Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Estimation of Parameters in Hidden Markov Models

W. Qian and D. M. Titterington

Phil. Trans. R. Soc. Lond. A 1991 **337**, 407-428
doi: 10.1098/rsta.1991.0132

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. Lond. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

This journal is © 1991 The Royal Society

Downloaded from rsta.royalsocietypublishing.org

Estimation of parameters in hidden Markov models

By W. Qian and D. M. Titterington

Department of Statistics, University of Glasgow, Glasgow G12 8QQ, U.K.

Parameter estimation from noisy versions of realizations of Markov models is extremely difficult in all but very simple examples. The paper identifies these difficulties, reviews ways of coping with them in practice, and discusses in detail a class of methods with a Monte Carlo flavour. Their performance on simple examples suggests that they should be valuable, practically feasible procedures in the context of a range of otherwise intractable problems. An illustration is provided based on satellite data.

1. Introduction

There is a wide class of problems in statistics that can be regarded as incomplete-data problems. The observed data, to be modelled as a realization of a random vector, Y, are interpreted as a partly observed version of a 'complete' set of data, which would be a realization of a random vector, Z. Often, Z has physical meaning, as in the context of data with missing values. Here

$$Z = (X, Y), \tag{1.1}$$

where X represents the missing values.

It should be emphasized that there is a range of incomplete-data problems that are not so obviously represented by the 'complete data = observed data + missing values' paradigm of (1.1). These include such topics as censored and truncated data; see Dempster *et al.* (1977) and Little & Rubin (1987) for many examples and much discussion. However, in the present paper, we concentrate on missing data problems. In particular, we assume that

$$Z = \{Z_i\},$$

where $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$. Here, n represents the number of individuals or items in a sample, or, in our later discussion, the number of pixels in an image. Thus each of the n 'observations' consists of multivariate data of which Y_i is observed and X_i is missing. In the present paper, we assume that each X_i corresponds to the same variable or variables in the multivariate data vector. Thus we are considering a more particular structure than is typical in the context of, say, sample-survey data with non-response a possibility in any of the individual variables. In the simplest version of our structure (which is also both very common in practice and the one we shall consider in detail throughout the paper), X_i is one dimensional, discrete and finite valued. Somewhat more generally, each X_i has the same sample space, and similarly for the Y_i. Without too much extra effort, one could cope with additional incompleteness within the Y_i, but we shall not introduce this extra complication.

Phil. Trans. R. Soc. Lond. A (1991) **337**, 407–428
Printed in Great Britain

The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to *Philosophical Transactions: Physical Sciences and Engineering.* www.jstor.org

Downloaded from rsta.royalsocietypublishing.org

Some particular examples of our structure are as follows.

(*a*) $\{X_i\}$ are the (unknown) age-categories of n fish and $\{Y_i\}$ are the observed lengths of the fish.

(*b*) $\{X_i\}$ are a time-sequence of n (unknown) configurations of an individual's vocal tract and $\{Y_i\}$ are the corresponding sequence of projected sounds.

(*c*) X denotes the true surface-typing of a pixellated area of land and Y denotes the corresponding image observed by a satellite.

Case (*c*) is just one illustration of a noisy image, other examples of which occur in many fields, such as petrology and medicine; see Aykroyd & Green (1991), for instance.

In all these illustrations, the X_i are 'real' physical quantities. However, in some applications of our models, the X_i are latent variables, familiar in factor analysis and, potentially, in statistical approaches to neural networks; see also §7.

The key feature of our work will be the proposal of a statistical model for Z in the form of the joint probability density

$$f(Z\,|\,\theta) = f(X, Y\,|\,\theta), \tag{1.2}$$

where the functional form, f, will be treated as being known. If we are provided with data $Y = y$, therefore, two quantities are uncertain, the associated missing values x and the unknown parameter(s), θ. It pays to recognize the different natures of these two unknown quantities, and this is reflected in the different terminology we shall use for the two operations of trying to identify them on the basis of y; we might wish to (i) impute values for the missing data, x, and/or (ii) estimate the unknown parameter(s), θ.

The importance of making this distinction is emphasized later in the paper and is also highlighted by Little & Rubin (1983). We shall make some reference to objective (i), but the principal aim of this paper is to discuss approaches to (ii) and, in particular, to examine, and attempt to deal with, complications with maximum likelihood estimation in a class of problems with important applications.

The layout of the paper is as follows. In §2, we discuss the class of models of interest and, in §3, we highlight the difficulty of obtaining maximum likelihood estimates of parameters. A discussion of various approaches that attempt to obviate the difficulties in §4 is followed by a detailed, illustrated investigation of a subclass of procedures, based on Monte Carlo methods, in §5. An example involving satellite data is presented in §6 and a small discussion in §7 concludes the paper.

2. Mixture models, dependence structures and applications

There are, formally, the following two ways of factorizing (1.2) in terms of a conditional distribution,

$$f(x, y\,|\,\theta) = p(y\,|\,x, \theta)\,\pi(x\,|\,\theta) \tag{2.1}$$

and

$$f(x, y\,|\,\theta) = \pi(x\,|\,y, \theta)\,p(y\,|\,\theta). \tag{2.2}$$

We generically denote 'densities' associated with X by π, and those associated with Y by p. Clearly, the marginal density for Y, $p(\cdot\,|\,\theta)$, which defines the likelihood function for the observed data, is also given by

$$p(y\,|\,\theta) = \int p(y\,|\,x, \theta)\,\mathrm{d}\Pi(x\,|\,\theta), \tag{2.3}$$

Downloaded from rsta.royalsocietypublishing.org

where $\Pi(\cdot\,|\,\theta)$ is the measure associated with $\pi(\cdot\,|\,\theta)$. Equation (2.3) displays $p(y\,|\,\theta)$ in the form of a mixture density, with $\Pi(\cdot\,|\,\theta)$ as mixing measure and X as the mixing variable. Formulation (2.2) is important, in that the factor $\pi(\cdot\,|\,y,\theta)$ is the natural basis for methods of imputing values for the missing X, given observed data y, but we concentrate largely on (2.1) and (2.3).

It will often be natural to write $\theta = (\phi, \beta)$, where ϕ and β are distinct sets of parameters, each set with its own parameter space, associated respectively with the two factors on the right-hand side of (2.1). Thus

$$f(x, y\,|\,\theta) = p(y\,|\,x, \phi)\pi(x\,|\,\beta). \tag{2.4}$$

Recall that models such as (2.4) are associated with data on n individuals or items. Thus, for instance,

$$(X, Y) = \{(X_i, Y_i), i = 1, \ldots, n\}.$$

Very often, the Y_i are conditionally independent, given the X_i, so that

$$p(y\,|\,x, \phi) = \prod_{i=1}^{n} p(y_i\,|\,x, \phi), \tag{2.5}$$

or, even more simply,

$$p(y\,|\,x, \phi) = \prod_{i=1}^{n} p(y_i\,|\,x_i, \phi). \tag{2.6}$$

So far as estimation is concerned, the major complications emanate from the factor $\pi(x\,|\,\beta)$ in (2.4). As described in Titterington (1990), a gradation in complexity can be identified by considering different dependence patterns among the X_i. Except when discussing Example 5.2, we assume throughout that each X_i takes one of a finite set of k qualitative values $\{c_1, \ldots, c_k\}$, although some of the formulation can be generalized to cover, for instance, the case of an AR(1) process with additive noise, equivalent to an ARMA(1, 1) process.

(*a*) *Standard mixture model*

Here, the X_i are independent, and we shall assume them to be identically distributed. Thus, if (2.6) obtains, the Z_i are independent and the Y_i are marginally so. As a result,

$$\pi(x\,|\,\beta) = \prod_{i=1}^{n} \pi(x_i\,|\,\beta),$$

and

$$p(y\,|\,\theta) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} p_j(y_i\,|\,\phi)\pi_j \right\}, \tag{2.7}$$

where

$$p_j(y_i\,|\,\phi) = p(y_i\,|\,X_i = c_j, \phi)$$

and

$$\pi_j = \text{prob}\,(X_i = c_j),$$

for all j. (In the context of images, the c_j may be a set of colours, usually representing some discrete classification of the pixels in the true scene.) the π_j are called mixing weights and $p(y\,|\,\theta)$ represents the joint density of a random sample from a finite, k-component mixture distribution. Since the mixing X_i are, marginally, samples from a multinomial distribution, the mixture model may, in this case, be thought of as a hidden multinomial model. There are very many applications of standard mixture models; see Titterington *et al.* (1985), Titterington (1990) and references therein.

Downloaded from rsta.royalsocietypublishing.org

(b) *Hidden Markov chain model*

We now admit a simple form of dependence among the x_i, namely that

$$\pi(x \mid \beta) = \pi(x_1 \mid \beta) \sum_{i=2}^{n} \{\pi(x_i \mid x_{i-1}, \beta)\}.$$

Thus the x_i form a Markov chain, which is generally assumed to be stationary. Since the X_i are now no longer independent, the marginal density for Y_i no longer simplifies in the way that (2.7) does. The Y_i are said to come from a hidden Markov chain model, and the study of such models has become extremely popular in the speech recognition literature (Juang & Rabiner 1991; Bourlard 1990), as well as in the analysis of neurophysiological data.

In the former context, the X_i form a sequence (in time) of underlying prototypical spectra, each representing one of a (finite) number of configurations of the vocal tract, and the Y_i are random functions thereof.

(c) *More general hidden Markov models*

Perhaps the currently most familiar application of more general hidden Markov models is the hidden Markov random field model for noisy images. Here, the subscript i identifies a pixel-site, or a line-site, in a pixellated image, X represents the true scene and Y the observed image, which is just a blurred and/or noisy version of X.

We allow a more general, but still markovian marginal model for the X_i, and represent their joint density by that of a Gibbs distribution. Thus,

$$\pi(Z \mid \beta) = (1/C(\beta)) \exp\{-U(Z, \beta)\},$$

where $C(\beta)$ is a normalizing constant (the partition function), and the energy function, U, takes the additive form

$$U(Z, \beta) = \sum_{c \in \Theta} V_c(Z, \beta),$$

where Θ is a class of subsets of the sites, and V_c is the potential function associated with subset c. Usually, each $c \in \Theta$ consists of only a few sites. For instance, in the case of pairwise-interaction models, Θ contains only the elementary subsets consisting of the individual sites themselves, along with some two-site subsets, each containing a pair of 'neighbouring' sites; for popular two-dimensional examples, see Geman & Geman (1984). In the context of images, the structure of the Gibbs distribution is intended to reflect plausible local, spatial correlation in the true scene. Thus, typically, each V_c involves only a very small subset of the Z_i. As in the hidden Markov chain case, the likelihood function defined by the marginal density, $p(y \mid \theta)$, does not take a simple form.

The seminal papers on this application are Geman & Geman (1984) and Besag (1986). Those papers and many others, including Aykroyd & Green (1991), contain much material about restoring the scene (equivalent in our phraseology to imputing values for the missing x). Along with Ripley (1988), they also have something to say about the other problem, namely, that of estimating θ, that is the main focus of the present paper.

So far as $p(y \mid x, \phi)$ is concerned, we shall assume that (2.6) holds. Note, however, that (2.6) does not cover the case of systematic blurring, although the general

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

methods we describe are not necessarily limited, in terms of applicability, to non-blurred images.

3. Maximum likelihood estimation: methods and difficulties

As a general rule, parameter estimation from incomplete data is more awkward than would be the case if the data were complete. Often, the additional difficulty is simply that, whereas explicit formulae exist for estimates in the complete-data case, iterative numerical methods are now required. Sometimes, the situation is even more awkward, particularly when the complete-data version is itself non-trivial. The maximum-likelihood treatment of the models in §2 illustrates these points very well; see Titterington (1990) for a more detailed exposition.

We restrict our discussion to the case where each $x_i \in \{c_1, \ldots, c_k\}$.

Even for the simplest (hidden multinominal) model, explicit maximum likelihood estimates are hardly ever (not quite never!) available. Were the x_i known, parameter estimation would often be trivial: the k mixing weights would be estimated by relevant relative frequencies and the parameters associated with, say, $p_j(\cdot \mid x_i, \phi)$ are often easily estimated. Recently, a particular numerical procedure, the EM-algorithm (Dempster *et al.* 1977) has pervaded the incomplete-data literature. In the mth iterative stage, current estimates, $\theta^{(m)}$, are updated to $\theta^{(m+1)}$ by the following double step:

E-step: compute $Q^{(m)}(\theta) \equiv E\{\ln f(X, y \mid \theta) \mid y, \theta^{(m)}\}$,

M-step: find $\theta = \theta^{(m+1)}$ to maximize $Q^{(m)}(\theta)$.

Typically, the M-step is as easy or hard as is the computation of maximum likelihood estimates from complete data; the E-step involves 'averaging out' over the missing values and is sometimes loosely equivalent to imputing for the missing values.

Repetition of the double step generates a non-decreasing sequence of values of the log-likelihood (associated with y) and convergence to a local maximum likelihood estimate can often be proved.

The levels of difficulty involved in the implementation of the EM-algorithm can be summarized as follows, in which 'explicit' means that an explicit formula exists, not requiring numerical integration or summation (E-step) or iterative solution (M-step).

(i) *Hidden multinomial.* E-step: explicit; M-step: explicit.

(ii) *Hidden Markov chain.* E-step: explicit, although it does require a forwards and a backwards recursion through the data; M-step: explicit.

Details are provided in Titterington (1990). Note that, in claiming that the M-step is explicit, it is assumed that we are dealing with a case in which the 'colour-conditional' densities p_j admit explicit maximum likelihood estimates. This is true, for instance, for the case of gaussian densities, but not for beta densities.

(iii) *Hidden Markov random field.* E-step: very difficult; M-step: very difficult.

In the above, 'very difficult' borders on the impossible, at least so far as exact calculation is concerned! To explain this, it is helpful to introduce a slight change of notation. Let X_i now denote an indicator vector, of length k, which has unity as the jth element, and zero elsewhere, if pixel i is of colour c_j. Let X_{ij} denote the jth element of X_i. Then

$$\ln f(X, y \mid \theta) = -\sum_c V_c(X, \beta) - \ln C(\beta) + \sum_i \sum_j X_{ij} \ln p_j(y_i \mid \phi).$$

Downloaded from rsta.royalsocietypublishing.org

In the E-step we must therefore compute, given $\theta^{(m)}$,

$$Q^{(m)}(\theta) = -\sum W_c^{(m)}(\beta) - \ln C(\beta) + \sum_i \sum_j X_{ij}^{(m)} \ln p_j(y_i \mid \phi),$$

where
$$W_c^{(m)}(\beta) = E\{V_c(X, \beta) \mid y, \theta^{(m)}\} \tag{3.1}$$

and
$$X_{ij}^{(m)} = E(X_{ij} \mid y, \theta^{(m)}). \tag{3.2}$$

Exact computation of both (3.1) and (3.2) is, typically, impossible.

So far as the M-step is concerned, computation of $\phi^{(m+1)}$ is often straightforward, as in the earlier models. Maximization of the part of $Q^{(m)}(\theta)$ depending on β is not easy, in general, largely because computation of $C(\beta)$ is hardly ever feasible. In other words, even if we have a pure realization from a Markov random field, maximum likelihood estimation of the underlying parameter, β, is not a practical proposition.

The inherent difficulties are well illustrated by the following simple example.

Example 3.1. Exponential family case

Suppose

$$f(x, y \mid \theta) = (1/D(\theta)) \exp\{H(x, y)^{\mathrm{T}} \theta\}. \tag{3.3}$$

Then the EM double step is as follows.

E-step: evaluate $E\{H(X, y) \mid y, \theta^{(m)}\}$.

M-step: solve $E\{H(X, y) \mid \theta\} = E\{H(X, y) \mid y, \theta^{(m)}\}$ to obtain $\theta = \theta^{(m+1)}$.

In general, neither of the above expectations can be computed explicitly, nor can the equation in the M-step be solved without numerical methods. Both expectations can be approximated by sample means created by multiple implementations of the two associated Gibbs samplers. (These Gibbs samplers are simply mechanisms for simulating realizations from the relevant Gibbs distributions. From an arbitrary initial configuration, a realization for a given site is simulated from its distribution conditional on the rest of the configuration. The configuration is updated, the procedure is repeated for each site and the whole operation iterated a 'large' number of times. Ultimately, a valid realization is created (see also Smith 1991).)

However, only in very simple cases is the procedure feasible in practice, because of the scale of computations. Geman & McClure (1987) describe a one-parameter problem, for which an off-line approximation to the function $E\{H(X, y) \mid \theta\}$ is computed at the outset, using a grid of θ-values.

One important idea with considerable promise, also involving Monte Carlo procedures, is the approach of Geyer & Thompson (1992). Although they concentrate on the case of non-noisy data, their idea is useful in the M-step of the EM algorithm for Example 3.1, where we have to maximize, with respect to θ,

$$E\{H(X, y) \mid y, \theta^{(m)}\}^{\mathrm{T}} \theta - \ln D(\theta). \tag{3.4}$$

The difficulty lies in the intractability of $D(\theta)$, but Geyer & Thompson (1992) note that, for any θ',

$$D(\theta) = D(\theta') \int \exp\{H(x, y)^{\mathrm{T}} (\theta - \theta')\} \, \mathrm{d}F(x, y \mid \theta'), \tag{3.5}$$

which is simply proportional to a moment generating function, may be approximated by the empirical counterpart

$$D_N(\theta) = D(\theta') \frac{1}{N} \sum_{r=1}^{N} \exp\{H(x_r, y_r)^{\mathrm{T}} (\theta - \theta')\},$$

Downloaded from rsta.royalsocietypublishing.org

where $\{(x_r, y_r), r = 1, \ldots, N\}$ are realizations from the distribution with density $f(x, y \mid \theta')$. Since (3.4) involves $\ln D(\theta)$, to be approximated by $\ln D_N(\theta)$, rather then $D(\theta)$ itself, it is not necessary to know $D(\theta')$ when computing the maximal θ. Geyer & Thompson (1992) give much more detail, including guidance about the choice of θ'. Clearly, the choice of $\theta' = \theta^{(m)}$ is a natural one when computing $\theta^{(m+1)}$ in the EM algorithm. Alternatively, one could use the same N simulated realizations throughout the EM algorithm.

As noted by Besag (1976), (3.5) has a history going back at least to Bartlett (1971).

4. Practical parameter estimation for hidden Markov models

A variety of more practicable procedures have been considered for dealing with data from hidden Markov models. Some of them are comparatively general in scope, whereas others are appropriate only for special cases.

4.1. *Methods based on decision-directed imputation*

Recall that, for most hidden Markov models, there are two difficulties: (*a*) X is missing; (*b*) even were X provided, maximum likelihood estimation of β would be difficult.

One general approach is to generate a sequence $\{(\theta^{(m)}, x^{(m)}), m = 0, 1, \ldots\}$ of pairs of iterates for θ and the missing X such that $x^{(m+1)}$ is created on the basis of y and $\theta^{(m)}$, and $\theta^{(m+1)}$ is 'estimated' from y and $x^{(m+1)}$.

The general form of the iterative stage in what we might call the restoration–maximization (RM) algorithm is as follows, given $\{\theta^{(m)}, x^{(m)}\}$.

R-step: create $x^{(m+1)}$ from $\pi(x \mid y, \theta^{(m)})$;

M-step: choose $\theta = \theta^{(m+1)}$ to maximize

$$p(y \mid x^{(m+1)}, \phi)\, \pi_p(x^{(m+1)} \mid \beta), \tag{4.1}$$

where $\pi_p(x^{(m+1)} \mid \beta)$ is an amenable alternative to $\pi(x^{(m+1)} \mid \beta)$, such as Besag's (1975) pseudo-likelihood, defined by

$$\prod_{i=1}^{n} \pi(x_i^{(m+1)} \mid x_{\partial i}^{(m+1)}, \beta), \tag{4.2}$$

in which $x_{\partial i}$ denotes the values of x on the neighbouring pixels to pixel i. The crucial simplifying feature of the pseudo-likelihood is that the partition function $C(\beta)$ is absent.

There are various specific versions of the R-step, including the following.

$\mathrm{R}_{\mathrm{ICM}}$ (Besag 1986): apply Besag's (1986) ICM (iterated conditional modes) algorithm to $\pi(\cdot \mid y, \theta^{(m)})$. This iterative procedure, which typically converges quickly, locates an $x^{(m+1)}$ that may be close to the true mode.

$\mathrm{R}_{\mathrm{MAP}}$: compute (or attempt to compute) the true mode of $\pi(\cdot \mid y, \theta^{(m)})$, thereby obtaining the maximum *a posteriori* (MAP) restoration. For very special situations (Greig *et al.* 1989), the MAP restoration may be obtained exactly. Otherwise, techniques such as simulated annealing (Geman & Geman 1984) have been proposed which may or may not attain the mode in practice.

There are, however, unsatisfactory aspects of this approach of trying to find a modal restoration and then behaving, in the M-step, as if the restoration is the true scene. We are effectively attempting to maximize $f(x, y \mid \theta)$ simultaneously with

Downloaded from rsta.royalsocietypublishing.org

respect to the missing x and the unknown θ. The finite mixture (hidden multinomial) version of this approach is well known to lead to biases in the estimates of θ (see, for example, Marriott 1975; Titterington 1984).

Modal restoration is in the same spirit as so-called decision-directed learning, familiar in the engineering literature on unsupervised learning. To explain this concept, let $\Delta(x)$ denote, in general, a randomized rule whereby restoration x is selected with probability $\Delta(x)$, for all x. Decision-directed rules correspond to Δ being degenerate. Thus modal restoration is representable by

$$\Delta(x) = \begin{cases} 1 & \text{if } x \text{ is the posterior mode,} \\ 0 & \text{otherwise.} \end{cases}$$

For further discussion of decision-directed learning, see, for instance, ch. 6 of Titterington *et al.* (1985).

The problem can be attacked by modifying either or both of the R-step and the M-step.

Qian & Titterington (1991) modified the latter as follows. If (2.6) applies, then (4.1), combined with (4.2), gives

$$\prod_{i=1}^{n} \{p(y_i \,|\, x_i^{(m+1)}, \phi)\, \pi(x_i^{(m+1)} \,|\, x_{\partial i}^{(m+1)}, \beta)\}.$$

Instead, Qian & Titterington (1991) used

$$\prod_{i=1}^{n} \{p(y_i \,|\, x_{\partial i}^{(m+1)}, \theta)\}. \tag{4.3}$$

There are various points to make about (4.3). Note that it can be written

$$\prod_{i=1}^{n} \left\{ \sum_{x_i} p(y_i \,|\, x_i, \phi)\, \pi(x_i \,|\, x_{\partial i}^{(m+1)}, \beta) \right\} = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{k} \pi_j^{(m+1,\,i)}(\beta) p_j(y_i \,|\, \phi) \right\},$$

where, for each i,

$$\pi_j^{(m+1,\,i)}(\beta) = \pi(x_i = j \,|\, x_{\partial i}^{(m+1)}, \beta), \quad j = 1, \ldots, k. \tag{4.4}$$

In spite of the tortuous notation, it can be seen that the k quantities in (4.4) are a set of mixing weights, and that (4.3) takes the form of a likelihood from independent observations from finite mixtures of the same component densities. The sets of mixing weights vary, but are clearly linked through the common β, and are locally associated, because of the dependence on $x_{\partial i}$. Qian & Titterington (1991) used a few steps of the EM algorithm to maximize (4.3) within the M-step of the RM algorithm. Clearly, within the ith factor of (4.3), the method still treats $x_{\partial i}^{(m+1)}$ as if it were the truth. However, acknowledgement that x_i itself is unknown appears to lead to improved estimates of the parameters and subsequently to robust restoration procedures; see the authors' reply to the discussion of Besag *et al.* (1991).

Qian & Titterington (1991) call their estimation algorithm, based on (4.3), the point-pseudo-likelihood (PPL)-EM algorithm. They also generalized the method.

4.2. *Methods based on probabilistic-teacher imputation*

It is also possible to modify the R-step in ways that are likely to improve the properties of the resulting estimators. Just as the use of restorations such as posterior modes is comparable with decision-directed imputation, so can one parallel the so-

Downloaded from rsta.royalsocietypublishing.org

called probabilistic-teacher procedures familiar in the unsupervised-learning literature; see ch. 6 of Titterington *et al.* (1985). One essentially chooses, for $x^{(m+1)}$, a realization from $\pi(x \mid y, \theta^{(m)})$. Thus the randomized rule $\Delta(\cdot)$ introduced in §4.1 is non-degenerate and, in this case, is simply $\pi(\cdot \mid y, \theta^{(m)})$. In general, it is much safer to treat such an $x^{(m+1)}$ as if it were the truth than it is in the decision-directed case. However, any convergence properties of $\theta^{(m+1)}$ will be in law rather than, say, in probability. For the latter, further modification is clearly necessary to create a sequence of parameter estimates with suitably ergodic behaviour. Simulation of $x^{(m+1)}$ can be achieved in various ways, by using the Gibbs sampler of Geman & Geman (1984), for example.

One such modified approach is the following. Choose a positive integer T.

R-step: simulate independent samples $\{x_t^{(m+1)};\ t = 1, \ldots, T\}$ from $\pi(x \mid y, \theta^{(m+1)})$.

M-step: for each t, find θ_t, based on $x_t^{(m+1)}$, using (4.1) or any desired variant thereof, and take

$$\theta^{(m+1)} = \frac{1}{T} \sum_{t=1}^{T} \theta_t. \tag{4.5}$$

Thus each simulated x_t creates an estimate of θ, and those estimates are then averaged. It will be helpful to consider this algorithm in more detail in the context of a particular example.

Example 4.1

Suppose that there is an explicit method for estimating θ in the M-step of the RM-algorithm, so that

$$\hat{\theta} = g(H(x, y)),$$

for some functions g and H. This would obtain if $f(x, y \mid \theta)$ were of exponential family form (cf. Example 3.1), with invertible likelihood equation given by

$$E\{H(X, Y) \mid \theta\} = H(x, y),$$

where $H(X, Y)$ are the sufficient statistics. Then (4.5) takes the form

$$\theta^{(m+1)} = \frac{1}{T} \sum_{t=1}^{T} g\{H(x_t^{(m+1)}, y)\}.$$

The following represent special cases.

Example 4.1.1

If $T = \infty$, and function g is linear, this takes us back to the EM algorithm, which can therefore be approximated by using the Gibbs sampler with a large but finite value for T. Chalmond (1989) describes a version of this for maximizing the pseudo-likelihood function, rather than the intractable likelihood. A somewhat similar approach was developed by Veijanen (1990).

Example 4.1.2

The case $T = 1$ corresponds to what Celeux & Diebolt (1985) call the Stochastic EM algorithm (SEM) algorithm. In that paper, they apply their method only to the case of standard finite-mixture data, but it is clearly applicable more generally.

In §5, we shall consider in much more detail the implementation and behaviour of these procedures.

Downloaded from rsta.royalsocietypublishing.org

Younes (1989) constructs a stochastic gradient algorithm for the case where the joint distribution of X and Y belongs to the exponential family with sufficient statistics $H(X, Y)$. Then the likelihood equation takes the form

$$E\{H(X, Y) \mid \theta\} = E\{H(X, Y) \mid Y = y, \theta\}.$$

Younes's approach is to generate a sequence of iterations $(\theta^{(m)})$ according to

$$\theta^{(m+1)} = \theta^{(m)} + [1/(m+1)U][H(X^{(m+1)}, Y^{(m+1)}) - H(X_y^{(m+1)}, y)], \tag{4.6}$$

for $m = 0, 1, \ldots$. In (4.6), U is a positive constant, $(X^{(m+1)}, Y^{(m+1)})$ are sampled from $f(X, Y \mid \theta^{(m)})$, and $X_y^{(m+1)}$ is sampled from $\pi(\cdot \mid y, \theta^{(m)})$. Of course, in principle, such samples themselves involve, if the Gibbs sampler is used, many passes over the frame.

Instead, one can run a version of the algorithm with $X^{(m+1)}$ created by applying a single cycle of the Gibbs sampler, with $\theta = \theta^{(m)}$, from $X^{(m)}$, and similarly for $X_y^{(m+1)}$.

The algorithm is clearly a Monte Carlo variant of the technique known as stochastic approximation. The procedure described in Younes (1989) for hidden Markov random fields is a direct development of his procedure for the noise-free Markov random field (Younes 1988). Convergence properties are somewhat complicated but are described in the original papers. For further work along these lines see Moyeed & Baddeley (1991).

4.3. *Other methods*

In view of the perceived difficulty of maximum likelihood estimation, various other approaches have been explored. The method of moments has been implemented on simple models, by Frigessi & Piccioni (1988, 1990) for two-dimensional Ising models with binary channel noise (two parameters altogether) and by Geman & McClure (1987) for one parameter in the context of single-photon-emission computed tomography (SPECT). Pickard (1987) develops asymptotic maximum likelihood estimation, but this approach is available only for the Ising model. Possolo (1986) and Derin & Elliott (1987) note that, for certain noise-free binary images, conditional logits are linear functions of the parameters. This motivates a least-squares approach to parameter estimation which is further discussed by Gray *et al.* (1991).

We now concentrate on and illustrate the Monte Carlo-based RM algorithm introduced in §4.1.

5. Monte Carlo restoration–estimation algorithms

We begin this section by proposing a more general framework for the RM algorithm described in §4.1. Suppose that, when the complete data, (x, y), are available, we estimate θ by

$$\hat{\theta} = g(x, y). \tag{5.1}$$

The function g may or may not be explicit, so that (5.1) formally includes any method such as maximum likelihood, maximum pseudo-likelihood or moment estimation.

Consider now the following Monte Carlo approach to deal with the case where only $Y = y$ is given. The procedure generates a sequence $\{\theta^{(m)}\}$ of iterates, starting from an initial guess, $\theta^{(0)}$.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

R-step (restoration): for specified T, generate samples $x_{(m+1)1}, \ldots, x_{(m+1)T}$ from $\pi(x \mid y, \theta^{(m)})$.

E-step (estimation): two possibilities are considered here,

E_A: take $\theta^{(m+1)} = \frac{1}{T} \sum_{j=1}^{T} g(x_{(m+1)j}, y)$;

E_B: take $\theta^{(m+1)} = g\left(\frac{1}{T} \sum_{j=1}^{T} x_{(m+1)j}, y\right)$.

We describe both of these algorithms as stochastic restoration–estimation (SRE) algorithms and denote them by SRE_A and SRE_B. There may be other variants of the SRE_B algorithm. When (5.1) defines the maximum (pseudo-) likelihood estimate, we refer to the algorithm as the SRM algorithm. Clearly, unless g is 'linear' the SRE_B algorithm, as stated, will be different from the SRE_A algorithm, which is more obviously motivated by (5.1), but this discrepancy often disappears asymptotically, as indicated later.

Example 5.1. Scalar exponential family

Suppose θ is scalar and

$$f(x, y \mid \theta) \propto (1/C(\theta)) \exp\{\theta H(x, y)\}.$$

Let
$$\theta = g(x, y) \tag{5.2}$$

denote the inverse of the likelihood equation, which is

$$C'(\theta)/C(\theta) = H(x, y).$$

Thus $g(x, y)$ is a function of the sufficient statistic, $H(x, y)$. With an abuse of notation, write (5.2) as

$$\theta = g(H(x, y)),$$

which motivates the following variant of the SRM_B algorithm,

$$\theta^{(m+1)} = g\left(\frac{1}{T} \sum_{j=1}^{T} H(x_{(m+1)j}, y)\right). \tag{5.3}$$

In general, g will often be a function only of the appropriate sufficient statistics, and certainly so if $\hat{\theta}$ in (5.1) is the maximum likelihood estimate.

Special cases, in the maximum likelihood context, are as follows.

(i) The case $T = 1$, for which E_A and E_B are equivalent, correspond to the SEM algorithm for mixture data described by Celeux & Diebolt (1985),

(ii) A further special case is discovered if $T \to \infty$ in example (5.1). In that case,

$$\frac{1}{T} \sum_{j=1}^{T} H(x_{(m+1)j}, y) \to E[H(X, y) \mid y, \theta^{(m)}],$$

so that the SRE_B algorithm corresponding to (5.3) is just the EM algorithm itself. For large T, the results should approximate those for the EM algorithm.

It is now appropriate to discuss the properties of the random sequence $\{\theta^{(m)}\}$. Let Ω_y denote the range of estimates $\theta^{(m)}$ for given y; Ω_y is assumed to be the same, for all m. Also, let

$$K(\cdot \mid \theta^{(m)}, y)$$

Downloaded from rsta.royalsocietypublishing.org

denote the conditional density of $\theta^{(m+1)}$, given $\theta^{(m)}$ and y. $K(\cdot\,|\,\theta, y)$ is then related to the conditional density of random variable $g(X, Y)$, given $Y = y$, with parameter θ, according to whichever of our algorithms is in use. Let

$$p_m(\cdot\,|\,y)$$

denote the conditional density of $\theta^{(m)}$ given $y, m = 1, 2, \ldots$. Then

$$p_{m+1}(\cdot\,|\,y) = \int_{\Omega_y} p_m(\tau\,|\,y)\, K(\cdot\,|\,\tau, y)\,\mathrm{d}\tau, \quad m = 1, 2, \ldots.$$

Often, $\{p_m\}$ will converge to the eigenfunction, corresponding to eigenvalue unity, of the integral operator

$$\int_{\Omega_y} K(\cdot\,|\,\tau, y)\, p(\tau\,|\,y)\,\mathrm{d}\tau,$$

in which $K(\lambda\,|\,\tau, y)$ can be regarded as the transition function of a homogeneous Markov chain.

By using the Hilbert projective metric, this convergence can be proved in the case where Ω_y is a closed, bounded region and $K(\lambda\,|\,\tau, y)$ is a positive-valued, continuous function of (λ, τ) on $\Omega_y \times \Omega_y$, provided the initial p_1 is continuous and positive-valued on Ω_y.

The required ergodicity of K cannot be established in any generality but the hope is that, in practice, $\{p_m\}$ converges to an eigendistribution p^* that satisfies

$$p^*(\cdot\,|\,y) = \int_{\Omega_y} K(\cdot\,|\,\tau, y)\, p^*(\tau\,|\,y)\,\mathrm{d}\tau \tag{5.4}$$

and that, 'eventually', $\theta^{(m)}$ can be regarded as being a random deviate from $p^*(\cdot\,|\,y)$. The resulting p^* can be used in many ways, of which the simplest is to use a sample mean of realizations as the 'final' estimate of θ. In general, it not possible to relate $p^*(\cdot\,|\,y)$ to the relationship between y and the true θ.

In this paper, we derive theoretical results only for a very simple example and merely illustrate the usefulness of the methods for Markov random fields.

Example 5.2

Suppose $X_i \sim N(\theta, \sigma_1^2)$, $Y_i\,|\,x_i \sim N(x_i, \sigma_2^2)$, $i = 1, \ldots, n$, that all $\{X_i\}$ are independent, and that (2.6) holds. We assume σ_1^2 and σ_2^2 are known, that the x_i are missing and that θ is of interest.

In this example, exact calculations are feasible, since, marginally, for each i,

$$Y_i \sim N(\theta, \sigma_1^2 + \sigma_2^2)$$

and

$$\hat{\theta} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

is the obvious estimate of θ from data $y_1, \ldots, y_n$.

SRM algorithm with $T = 1$

R-step: for $i = 1, \ldots n$, generate $x_i^{(m+1)}$ from $p(x_i\,|\,y_i, \theta^{(m)})$;

M-step: take $\theta^{(m+1)} = \dfrac{1}{n} \sum_{i=1}^{n} x_i^{(m+1)}$.

Downloaded from rsta.royalsocietypublishing.org

Table 1. *Values of $\bar{y}$, and sample mean and variance of late iterates of $\theta^{(m)}$, for two data-sets corresponding to Example 5.2*

data-set	$\bar{y}$	$\bar{\theta}^{(m)}$	$V(\theta^{(m)})$
1	1.4358	1.4266	0.0363
2	0.3315	0.3379	0.0292

Thus, in the general notation,

$$g(x, y) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

It is straightforward to show that, given y and $\theta^{(m)}$,

$$\theta^{(m+1)} \sim N\left[\frac{\sigma_2^2 \theta^{(m)} + \sigma_1^2 \bar{y}}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{n(\sigma_1^2 + \sigma_2^2)}\right]. \tag{5.5}$$

Relationship (5.5) defines the density $K(\cdot \mid \tau, y)$ and it is easy to check that the corresponding stationary density, $p^*(\cdot \mid y)$ satisfying (5.4) is that of the

$$N\left[\bar{y}, \frac{\sigma_2^2(\sigma_1^2 + \sigma_2^2)}{n(\sigma_1^2 + 2\sigma_2^2)}\right]$$

distribution. If, therefore, the SRM algorithm is carried out to convergence and a sample average of late iterates of $\theta^{(m)}$ is used as an estimate of the parameter, then that sample average should differ from the true maximum likelihood estimate only by an amount quantified by the above stationary distribution.

In a very simple illustration with $n = 20$, $\theta = 1$, and $\sigma_1^2 = \sigma_2^2 = 1$, two sets of data $\{y_i\}$ were generated. For each set, $\{\theta^{(m)} : m = 1, \ldots, 2000\}$ were generated and the sample mean and variance calculated for the last 1000 iterates are as given in table 1.

Note the proximity of $\bar{\theta}^{(m)}$ to $\bar{y}$ in each case, and that the theoretical variance is 0.0333.

SRM algorithm with general T

In this case, $\{x_{it}^{(m+1)} : i = 1, \ldots, n, t = 1, \ldots T\}$ are generated,

$$\theta^{(m+1)} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}^{(m+1)}.$$

Here $K(\cdot \mid \tau, y)$ is the density corresponding to a

$$N\left[\frac{\sigma_2^2 \tau + \sigma_1^2 \bar{y}}{\sigma_1^2 + \sigma_2^2}, \frac{\sigma_1^2 \sigma_2^2}{nT(\sigma_1^2 + \sigma_2^2)}\right]$$

random variable, and $p^*(\cdot \mid y)$ is that for a

$$N\left[\bar{y}, \frac{\sigma_2^2(\sigma_1^2 + \sigma_2^2)}{nT(\sigma_1^2 + 2\sigma_2^2)}\right]$$

random variable. As $T \to \infty$, $p^*(\cdot \mid y)$ becomes degenerate at $\bar{y}$, and the SRM algorithm becomes the EM algorithm.

Downloaded from rsta.royalsocietypublishing.org

We now present a case with a small sample size to show the difference that can arise between the SRM_A and SRM_B algorithms.

Example 5.3. Markov chain with periodic boundary condition and parameter β

Here

$$p(x_1, \ldots, x_5 \,|\, \beta) = \frac{1}{C(\beta)} \exp\left\{\beta \sum_{i=1}^{5} x_i x_{i+1}\right\},$$

where $x_i \in (-1, 1\}, i = 1, \ldots, 5, x_6 = x_1$, and $\beta > 0$. It is easy to show that

$$C(\beta) = 2(\mathrm{e}^{5\beta} + 10\mathrm{e}^{\beta} + 5\mathrm{e}^{-3\beta}).$$

Although we can clearly deal with the likelihood itself, consider estimating β by maximizing the pseudo-likelihood,

$$\prod_{i=1}^{5} p(x_i \,|\, x_{i-1}, x_{i+1}).$$

If

$$S_1(x) = \sum_{i=1}^{5} x_i x_{i+1}, \quad S_2(x) = \sum_{i=1}^{5} \delta(x_{i-1}, x_{i+1}),$$

it is easy to show that the maximum pseudo-likelihood estimate of β satisfies

$$\theta \equiv (\mathrm{e}^{2\beta} - \mathrm{e}^{-2\beta})/(\mathrm{e}^{2\beta} + \mathrm{e}^{-2\beta}) = S_1(x)/S_2(x).$$

Suppose now we observed, not x, but noisy data

$$y = (y_1, \ldots, y_5) = (0.80,\ 0.98,\ 1.01,\ -1.15,\ -0.95),$$

assumed to be modelled by $y_i = x_i + \epsilon_i$, where the $\epsilon_i \sim N(0, 0.36)$, independently.

In this very simple example, it is easy to generate samples from the posterior distribution, in order to implement the following iterative steps of the algorithms:

SRM_A:

$$\theta^{(m+1)} = \frac{1}{T} \sum_{t=1}^{T} \frac{S_{1t}^{(m+1)}}{S_{2t}^{(m+1)}},$$

SRM_B:

$$\theta^{(m+1)} = \sum_{t=1}^{T} S_{1t}^{(m+1)} \Big/ \sum_{t=1}^{T} S_{2t}^{(m+1)},$$

where $S_{1t}^{(m+1)}$ and $S_{2t}^{(m+1)}$ are S_1 and S_2 as obtained from the tth simulated x. The iterative step in the corresponding EM-algorithm is as follows,

EM:

$$\theta^{(m+1)} = \frac{E\{S_1(X) \,|\, y, \theta^{(m)}\}}{E\{S_2(X) \,|\, y, \theta^{(m)}\}}.$$

For the data provided as above, the EM algorithm (as given by the above formula) leads to the maximum (pseudo-)likelihood estimate $\hat{\theta}_{ML} = 0.9866$.

Both the SRM_A and SRM_B algorithms were run for 60 cycles, with $T = 1000$. The empirical means of the last 40 iterates were

$$\hat{\theta}_\text{A} = 0.9954 \quad \text{and} \quad \hat{\theta}_\text{B} = 0.9852,$$

in obvious notation. The latter is clearly close to $\hat{\theta}_{ML}$. Not surprisingly, $\hat{\theta}_\text{A}$ is different from $\hat{\theta}_\text{B}$, but is very close to the solution of

$$\theta^* = E\{S_1(X)/S_2(X) \,|\, \theta^*, y\},$$

namely, $\theta^* = 0.9955$.

Downloaded from rsta.royalsocietypublishing.org

Here, we did not estimate the variance of the noise. If the variance is larger, the difference between the estimates from the two procedures is larger.

Note that the sample size of $n = 5$ is very small. As $n \to \infty$, the ergodic properties of $S_1(X)$ and $S_2(X)$ will imply that $\hat{\theta}_{ML}$ and θ^* will ultimately be the same, so that the results of the SRM_A and SRM_B algorithms will be comparable.

Example 5.4. Binary first-order Markov chain

Suppose

$$f(x\,|\,\beta) = \frac{1}{C(\beta)} \exp\left\{\beta \sum_{i=1}^{n-1} \delta(x_i, x_{i+1})\right\},$$

where $x_i \in \{1, 2\}$ for all i. Given x_i, $Y_i \sim N(x_i, \sigma^2)$, independently for all i. Both β and σ^2 require to be estimated. For illustrative purposes, we consider maximum pseudo-likelihood estimation as the basis of the estimation of β. The log-pseudo-likelihood is

$$\beta S_1(x) - S_2(x) \ln(\mathrm{e}^{2\beta} + 1) - S_3(x)(\beta + \ln 2), \tag{5.6}$$

where

$$S_1(x) = \delta(x_1, x_2) + 2 \sum_{i=2}^{n-2} \delta(x_i, x_{i+1}) + \delta(x_{n-1}, x_n),$$

$$S_2(x) = \sum_{i=2}^{n-1} \delta(x_{i-1}, x_{i+1}),$$

$$S_3(x) = n - 2 - S_2(x).$$

Maximization of (5.6) gives

$$\theta \equiv \mathrm{e}^{2\beta}/(\mathrm{e}^{2\beta} + 1) = (S_1(x) - S_3(x))/2S_2(x). \tag{5.7}$$

Given x and y, σ^2 is estimated by

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2. \tag{5.8}$$

For this example, the partition function can again be evaluated, and is

$$C(\beta) = 2(\mathrm{e}^{\beta} + 1)^{n-1} = 2[\sqrt{((1-\theta)^{-1} - 1)} + 1]^{n-1}.$$

Thus, the likelihood equation results in

$$\mathrm{e}^{\beta}/(\mathrm{e}^{\beta} + 1) = S_4(x)/(n-1),$$

where

$$S_4(x) = \sum_{i=1}^{n-1} \delta(x_i, x_{i+1}).$$

In terms of θ, this gives

$$\theta = \frac{S_4^2(x)}{[n - 1 - S_4(x)]^2 + S_4^2(x)}. \tag{5.9}$$

Qian & Titterington (1990) developed, for this example, a recursive technique for maximizing the likelihood corresponding to the observed data, y. As a result, it is possible to compare the EM, SRM_A and SRM_B algorithm based on both the likelihood and the pseudo-likelihood functions. Results in Qian & Titterington (1990) showed

Downloaded from rsta.royalsocietypublishing.org

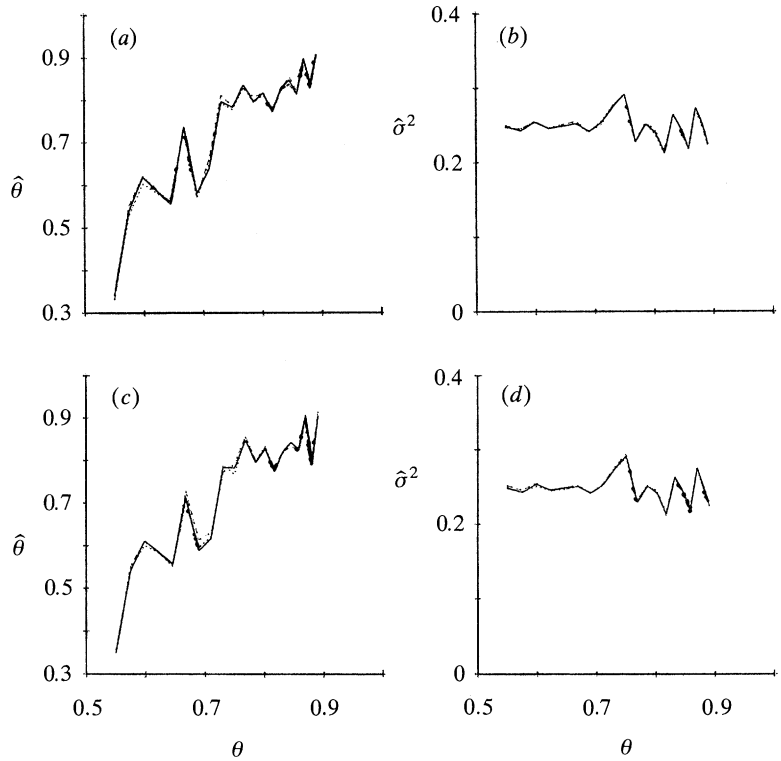

Figure 1. Comparison of EM, SRM_A and SRM_B algorithms for the estimation of θ and σ^2 in Example 5.4 with $\sigma^2 = 0.25$ and $n = 512$; (*a*), (*b*) based on likelihood; (*c*), (*d*) based on pseudo-likelihood. ——, EM; - - - - -, SRM_A; – – – –, SRM_B.

that the EM algorithm associated with the likelihood function provided very good parameter estimates. Here, we make comparisons with the other methods. (The comparison with the EM algorithm may offer useful extrapolation to the case of two-dimensional Markov random fields, for which it is impossible to carry out the EM algorithm.)

The ergodic properties of the statistics that appear in (5.7)–(5.9) should ensure that, for large n, the EM, SRM_A and SRM_B algorithms should perform similarly, whether the likelihood or the pseudo-likelihood is used. Accordingly, the value, 5, chosen for T was much less than was the case in Example 5.3.

Figure 1 displays results for $\sigma^2 = 0.25$ and $n = 512$. The value of $\theta = \mathrm{e}^{2\beta}/(\mathrm{e}^{2\beta}+1)$ was varied between 0.55 and 0.9. For each choice of parameters a single realization for x was generated and noise added to create y. It is clear that the results from the EM, SRM_A and SRM_B algorithms are very similar, and that the results based on likelihood are very similar to those based on pseudo-likelihood. For each procedure, except that associated the EM algorithm, 50 iterative cycles were carried out, and the quoted estimates are the averages of the last 40 cycles.

Figure 2 reports averages of the results based on 40 realizations of (x, y) for each choice of parameters, this time with $\sigma^2 = 0.09$ and $n = 64$. In this case, $T = 15$, 50 iterative cycles were carried out for each realization and the parameters were estimated by the averages of the final 30 iterates. Once again, all methods gave very similar results.

Downloaded from rsta.royalsocietypublishing.org

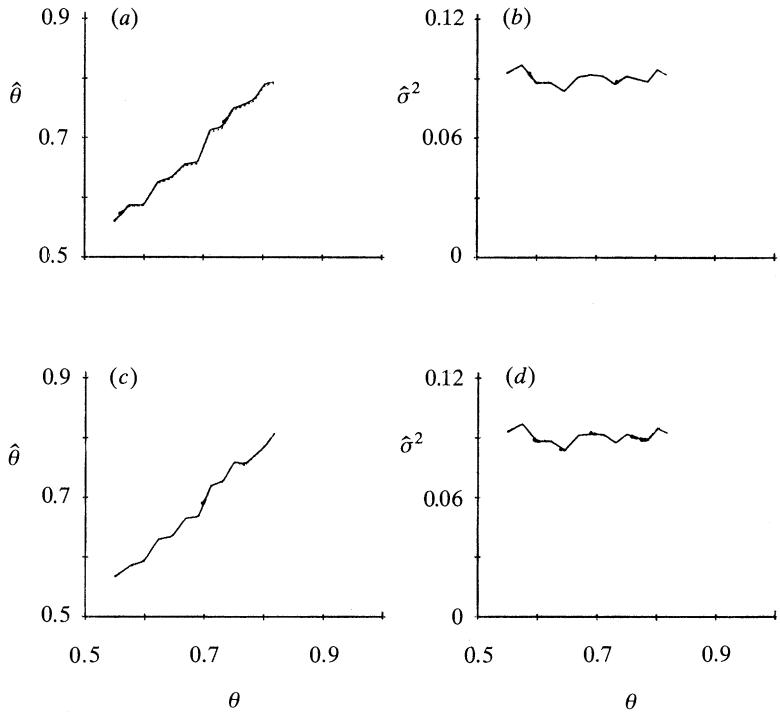

Figure 2. Comparison of EM, SRM_A and SRM_B algorithms for the estimation of θ and σ^2 in Example 5.4 with $\sigma^2 = 0.09$ and $n = 64$; (*a*), (*b*) based on likelihood; (*c*), (*d*) based on pseudo-likelihood. ——, EM; - - - - -, SRM_A; – – – –, SRM_B.

On a theoretical level, there are consistency results for maximum pseudo-likelihood estimation in the noise-free case (Geman & Graffigne 1987) but little has been established, as yet, for the case of noisy data.

6. An illustration using satellite data

The top four pictures in figure 3 display a four-band satellite image of the Lake of Menteith (Scotland's only 'Lake' rather then 'Loch'!) in Perthshire. As an illustrative exercise, we shall use our methodology to classify the pixels in 64×64 frame into a six-state scene. Thus, for each i, $x_i \in \{1, 2, \ldots, 6\}$. As a prior for X we use the first-order Markov random field, with probability function

$$p(x \mid \beta) = \frac{1}{C(\beta)} \exp\left\{\beta \sum_{i \sim j} \delta(x_i, x_j)\right\}, \tag{6.1}$$

where the summation is over nearest-neighbour pairs. It is arguable that this prior is unrealistically simple, in that it treats all pixel states symmetrically. The choice of six for the number of states is also somewhat arbitrary but serves well for illustrative purposes.

Each y_i is four-dimensional and it is assumed that, for a pixel in state $k (x_i = k)$,

$$y_i \sim N(\mu_k, V),$$

where V is a 4×4 covariance matrix. There are therefore 24 unknown parameters

Downloaded from rsta.royalsocietypublishing.org

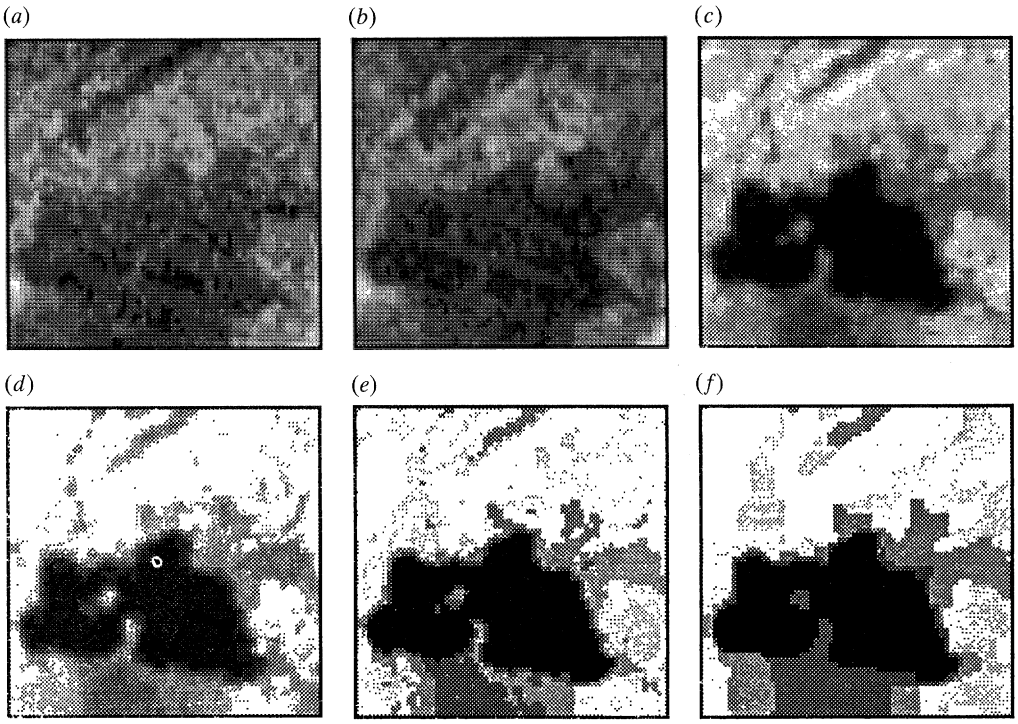

Figure 3. Satellite data (*a*)–(*d*) data from bands 1–4; (*e*) initial restoration from band 3 data; (*f*) final SRM restoration from band 3 data.

associated with the $\{\mu_k\}$ and 10 parameters including in V. Altogether, therefore, including the parameter β, there are 35 unknown parameters.

In the iterative procedure, we shall deal with β itself, rather than any transformation thereof, as we did for Examples 5.3 and 5.4.

The procedure is initiated by choosing a starting classification. In our illustration, we achieved this by taking the band 3 image alone, dividing the observed range into six equal ranges, and regarding these as the six states (a multi-state thresholding exercise.) The resulting restoration is presented in figure 3*e*. We report here only results from the SRM_A method, starting from the above restoration and using only the band-3 data. In this case there are only eight parameters (six means, one variance and β). The starting parameters were obtained by regarding the initial restoration as the true scene. The routine in each iterative stage was as follows.

(i) Run 50 cycles of the (posterior) Gibbs sampler starting from the current restoration.

(ii) Run the Gibbs sampler for further 100 cycles, estimating the parameters after each 10 cycles.

(iii) Obtain the average of the 10 sets of estimates and use them to start the next iteration. (Thus, in the previous notation, $T = 10$.)

Note that, at any stage, there is a current restored image. The restoration achieved after five of the above iterative cycles is displayed in figure 3*f*.

Next, we used this segmentation to initiate the SRM_A procedure using the four-dimensional data. Again, initial parameters were estimated by using this initial image. Note that the estimation of those mean-parameters and the covariance-

Downloaded from rsta.royalsocietypublishing.org

Table 2. *Satellite data: parameter estimates for mean vectors μ_k corresponding to six states $\{S1, \ldots, S6\}$, and for the covariance matrix V using (a) the* SRM$_A$ *algorithm, (b) the* PPL-ICM *algorithm, (c) the* PPL-EM-ICM *algorithm*

	S1	S2	S3	S4	S5	S6
(*a*) SRM parameters; $\hat{\beta} = 2.571$; $\{\hat{\mu}_k\}$						
band 1	27.49	27.80	28.53	31.96	30.53	31.58
band 2	20.72	21.85	23.40	28.27	25.54	27.03
band 3	33.69	58.82	75.55	88.78	97.06	110.52
band 4	14.65	46.73	69.85	85.09	95.88	114.40
$\hat{V}$						
		1.72	1.49	2.17	1.69	
		1.49	3.96	0.72	−0.75	
		2.17	0.72	27.42	45.81	
		1.69	−0.75	45.81	66.02	
(*b*) PPL-ICM parameters; $\hat{\beta} = 2.736$; $\{\hat{\mu}_k\}$						
band 1	27.46	27.83	28.21	31.17	31.21	31.69
band 2	20.69	21.80	22.86	27.16	26.55	27.11
band 3	33.47	57.75	71.53	85.97	99.24	114.76
band 4	14.42	45.11	64.85	82.18	98.55	118.92
$\hat{V}$						
		1.98	1.88	1.43	0.79	
		1.88	4.49	0.01	−1.55	
		1.43	0.01	24.72	32.10	
		0.79	−1.55	32.10	51.47	
(*c*) PPL-EM-ICM parameters; $\hat{\beta} = 2.606$; $\{\hat{\mu}_k\}$						
band 1	27.47	27.82	28.23	31.21	31.18	31.65
band 2	20.70	21.80	22.89	27.23	26.48	27.03
band 3	33.54	57.91	71.80	86.23	99.30	114.61
band 4	14.50	45.33	65.16	82.46	98.65	118.84
$\hat{V}$						
		1.98	1.88	1.62	1.02	
		1.88	4.47	0.34	−1.16	
		1.63	0.34	26.21	33.63	
		1.02	−1.16	33.63	52.98	

parameter is trivial. The algorithm was as above except that 30 iterative cycles were carried out, rather than 5. The averages of the parameter estimates obtained in the last 25 cycles are listed in table 2*a*.

It should be remarked that, in the SRM$_A$ algorithm as originally described, the T realizations of the posterior distribution are expected to be mutually independent. In the case of a two-dimensional Markov random field, it is computationally very expensive to generate many independent samples. However, two realizations, suitably far apart in a single operation of the Gibbs sampler, will be only loosely dependent, a phenomenon which we exploited in our calculations.

Once we have settled on a 'final' set of estimates of the parameters, in the present case the values in table 2*a*, there is more than one way of producing a 'final'

Downloaded from rsta.royalsocietypublishing.org

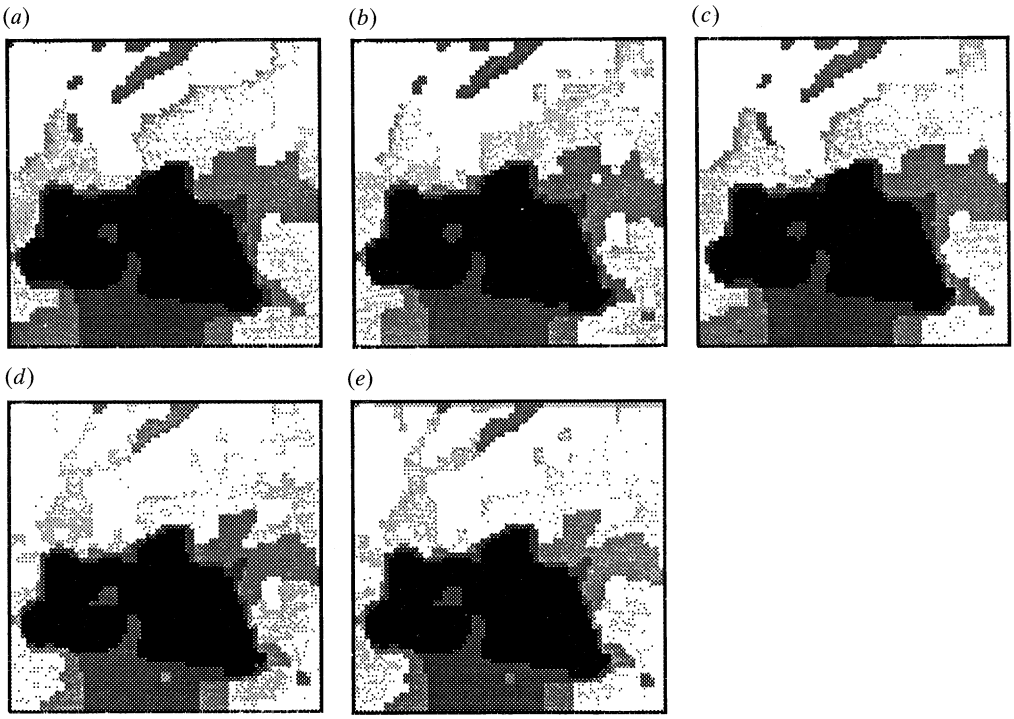

Figure 4. Satellite data: (*a*) final restoration from SRM phase; (*b*) ICM restoration using SRM estimates; (*c*) Gibbs sampler restoration using SRM estimates; (*d*) PPL-ICM restoration; (*e*) PPL-EM-ICM restoration.

classification for the pixels. Here, we report two methods, the former being the result of applying 25 cycles of Besag's (1986) ICM algorithm and the second involving implementation of an idea that originates in Besag *et al.* (1991). In the latter, many samples from the estimated posterior distribution are generated (we ran the Gibbs samplers for 500 cycles), and the most frequently generated state at any given pixel is used to create the restoration; in fact, we noted the states only after every fifth sampling cycle. We initiated both these procedures with the 'final' restoration created during the estimation phase. That particular image and the results of implementing the two restoration stages are presented in figure 4. The three images show very few mutual differences.

Also displayed in figure 4 are the results of two further estimation–restoration algorithms, the decision-directed PPL-ICM algorithm of Besag (1986) and the PPL-EM-ICM algorithm of Qian & Titterington (1991). Both were initiated from the initial restoration from the band 3 data, shown as figure 3*e*. The corresponding sets of parameter estimates resulting from the two procedures are shown in table 2*b*, *c*. The two sets of estimates are very similar, as are the two restorations. The reason for this may be that the noise is small compared with the 'original' image, bearing in mind the estimated covariance matrix and the mean-parameters; for example, for all three procedures, the differences of means between state 5 and state 6, associated with band 4 data, are about 20.0, while the standard deviations associated with band 4 are about 8.0 or less.

Downloaded from rsta.royalsocietypublishing.org

7. Discussion

In this paper, we have highlighted the difficulties involved in estimating parameters in hidden Markov models and we have reviewed methods that have been proposed for dealing with them. Detailed discussion and illustration have been presented of a subclass of the methods that have a strongly Monte Carlo flavour. Further research is required to consolidate these methods, in term of theory and implementation. However, the results reported here and, in particular, the account of Example 5.4, suggest that SRM algorithms based on pseudo-likelihood may be valuable alternatives to EM algorithms in problems in which the latter are totally impracticable.

The ideas should also be useful for the estimation of parameters in other contexts where models with interactive characteristics are essential. Prime examples are the fields of probabilistic expert systems (Lauritzen & Spiegelhalter 1988) and the use of statistical modelling in neural networks (see for instance Bourlard 1990).

W. Q. was supported by a research grant from the UK Science and Engineering Research Council, awarded under the Complex Stochastic System Initiative. The same agency funded the equipment on which the computational work was carried out. The authors are very grateful to Professor Julian Besag for his helpful comments on an earlier draft of the paper, and to Dr Jim Kay for kindly providing the data for the Lake of Menteith example.

References

Aykroyd, R. G. & Green, P. J. 1991 Global and local priors, and the location of lesions using gamma-camera imagery. *Phil. Trans. R. Soc. Lond.* A **337**, 323–342. (This volume.)

Bartlett, M. S. 1971 Physical nearest-neighbour models and non-linear time-series. *J. appl. Prob.* **8**, 222–232.

Besag, J. E. 1975 Statistical analysis of non-lattice data. *Statistician* **24**, 179–195.

Besag, J. E. 1976 Parameter estimation for Markov fields. Tech. Rep. 198, Series 2. Dept Statistics, Princeton Univ.

Besag, J. E. 1986 On the statistical analysis of dirty pictures (with discussion). *Jl R. statist. Soc.* B **48**, 259–302.

Besag, J. E., York, J. & Mollie, A. 1991 Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Bourlard, H. E. 1990 How connectionist models could improve Markov models for speech recognition. In *Advanced neural computers* (ed. R. Eckmiller), pp. 247–254. Amsterdam: North Holland.

Celeux, G. & Diebolt, J. 1985 The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Q.* **2**, 73–82.

Chalmond, B. 1989 An iterative Gibbsian technique for reconstruction of M-ary images. *Pattern Recognition* **22**, 747–761.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Jl R. statist. Soc.* B **39**, 1–38.

Derin, H. & Elliott, H. 1987 Modelling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Trans. Pattern Analysis Machine Intell.* **PAMI-9**, 39–55.

Frigessi, A. & Piccioni, M. 1988 Parameter estimation for two-dimensional Ising fields corrupted by noise. Quaderno no. 12, IAC-CNR, Rome.

Frigessi, A. & Piccioni, M. 1990 Consistent parameter estimation for 2-D Ising fields corrupted by noise: numerical experiments. In *Proc. IMS Conf. Spatial Statistics and Imaging* (ed. A. Possolo). Haywood, California: Institute of Mathematical Statistics.

Geman, S. & Geman, D. 1984 Statistical relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis Machine Intell.* **PAMI-6**, 721–741.

Downloaded from rsta.royalsocietypublishing.org

Geman, S. & Graffigne, C. 1987 Markov random field image models and their applications to computer vision. In *Proc. Int. Congress of Mathematicians* (ed. A. M. Gleason), pp. 1498–1517. Berkeley, California: American Mathematical Soc.

Geman, S. & McClure, D. 1987 Statistical methods for tomographic image reconstruction. *Bull. Int. statist. Inst.* (BK. 4) **52**, 5–21.

Geyer, C. J. & Thompson, E. A. 1992 Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Jl R. statist. Soc.* B **54**. (In the press.)

Gray, A. J., Kay, J. W. & Titterington, D. M. 1991 On the estimation of noisy binary Markov random fields. (Submitted.)

Greig, D. M., Porteous, B. T. & Seheult, A. H. 1989 Exact maximum a posterior estimation for binary images. *Jl R. statist. Soc.* B **51**, 271–279.

Juang, B. H. & Rabiner, L. R. 1991 Hidden marker models for speech recognition. *Technometrics* **33**, 251–272.

Lauritzen, S. L. & Spiegelhalter, D. J. 1988 Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Jl R. statist. Soc.* B **50**, 157–224.

Little, R. J. A. & Rubin, D. B. 1983 On jointly estimating parameters and missing values by maximizing the complete data likelihood. *Am. Statist.* **37**, 218–220.

Little, R. J. A. & Rubin, D. B. 1987 *Statistical analysis with missing data*. New York: Wiley.

Marriott, F. H. C. 1975 Separating mixtures of normal distributions. *Biometrics* **31**, 767–769.

Moyeed, R. A. & Baddeley, A. J. 1991 Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18**, 39–50.

Pickard, D. K. 1987 Inference for discrete Markov fields: the simplest nontrivial case. *J. Am. statist. Ass.* **82**, 90–96.

Possolo, A. 1986 Estimation of binary Markov random fields. Tech. Rep. no. 77, Dept Statistics, University Washington, Seattle.

Ripley, B. D. 1988 *Statistical inference for spatial processes*. Cambridge University Press.

Qian, W. & Titterington, D. M. 1990 Parameter estimation for hidden Gibbs chains. *Statist. Prob. Lett.* **10**, 49–58.

Qian, W. & Titterington, D. M. 1991 Stochastic relaxations and EM algorithms for Markov random fields. *J. statist. Comput. Simul.* **41**. (In the press.)

Smith, A. F. M. 1991 Bayesian computational methods. *Phil. Trans. R. Soc. Lond.* A **337**, 369–386. (This volume.)

Titterington, D. M. 1984 Comments on a paper by S. C. Sclove. *IEEE Trans. Pattern Analysis Machine. Intell.* **PAMI-6**, 656–658.

Titterington, D. M. 1990 Some recent research in the analysis of mixture distributions. *Statistics* **21**, 619–641.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. 1985 *Statistical analysis of finite mixture distributions*. London and New York: Wiley.

Veijanen, A. 1990 An estimator for imperfectly observed Markov fields. Research Rep. no. 74, Dept Statistics, University of Helsinki.

Younes, L. 1988 Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincare* **24**, 269–294.

Younes, L. 1989 Parametric inference for imperfectly observed Gibbsian fields. *Prob. Theory Rel. Fields* **82**, 625–645.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

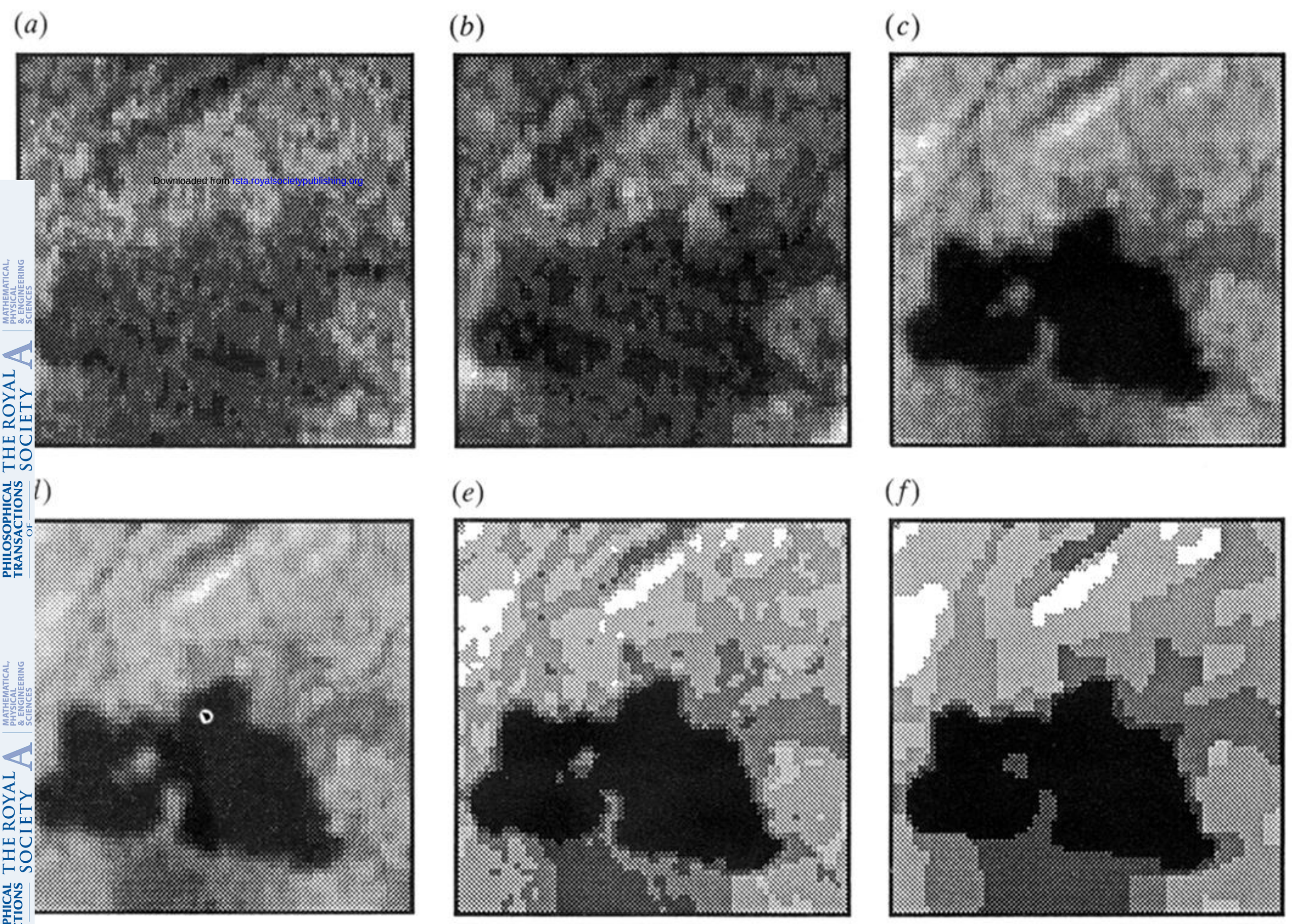

Figure 3. Satellite data (*a*)–(*d*) data from bands 1–4; (*e*) initial restoration from band 3 data; (*f*) final SRM restoration from band 3 data.

Downloaded from rsta.royalsocietypublishing.org

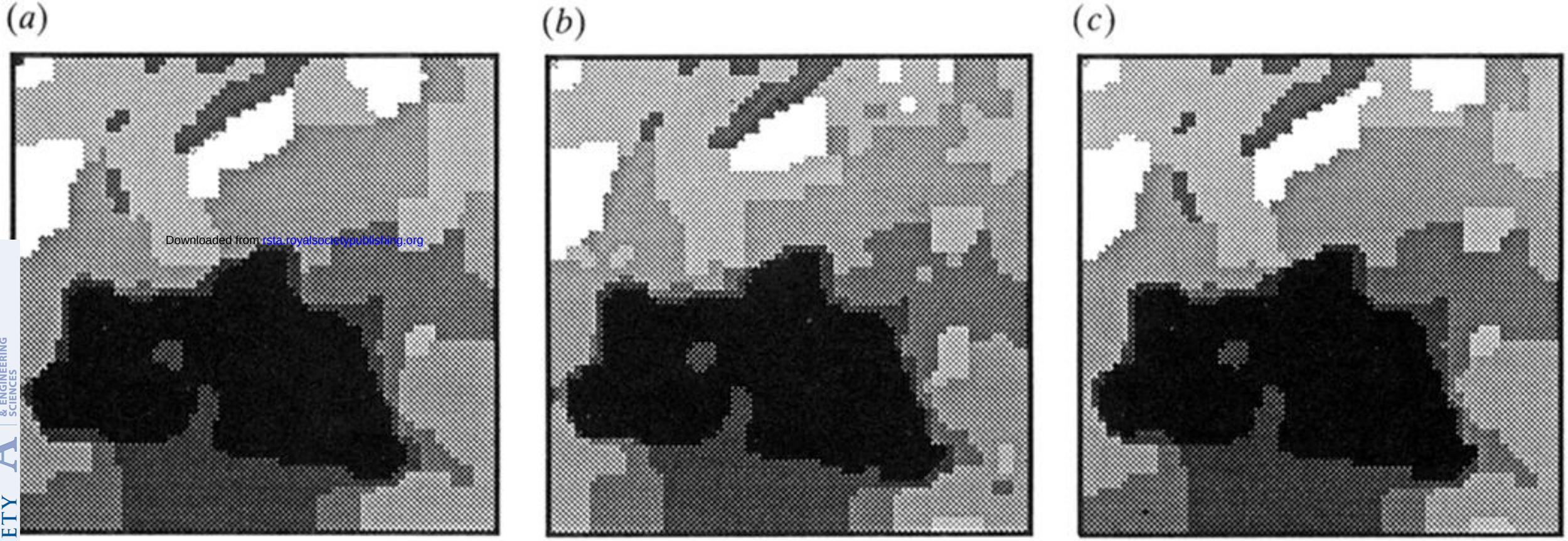

Figure 4. Satellite data: (*a*) final restoration from SRM phase; (*b*) ICM restoration using SRM estimates; (*c*) Gibbs sampler restoration using SRM estimates; (*d*) PPL-ICM restoration; (*e*) PPL-SRM-ICM restoration.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES