameters as those underlying the Q-factor results presented in figure 3. In this case, the evidence maximum correctly identifies the generating parameters and thus coincides with the maximum of the quality factor. The large negative evidence characteristic of the region where $\tilde{K}$ and $\tilde{h}$ are *both* large is striking; it can be understood as follows. The underlying data (generated from a near critical coupling) has approximately *zero* bias; in the large $\tilde{K}$ and $\tilde{h}$ region the hypothesis being evaluated is that the source coupling is a *close representation* (h is 'large') of a source of large coupling, and thus *large* bias. The hypothesis is thus particularly poor; this is what the evidence shows.

When $\tilde{P}(\boldsymbol{S})$ and $P(\boldsymbol{S})$ are ill matched the results are more problematic. Figure 11 plots the log evidence for the parameters of the same edge-density model but with a range of *chequerboard* sources and noise levels. Comparison with figure 4 shows that there is *qualitative* similarity between the *most likely* parameters identified by the maximum of the evidence and the *optimal* parameters identified by the maxima of the Q-factor. (It should be recalled that the former—the evidence—utilizes only instances of the actual data

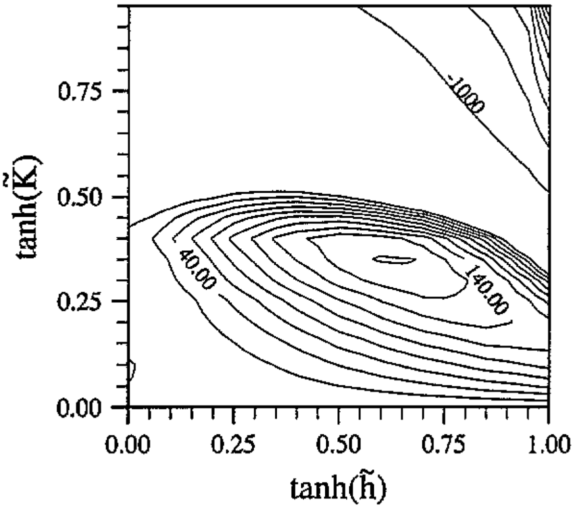

Figure 10. The log evidence for the Ising prior with Ising source: the contours show the value of $E \equiv \ln \tilde{P}(D|\tilde{K},\tilde{h}) + N \ln 2$ averaged over 50 realizations of data constructed from an $N = 64^2$ Ising source with the density of edges $\epsilon_S = 0.25$ and noise $q = 0.2$ ($\tanh(K) \simeq 0.36, \tanh(h) = 0.6$), as in figure 3. The positive contours are spaced every 20 units. The negative contours in the top right are spaced every 1000 units.

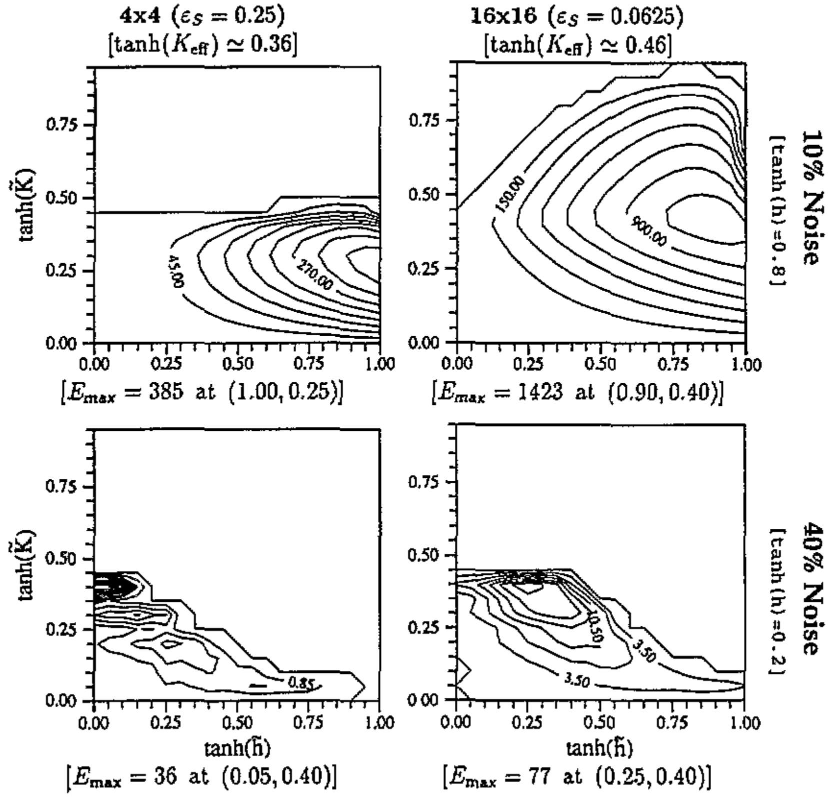

Figure 11. The log evidence for the Ising prior, with chequerboard sources: the contours show the value of $E \equiv \ln P(D|\tilde{K}, \tilde{h}) + N \ln 2$ (for $N = 64^2$), averaged over 50 realizations of data constructed from chequerboards with various chequer sizes (density of edges ϵ_S) and noise levels (q); only positive contours are plotted. The values and locations of the evidence maxima $E_{\rm max}$ are noted in each case. (For comparison, the evidence values for chequerboard priors, optimally tuned to these chequerboard sources (i.e. $\tilde{h} = h$ and $\tilde{c} = c$) follow from (6.4) as $E_{\rm max} = 1508$ when $\tanh(h) = 0.8$ and $E_{\rm max} = 82$ when $\tanh(h) = 0.2$).

D; while the latter—the Q-factor—utilizes knowledge of the full true source distribution $P(S)$.) Thus, in particular, both the Q-optimal and evidence-maximal couplings $\tilde{K}$ increase with increasing chequer size; both Q-optimal and evidence-maximal fields $\tilde{h}$ increase with decreasing source noise. But the overall level of agreement is poor—particularly in the case of the largest chequer size and the lower noise level.

The lesson (already apparent from the toy example discussed above, and explored in a different context elsewhere [51]) is clear: evidence does not provide a reliable performance measure (and thus guide) when the modelling assumptions (structures of priors) are intrinsically poor. On a more positive note, however, comparison of the results of figure 11 with those of equation (6.4) readily shows that (cf caption to figure 11)) the evidence for an Ising source—even if assigned the evidence-optimal parameters—is always less than the evidence for the (actual) chequerboard source. It may be that it is in such model comparison (as distinct from parameter estimation) that the evidence procedure ultimately proves most fruitful. Expanded ensemble methods of free energy calculation [48] may prove useful here.

7. Conclusions

Image restoration is a hard problem, with many unsolved aspects. Over the years, it has been addressed in a wide variety of disciplines, ranging from signal-processing to applied statistics. In this paper, building on the seminal work by Geman and Geman [9], we have endeavoured to show that important aspects of the problem may be illuminated by appeal to the methods and concepts of statistical mechanics.

Our mean-field analysis does much to explain the dependence of the quality of the posterior distribution on the restoration parameters. The competition between an image-smoothing prior and a data-binding likelihood gives rise to a posterior that supports phase transitions between data-dominated and prior-dominated regions. In the prior-dominated regions the posterior (and any restoration scheme based on it) will be useless, unless one can capitalize on the metastable data-like states which persist beyond the phase boundary.

In our comparison of the different ways in which one can use the posterior to identify a single binary image, we have found that almost invariably TPM provides a more reliable estimate of the original source image than does MAP. Although choosing the optimal values for the restoration parameters remains problematical, it is easy to avoid the region of parameter space where the TPM estimate offers no improvement, and it is only in a part of this region that MAP may do better than TPM. We have seen that the process of finding the MAP image can fail badly if the annealing process traverses the data-prior phase transition. Such 'gains' as are made along the annealing path are lost when the phase boundary is crossed, and much of the computation time is wasted annealing in the 'wrong' part of the phase diagram. The existence of metastable data-like states accounts for the fact that simulated annealing may fail to reproduce the exact MAP estimate (and may do all the better thereby). The TPM estimate is free of such problems, and may be computed in a fraction of the Monte Carlo time. We believe that it should be the favoured estimate for image reconstruction problems.

Our studies of the evidence procedure for assigning restoration parameters show that the effectiveness of the method is limited by the quality of the *form* of the prior. If the prior is well matched to the source, the most probable parameters identified by the evidence maximum coincide with the data-generating parameters, and secure optimal quality for the posterior. But if the prior models the source poorly, the evidence optimal parameters are not, in general, reliable guides to the parameters which will optimize the quality of the posterior. Nevertheless the development of techniques for Monte Carlo calculations of free energy holds out interesting possibilities for the use of evidence in comparing model priors.

Acknowledgment

JMP acknowledges the support of a SERC research studentship.

References

[1] Papoulis A 1986 *Probability, Random Variables and Stochastic Processes* 2nd edn (Singapore: McGraw-Hill)
[2] Habibi A 1972 *Proc. IEEE* **60** 878–83
[3] Hunt B R 1977 *IEEE Trans. Comput.* **26** 219–29
[4] Nahi NE and Assefi T 1972 *IEEE Trans. Comput.* **21** 734–8
[5] Richardson W H 1972 *J. Optical Soc. Am.* **62** 55–9
[6] Besag J 1986 *J. R. Stat. Soc.* B **48** 259–302
[7] Chellappa R and Jain A (eds) 1993 *Markov random fields: theory and application.* (Boston, MA: Academic)

[8] Hammersley J M and Clifford P 1968 Markov fields of finite graphs and lattices *Preprint* University of California—Berkeley

[9] Geman S and Geman D 1984 *IEEE Trans. Pattern Analysis Machine Intell.* 6 721–41

[10] Geman S and Graffigne C 1986 Markov random field image models and their applications to computer vision *Proc. Int. Congress of Mathematicians* ed A M Gleason (Providence RI: American Mathematical Society)

[11] Marroquin J L, Mitter S and Poggio T 1987 *J. Am. Stat. Assoc.* 82 76–89

[12] Smith A F M and Roberts G O 1993 *J. R. Stat. Soc.* B 55 3–23

[13] Gidas B 1989 *IEEE Trans. Pattern Anal. Machine Intell.* **11** 164–80

[14] Geiger D and Girosi F 1991 *IEEE Trans. Pattern Anal. Machine Intell.* **13** 401–12

[15] Frigessi A, di Stefano P, Hwang C-R and Sheu S-J 1993 *J. R. Stat. Soc.* B 55 205–19

[16] Pryce J 1993 Statistical mechanics of image restoration *PhD Thesis* University of Edinburgh

[17] Greig D M, Porteous B T and Seheult A H 1989 *J. R. Stat. Soc.* B 51 271–9

[18] Marroquin J L 1985 *A. I. Lab. Memo* **839** MIT

[19] Jeng F C and Woods J W 1991 *IEEE Trans. Signal Proc.* 39 683–97

[20] Kashyap R L and Chellappa R 1983 *IEEE Trans. Info. Theory* **29** 60–72

[21] Pickard D K 1977 *J. Am. Stat. Assoc.* 82 90–6

[22] Simchony T, Chellappa R and Lichtenstein Z 1990 *IEEE Trans. Info. Theory* **36** 608–13

[23] Frigessi A and Piccioni M 1989 *Stoch. Proc. App.* 34 297–311

[24] Gull S F 1989 Developments in maximum entropy data analysis *Maximum Entropy and Bayesian Methods, Cambridge 1988* ed J Skilling (Dordrecht: Kluwer)

[25] MacKay D J C 1992 *Neural Computation* 4 415–47

[26] Neal R M 1993 Probabilistic inference using Markov chain Monte Carlo methods University of Toronto, Dept of Computer Science *Technical Report* CRG-TR-93-1

[27] Baierlein R 1971 *Atoms and Information Theory: an Introduction to Statistical Mechanics* (San Francisco: Freeman)

[28] Binder K and Heermann D W *Monte Carlo Simulation in Statistical Mechanics* (Berlin: Springer)

[29] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* 21 1087–92

[30] Itzykson C and Drouffe J M 1989 *Statistical Field Theory* (New York: Cambridge University Press)

[31] Derin H and Elliott H 1987 *IEEE Trans. Pattern Anal. Machine Intell.* 9 39–55

[32] Ripley B D 1986 *Canadian J. Stat.* 14 83–111

[33] Jeng F C, Woods J W and Rastogi S 1993 Compound Gauss–Markov random fields for parallel image processing in [7] pp 11–38

[34] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 *Science* 220 671–80

[35] Wahba G 1977 *SIAM J. Numer. Anal.* 14 651–67

[36] Besag J 1974 *J. R. Stat. Soc.* B 36 192–225

[37] Besag J and Moran P A P 1975 *Biometrika* 62 555–62

[38] Besag J 1975 *The Statistician* 24 179–95

[39] Besag J 1977 Efficiency of pseudolikelihood estimation for simple Gaussian fields *Biometrika* **64** 616–8

[40] Ord K 1975 *J. Am. Stat. Assoc.* 70 120–6

[41] Dempster A P, Laird N M and Rubin D B 1977 *J. R. Stat. Soc.* B 39 1–38

[42] Zhauo Y, Zhuang X, Atlas L and Anderson L 1992 *CVGIP: Graphical Models Image Processing* 54 187–97

[43] Lakshmanan S and Derin H 1989 *IEEE Trans. Pattern Anal. Machine Intelli.* 11 799–813

[44] Qian W and Titterington D M 1991 *Phil. Trans. R. Soc.* A 337 447–28

[45] Buntine W L and Weigand A S 1991 *Complex Systems* 5 603–43

[46] MacKay D J C 1992 *Neural Computation* 4 448–72

[47] Ackley D H Hinton G E and Sejnowski T J 1985 *Cogn. Sci.* 9 147–69

[48] Lyubartsev A P, Martsinovski A A, Shevkunov S V and Vorontsov-Velyaminov P N 1992 *J. Chem. Phys.* 96 1776–83

[49] Berg B A and Neuhaus T 1992 *Phys. Rev. Lett.* 68 9–12

[50] Frenkel D 1986 Free-energy computation and first-order phase transitions *Molecular Dynamics Simulation of Statistical-Mechanical Systems* (Bologna: XCVII Corso Soc. Italiana di Fisica)

[51] Bruce A D and Saad D 1994 *J. Phys. A: Math. Gen.* 27 3355–63