

SOME ASPECTS OF STATISTICAL IMAGE MODELLING AND RESTORATION

D. M. TITTERINGTON

Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland

E-mail: mike@stats.gla.ac.uk

A review is provided of some ways in which statistical ideas have influenced research into image analysis, in particular the problems involved in making inferences about the true scene and any parameters in the underlying model. The emphasis will be on the application of general statistical paradigms such as maximum likelihood, implemented using tools such as the EM algorithm, and Bayes' Theorem. The resulting procedures can be regarded as particular recipes for regularisation or deconvolution, according to the context.

1 Introduction

Statisticians have made a substantial impact on image modelling and analysis. The purpose of this paper is to give a quick overview of some of these contributions, to emphasise the key statistical issues, to highlight some points of contact with Physics and to mention a few applications, including one or two from contexts within Physics. The account cannot claim to cover all methodological approaches or all application areas, and there is a strong element of selectivity in the reference list! For instance, a large body of recent work in shape analysis (Dryden and Mardia ¹) is largely overlooked.

Much of the paper will concern pixellated images, and the associated notation is as follows: the true scene is denoted by x , the observed image by y , and each can be written as a vector of length N , where N denotes the number of pixels. However, it is also natural to regard the elements of both x and y as being originally arrayed as $R \times C$ matrices, where R and C denote the numbers of rows and columns of pixels and $N = RC$.

The essence of the statistical approach is to introduce probabilistic models that might plausibly represent the relationship between x and y , and possibly also the structure of x itself, and to use 'standard' statistical paradigms to make inference about the unknown x . In Section 2, we concentrate on a particularly simple model for the way in which y is created as a distorted version of x , and a notional model for x appears almost incidentally. Section 3 reviews some now-classical material in which the model for x is proposed from the outset. Section 4 returns to a starting point similar to that of Section

2 and reviews some material under the heading of 'deconvolution'.

2 A Regularization Approach to Image Restoration

The simplest and most usual model that underlies this approach is that, for some square matrix H ,

$$y = Hx + \epsilon. \quad (1)$$

According to this model, the true scene is deterministically blurred by H and subjected to additive noise ϵ , often assumed to be white noise, so that $\epsilon \sim N(0, \sigma^2 I)$, in which σ^2 is a variance parameter and I denotes the identity matrix. The model makes sense only if the true scene and the observed image can be regarded as continuous intensities; this might be a reasonable approximation for grey-level images, but obviously not if, for example, the true scene is binary. If H , which is a characteristic of the observing instrument, is known, then the natural estimator of the true intensities is

$$\hat{x} := H^{-1}y = x + H^{-1}\epsilon,$$

which corresponds to the least-squares estimator of x , the minimiser of

$$\Delta(x, y) := \|y - Hx\|^2.$$

Since $E\hat{x} = x$, \hat{x} is what is called an unbiased estimator of the true scene. The matrix H might be fairly sparse, if the amount of blurring is small. However, the high dimensionality of the problem, in imaging contexts, can lead to \hat{x} being a very unstable estimator of high variability, and a common solution is to minimise instead

$$\Delta(x, y) + \beta\Phi(x), \quad (2)$$

where Φ is a measure that is intended to penalise 'roughness' in \hat{x} , and β is usually a positive scalar,

although a version involving a high-dimensional β is described by MacKay² under the nomenclature of automatic relevance determination. In a computationally-convenient simplest version of this method, $\Phi(x) = x^\top Cx$, in which C is typically positive-definite, and the ‘optimum’ restoration is given explicitly by

$$\hat{x}_\beta := (H^\top H + \beta C)^{-1} H^\top y. \quad (3)$$

Such an estimator is biased, in that $E\hat{x}_\beta \neq x$, for any $\beta \neq 0$, but its instability, as measured by variance, is greatly reduced. A key decision is to choose β appropriately, and a variety of rationales have been investigated. The choice of C reflects the type of smoothness imposed, and is usually designed to penalise differences between intensities on neighbouring pixels or simply, with $C = I$, to penalise ‘large’ x .

Some of the approaches to the choice of β seek to compromise between low bias and low variance, and a natural criterion for choice is a ‘minimum risk’ criterion of the form

$$\min_\beta E_{y|x} \delta(x, \hat{x}_\beta),$$

where δ is a measure of distance. If $\delta(x, x') = \|x - x'\|^2$ then this amounts to a minimum total mean squared error criterion. The operational difficulty with this is that the criterion and therefore the minimising β are functions of the true x , which of course is unknown! One possibility is to substitute a preliminary estimate \tilde{x} for x at this point, or to apply a method of crossvalidation, which leads to the use of a criterion that measures the ability to predict individual (pixel) observations, given the data on all other pixels. For example, one might define

$$CV(\beta) := N^{-1} \sum_{i=1}^N \{y_i - E(y_i | \hat{x}_\beta^{(i)})\}^2,$$

in which y_i is the i th element of y and $\hat{x}_\beta^{(i)}$ is the restoration computed from all observations except for y_i , and choose β to minimise $CV(\beta)$. A slight modification of this which has been very popular in practice is the generalised crossvalidation function of Golub et al.³ When δ is a simple quadratic loss function this crossvalidation function is given by

$$GCV(\beta) := RSS(\beta) / [\text{tr}\{I - K(\beta)\}^2],$$

in which $K(\beta) = H(H^\top H + \beta C)^{-1} H^\top$, $K(\beta)y = E(y | \hat{x}_\beta)$ denotes the set of ‘fitted values’, and

$RSS(\beta) := \|\{I - K(\beta)\}y\|^2$ is the residual sum of squares. Generalised crossvalidatory choice selects $\beta = \hat{\beta}_{GCV}$ to minimise $GCV(\beta)$. Other methods exist, including the so-called ‘empirical degrees of freedom’ choice, $\hat{\beta}_{EDF}$, defined as the solution of the equation

$$RSS(\beta) / \text{tr}\{I - K(\beta)\} = \sigma^2,$$

that is,

$$RSS(\beta) = \{n - \text{tr}K(\beta)\}\sigma^2,$$

provided σ^2 is known or can be estimated reliably externally. The quantity $n - \text{tr}K(\beta)$ is called the equivalent degrees of freedom for error, by analogy with the corresponding version in ordinary linear models. One justification for this is that Wahba⁴ recommends $RSS(\hat{\beta}_{GCV}) / \text{tr}\{I - K(\hat{\beta}_{GCV})\}$ as an estimator of σ^2 .

So far as the non-statistical regularisation literature is concerned, a traditional way of choosing β is to solve

$$RSS(\beta) = n\sigma^2,$$

justified on the grounds that the *true* $RSS(\beta)$ has expectation $n\sigma^2$. This method will clearly oversmooth relative to $\hat{\beta}_{EDF}$.

In its most basic form the above methodology corresponds to ridge-regression, in general it amounts to an approach to solving potentially ill-posed inverse problems, and the structure appears in the development of smoothing splines as well as in the image-analysis context. My own involvement has included investigation of the various form of choosing β in the contexts of smoothing splines (Hall and Titterton⁵) and images (Hall and Titterton⁶; Thompson et al.⁷).

3 Bayesian Image Analysis

The foundations of what became known as Bayesian image analysis are the seminal papers by Geman and Geman⁸ and Besag⁹. In these papers, a ‘prior’ (marginal) model, $p(x|\beta)$, was assumed for the true scene, x , and a model was also assumed for the observed image, y , conditional on x , to represent the noise and/or blurring process; this model will be denoted by $p(y|x, \theta)$. Thus, for example, for the model defined in (1) $p(y|x, \theta)$ is the multivariate Gaussian

density with mean vector Hx and covariance matrix $\sigma^2 I$, and $\theta = (H, \sigma^2)$. (Note that we are using ‘ p ’ generically to denote a probability density function.)

The marginal model chosen for x typically reflects local spatial correlation, and usually corresponds to a Markov random field. The simplest scenario is that of a binary (black/white) image, with each $x_i = -1$ or $+1$. In this case a possible choice is to take $p(x|\beta)$ to correspond to the Ising model:

$$p(x|\beta) = \{C(\beta)\}^{-1} \exp(\beta \sum_{i \sim j} x_i x_j),$$

in which the sum is over neighbouring pairs of pixels. This corresponds to a so-called first-order Markov random field, and the observed image to a hidden Markov random field. The quantity $C(\beta)$ is a normalising constant for $p(x|\beta)$, also called a partition function. Its calculation is at best a complicated computational problem and leads to difficulties in inference, as mentioned later.

The quantities β and θ represent parameters and, for brevity, we shall denote the complete set of parameters by $\psi = (\beta, \theta)$. (In practice part or all of θ may be known from the specification of the observing instrument.)

There are two other probability distributions of interest, which we shall denote by $p(x|y, \psi)$ and $p(y|\psi)$. Both of these can be expressed in terms of the joint probability function for x and y , which is given by the product of $p(y|x, \theta)$ and $p(x|\beta)$:

$$p(x|y, \psi) \propto p(x, y|\psi) = p(y|x, \theta)p(x|\beta) \quad (4)$$

and

$$p(y|\psi) = \int p(y|x, \theta)p(x|\beta)dx, \quad (5)$$

where, in (5), the integration is over x and represents a summation if x is discrete.

Relationship (4) is the source of ‘Bayesian’ inference about the underlying (hidden) true scene, whereas (5) is the likelihood function corresponding to the observed data, and is important in making inferences about the underlying, and usually unknown, parameters ψ . I have put ‘Bayesian’ in inverted commas in the previous sentence because it is arguably in conflict with what Bayesian inference means in

statistical science. The Bayesian paradigm is characterised by the assignment of probability distributions to *parameters*, which are fixed but unknown, as well as to random variables, realised values of which represent a major component of the experimental data. Pre-experiment ideas about the parameters are summarised by the ‘prior’ distributions and Bayes’ Theorem is used to combine the prior information with that provided by the experimental data to give the ‘posterior’ distribution of the parameters. In what has become known as ‘Bayesian image analysis’, the key unknowns are not really parameters but are the true scene, x , which are perhaps better referred to as hidden or missing values. Bayes’ Theorem is used to construct the conditional distribution $p(x|y)$ from the reverse conditional distribution $p(y|x)$, corresponding to the distortion/noise model, together with the marginal model $p(x)$ for x . Of course, Bayesian inference, as statisticians know it, is one way of dealing with any unknown parameters within ψ .

If for the time being the parameters ψ are assumed known, then, ideally, one should make inferences about the true scene on the basis of $p(x|y, \psi)$. Early work concentrated on obtaining point estimates, such as the mode, using simulated annealing techniques (Geman and Geman⁸), or mode-like quantities (Besag⁹), but in principle the whole joint posterior distribution of x is available for exploitation. Just what is feasible in practice depends to some extent on what is meant by x . In low-level, pixel-based modelling, which was the case considered by Geman and Geman⁸ and Besag⁹ and which we are dealing with in this paper, x contained values, such as colours or intensities, associated with all individual pixels, possibly supplemented by inter-pixel edge indicators. Thus x is of extremely high dimension and it is not feasible to look at complicated features of $p(x|y, \psi)$. Modelling at a higher level is typified by the deformable-templates approach, originally conceived of by Grenander (Grenander *et al.*¹⁰; Grenander and Miller¹¹), in which features in images are represented by skeletal frameworks summarised by a comparatively small (at least relative to the number of pixels!) number of quantities. One would also like to obtain interval estimates concerning important features of the true scene. Typically, $p(x|y, \psi)$ is not of a form that is amenable to ex-

act analysis, but, in principle but still a daunting prospect in practice, Markov chain Monte Carlo methods allow realisations to be simulated from the distribution to be generated and quantities of interest to be estimated by empirical counterparts.

There are clear links between this formulation and Physics. As mentioned earlier, it is natural for the ‘prior’ $p(x|\beta)$ to reflect local association and, for a pixellated binary scene, the Ising model from statistical physics is often used as the prior; if the underlying scene is defined in terms of a known finite number of colours or land-types then a Potts model might be used. Furthermore many of the Markov chain Monte Carlo methods have their origins in physics; for example, what statisticians know as the Gibbs sampler is the same as the heat-bath method.

If the parameters ψ are unknown, then they have to be estimated from the available data, namely y . Various ad hoc methods have been used in the image-analysis context, but the statistician would prefer to implement a general paradigm, either likelihood-based or Bayesian.

In the likelihood approach, the appropriate estimator of ψ is the maximiser of $p(y|\psi)$, and the interpretation of the problem as a missing-data problem, with the true scene x being missing, makes available the general iterative EM algorithm of Dempster *et al.*¹² Let $L(x, y|\psi)$ denote the complete-data log-likelihood, given by

$$L(x, y|\psi) = \log\{p(y|x, \theta)p(x|\beta)\}.$$

Then the EM algorithm is as follows, if we envisage an iteration at stage m , with $\psi^{(m-1)}$ as the current approximation to the maximum likelihood estimate.

1. **E-step:** calculate $Q(\psi) = E_m L(x, y|\psi)$, where the expectation is respect to the conditional distribution represented by $p(x|y, \psi^{(m-1)})$.
2. **M-step:** find $\psi = \psi^{(m)}$ to maximise $Q(\psi)$.

In the E-step, therefore, we evaluate the expectation of the complete-data log-likelihood, conditional on the observed data and using the model based on the current estimates of the parameters to do the averaging, and then in the M-step we maximise that expected log-likelihood in order to obtain the next

set of estimates. One hopes that the sequence $\{\psi^{(m)}\}$ converges to the maximiser of $p(y|\psi)$; it is generally true that the sequence of likelihoods $\{p(y|\psi^{(m)})\}$ is monotonically increasing. As a result, convergence to at least a local maximum is ensured, except in very pathological circumstances.

For the EM algorithm to be easy, both the E-step and the M-step have to be straightforward, and unfortunately in the case of a hidden Markov random field this is true of neither step. It is not possible to obtain an explicit formula for the expectation in the E-step. One approximating alternative is to use a sample average, based on a number of realisations from the relevant distribution, but generation of each of these realisations requires a Markov chain Monte Carlo procedure. Another approach is to use an approximating measure based on so-called mean-field approximations. Here, the averaging measure is a suitably chosen fully-factorised independence model for the individual elements in x . (The mean-field approximation is another tool with its origins in Physics.) Although the use of an independence model might seem to represent a gross approximation to a typically highly complex multivariate distribution, its performance within the E-step of the EM Algorithm can be uncannily effective; see, for instance, Zhang^{13,14}.

Difficulties also arise in the M-step, although, maximisation with respect to θ , the parameters within the noise model, is often easy. However, this is not the case for β , the prior parameters, because $p(x|\beta)$ of the normally intractable β -dependent normalisation constant or partition function present in $p(x|\beta)$. Zhang suggests using mean-field approximations at this stage too. Other possibilities are to approximate the normalisation constant by an empirical average, as explained by Geyer and Thompson¹⁵, or to replace $p(x|\beta)$ by Besag’s¹⁶ pseudo-likelihood, which is defined as

$$p_{PL}(x|\beta) = \prod_i p(x_i|x_{\partial i}, \beta),$$

where $x_{\partial i}$ denotes values associated with the *neighbouring* pixels to pixel i , according to the neighbourhood system defined by the Gibbs distribution $p(x|\beta)$. Thus, p_{PL} is defined by the product of the full conditional distributions of the individual x_i ’s,

and the problem of the intractable partition function disappears. Maximum pseudo-likelihood estimators are often consistent, in that for large lattices the estimator is likely to be close to the true β , but may have rather low efficiencies. One application of the pseudo-likelihood is to use it in the M-step of the EM-algorithm as a replacement for the correct but intractable $p(x|\beta)$; see Qian and Titterton¹⁷ for this and other ways of making the EM-algorithm practicable.

At this point we mention a few practical applications.

Qian and Titterton¹⁷ considered four-band satellite image data of a view of the Lake of Menteith in Perthshire, the only substantial body of water in Scotland referred to as a ‘Lake’ rather than a ‘Loch’! A six-state Potts model was assumed for the true scene and additive Gaussian noise was assumed. An ad hoc initial six-state segmentation was constructed, based on the band-3 data alone, this led to a more refined restoration, again based only on the band-3 data, and then to a number of restorations based on all the data, obtained using various versions of the above EM-type methodology.

Qian and Titterton¹⁸ analysed magnetic induction data corresponding to cobalt-nickel evaporated tape, a high-density magnetic recording material. The image included a transition boundary and it is important to identify the boundary as precisely as possible. Altogether three ‘restorations’ were created, based on different ways of modelling the surfaces on either side of the boundary, with the estimated boundary identified.

Data from a transmission electron microscopy image were also examined in Qian *et al.*¹⁹ The image depicted a magnetic domain, the ideal shape of which would be that of a tilted circle, that is, an ellipse. In the paper a number of models were proposed, and restorations were obtained. A key feature of the ‘prior’ model reflected the notion that there was local radial association in the true scene, bearing in mind the knowledge that the image was indeed that of a noisy ellipse.

Mean-field-like approximations have also been

used in a somewhat different approach to the maximisation of a complicated likelihood such as $p(y|\psi)$, exploiting the fact that

$$\log p(y|\psi) = \log\left\{\sum_x p(x, y|\psi)\right\} \quad (6)$$

$$\geq \sum_x q(x) \log\{p(x, y|\psi)/q(x)\}, \quad (7)$$

by Jensen’s inequality, where $q(x)$ is any probability distribution for x . In practice q is chosen to have a form that facilitates computation, with a fully-factorised independence model being the simplest option, and ‘hyperparameters’ within that form are chosen so as to maximise the lower bound to the log-likelihood given in (7). For details of this approach see Jordan *et al.*²⁰ Note that, so far as choice of q or its hyperparameters is concerned, maximisation of the lower bound is equivalent to minimisation of the Kullback-Leibler directed divergence between q and the ‘target’, $p(x|y, \psi)$, defined by

$$KL(q, p) := \sum_x q(x) \log\{q(x)/p(x|y, \psi)\}.$$

For a fully Bayesian analysis, (hyper)priors must be imposed on $\psi = (\theta, \beta)$, and inference about θ , β and x should be made on the basis of $p(x, \theta, \beta|y)$, and the associated marginal distributions. Needless to say, in most image-analysis contexts, and certainly in the familiar pixel-based models, there is no practically useful closed form for $p(x, \theta, \beta|y)$:

$$p(x, \theta, \beta|y) \propto p(y|x, \theta)p(x|\beta)p(\theta)p(\beta),$$

where we are assuming that θ and β are independent, a priori, with prior densities $p(\theta)$ and $p(\beta)$.

What has become the standard statistical approach is to use Markov chain Monte Carlo methods to generate a set of realisations from the above joint distribution and to make inferences about the unknown quantities, both parameters (ψ) and missing values (x), on the basis of empirical summaries of the simulated quantities. However, in the case of hidden Markov random fields, the intractable partition function within $p(x|\beta)$ once more causes problems, in that the first step in the simulation cycle is not straightforward. As a result, approximate methods have been tried. One such approach, mentioned by Heikkinen and Högmänder²¹ and investigated in some detail by Rydén and Titterton²², is to replace $p(x|\beta)$ by the pseudo-likelihood function when generating the next value of β . Rydén and

Titterington²² comment that the ‘Gibbs’ sampling scheme that results does converge, but that it is not clear how to characterise the limiting distribution. Rydén and Titterington also report some simulation experiments involving realisations from the Ising model, corrupted by Gaussian noise. The parameters of the noise model are estimated quite well by the resulting marginal means of the simulated sample from the posterior distribution, but there can be small but perceptible biases in the corresponding estimates of the Ising parameter, β . On the other hand, their attempts at alternative ways of dealing with the partition function, in the spirit of Geyer and Thompson¹⁵, were distinctly unsuccessful because of computational difficulties.

As in the likelihood approach, there is a technique involving deterministic variational approximations for use in the fully-Bayesian context. In this case an approximation $q(x, \theta, \beta)$ to $p(x, \theta, \beta|y)$ is sought to minimise $KL(q, p)$ subject to q having some special structure that simplifies the analysis. Generally, q is taken to have the factorised form $q(x, \psi) = q_x(x)q_\psi(\psi)$. In many specific implementations $q_\psi(\psi)$ then takes the same parametric form as obtains in the case in which the true x is given. Since the correct, if inaccessible, $p(\psi|y)$ certainly does not take the same form, these variational approximations inevitably lead to error, but in some cases it is at least possible to show that the modes of the correct and approximate distributions are asymptotically the same; see for example Wang and Titterington²³. For more review and references on variational Bayesian approximations see Jordan²⁴ and Titterington²⁵.

Before leaving this section about the Bayesian approach, it is appropriate to return to the models discussed in Section 2 and to note that the regularised estimator has an obvious Bayesian interpretation for grey-level images. If the noise model is given by (1), with $\epsilon \sim N(0, \sigma^2 I)$, and if the prior/marginal distribution for x is that of $N(0, \sigma^2 \beta^{-1} C^{-1})$, then the negative of the logarithm of $p(x|y)$ is, apart from additive and multiplicative constants, given by

$$\|y - Hx\|^2 + \beta x^\top C x,$$

so that the mode is given by \hat{x}_β as defined in (3). This interpretation then stimulates other ways of

choosing the regularisation parameter β , such as maximum likelihood, in which β is chosen to maximise

$$p(y|\beta) = \int p(y|x)p(x|\beta)dx,$$

under the assumption that H and σ^2 are known from the specification of the observing instrument. The integration can be done explicitly and the resulting $p(y|\beta)$ can be maximised numerically.

4 Deconvolution

If the noise vector is omitted from equation (1) then we are left with the problem of solving the inverse problem

$$y = Hx, \quad (8)$$

which can be thought of as a discrete deconvolution problem. With pixellated images the discreteness is achieved automatically, but it might be imposed as a way of dealing with more general scenarios governed by the integral equation

$$y(t) = \int h(s, t)x(s)ds,$$

for t and s ranging over specified domains. This corresponds to deconvolution, especially if $h(s, t)$ is a function of $s - t$. For simplicity we shall concentrate on the discrete form of the problem, although ways of dealing with the integral-equation version are covered in many of the referenced papers. If H is square and nonsingular, then the formal solution is $x = H^{-1}y$, but this may be impracticable if the original problem is ill-posed, as discussed already. Furthermore, x is likely to have to satisfy nonnegativity constraints, a fact we have not yet recognised in this paper, and typically y and H will also consist of nonnegative elements.

This type of problem is of course very well researched, and here we concentrate on just a few approaches from the statistical literature. A key source is the discussion paper of Vardi and Lee²⁶. They note that, by a scaling argument, without loss of generality it can be assumed that y and x sum to 1, as do the columns of H . They derive the following iterative algorithm for obtaining a nonnegative solution for (8), starting from a positive-valued $x^{(0)}$ that satisfies the unit-sum constraint: for $m = 1, \dots$, and for

each i th element of x , obtain

$$x_i^{(m)} = x_i^{(m-1)} \sum_j (h_{ij} / \sum_k x_k^{(m-1)} h_{kj}) y_j. \quad (9)$$

Clearly, for all m , the elements of $x^{(m)}$ are non-negative and sum to 1. Then the algorithm converges to the probability measure x^* that maximises $\sum_i y_i \log z_i$, where $z_i = (\sum_j h_{ij} x_j)$. This is equivalent to minimising

$$\sum_i y_i \log(y_i/z_i) = KL(y, z).$$

When equation (8) has a nonnegative solution then the algorithm converges to such a solution. Otherwise, it converges to the closest approximation in the above KL sense.

To statisticians, the algorithm has the appealing interpretation as a limiting version of the EM algorithm. In the context of this example, the E-step and the M-step are as follows.

- **E-step:** for each i and j calculate

$$z_{ij}^{(m-1)} = \frac{x_i^{(m-1)} h_{ij}}{\sum_k x_k^{(m-1)} h_{kj}} y_j.$$

- **M-step:** for each i , calculate

$$x_i^{(m)} = \sum_j z_{ij}^{(m-1)}.$$

The combination of these two formulae clearly amounts to equation (9). Informally, the E-step ‘distributes’ each y_j over the individual pixel sites and the M-step accumulates all the contributions corresponding to pixel i . That this discrete form of the algorithm was an EM algorithm was noted by Titterton and Rossi²⁷, stimulated by the algorithm’s appearance as an ad hoc procedure in Di Gesu and Maccarone²⁸.

Vardi and Lee²⁶ list a number of disparate manifestations of the general structure, including emission tomography image reconstruction, in which x denotes pixelwise emission intensities, elements of y are event counts at a set of detectors, emissions are assumed to follow Poisson distributions and h_{ij} is the probability that a particle emitted from pixel j is picked up by detector i . Another image-based special case is that of motion de-blurring. Given the EM interpretation of the algorithm and the knowledge

that unregularised maximum likelihood estimates might be ill-conditioned, it is not surprising that modified versions have been developed that involve some sort of smoothing, especially in the context of emission tomography. Such modifications include the smoothed EM algorithm of Silverman *et al.*²⁹, in which the $\{x_i^{(m)}\}$ obtained in the M-step are locally smoothed before being fed into the next E-step, and the modified EM algorithm of Green³⁰, in which a roughness penalty on the $\{x_i\}$ is included when the M-step is carried out. Hudson and Larkin³¹ provide another variation of EM, applied to tomography. The papers by Green³⁰ and Hudson and Larkin³¹ both won IEEE awards for their high levels of citation.

A different algorithm for the same purpose is the so-called Iterative Image Space Restoration Algorithm, for which the iteration is

$$x_i^{(m)} = x_i^{(m-1)} (\sum_j h_{ij} y_j) / \{ \sum_j h_{ij} (\sum_k x_k^{(m-1)} h_{kj}) \},$$

for each i . This algorithm was introduced by Daube-Witherspoon and Muehllehner³² and convergence properties were investigated by De Pierro³³ and Titterton³⁴, the latter of whom noted that the algorithm could be interpreted as an iterative approach to the calculation of least squares estimates of x . Further references and illustrations in the context of motion-blur, together with extensions to incorporate roughness penalties, thereby obtaining minimisers of (2), are available in Archer and Titterton³⁵. Key references from the non-statistical literature include Byrne³⁶ and Eggermont³⁷.

Vardi and Lee²⁶ present a number of illustrations, one of which concerns a motion-blurred moving toy cart. Part of the image was also treated by Archer and Titterton³⁵. They implemented both the EM and ISRA algorithms, running each of them for totals of 40 and 106 iterations. The restoration obtained after 40 iterations was arguably better defined, which suggests that the underlying inverse problem is somewhat ill-posed; stopping the algorithm early is one way of avoiding an ill-posed solution.

Although much of the research discussed in this section is somewhat dated, statistical research into

deconvolution, with applications relevant to this Conference, is certainly continuing. For example, Hall and Yin³⁸ consider a model, for a signal y observed at n time-points, given by

$$y_i = g(t_i) + \epsilon_i = \mu + \sum_{j=1}^r g_j(t_i) + \epsilon_i,$$

for $i = 1, \dots, n$, where the g_j are periodic components with minimal periods $0 < \theta_1 < \dots < \theta_r$. The objective is to estimate the unknown periods $\theta = \{\theta_j\}$ and the unknown functions $\{g_j\}$, without imposing simple parametric forms on the latter. To estimate the $\{\theta_j\}$, Hall and Yin use the minimiser $\hat{\theta}$ of the residual sum of squares function

$$S(\theta) = \sum_i \{y_i - \hat{g}(t_i|\theta)\}^2,$$

in which $\hat{g}(t|\theta)$ is a (preliminary) nonparametric estimator of $g(t)$. In particular, Hall and Yin use

$$\hat{g}(t|\theta) = \left\{ \sum_i y_i K(t, t_i) \right\} / \sum_i K(t, t_i),$$

in which $K(t, t')$ is a kernel function, defined as a function of θ : the kernel function is defined in such a way that $\hat{g}(t|\theta)$ is a weighted average of the $\{y_i\}$, with weights that are monotonic decreasing functions of $|t - t_i|$. Given the $\{\hat{\theta}_j\}$, the $\{g_j\}$ are estimated using orthogonal series methods. The overall response function is written as

$$g(t) = \mu + \sum_{j=1}^r \sum_{k=1}^m a_{jk} \psi_k(t/\hat{\theta}_j),$$

in which the $\{a_{jk}\}$ are generalised Fourier coefficients, μ is a constant, the $\{\psi_k\}$ are orthonormal functions and m is a truncation point. Least squares is again used, this time to estimate μ and the $\{a_{jk}\}$. Hall and Yin³⁸ discuss ways of choosing the $\{\hat{\theta}_j\}$, m and smoothing parameters within the kernel function, they investigate theoretical properties, and they fit the model to radiation measurements from the slowly-pulsating B-star HD 123515, showing that a multiperiodic function with $r = 4$ periods gives a good fit.

Acknowledgments

Appendix

References

1. I.L. Dryden and K.V. Mardia, *Statistical Shape Analysis*. (Wiley, 1998).
2. D.J.C. MacKay, *Network: Computation in Neural Systems* **6**, 469 (1995).
3. G.H. Golub *et al.*, *Technometrics* **21**, 215 (1979).
4. G. Wahba, *J. R. Statist. Soc. B* **45**, 133 (1983).
5. P. Hall and D.M. Titterington, *J. R. Statist. Soc. B* **49**, 184 (1987).
6. P. Hall and D.M. Titterington, *J. R. Statist. Soc. B* **48**, 330 (1986).
7. A.M. Thompson *et al.*, *IEEE Trans. Pattern Anal. Machine Intell.* **13**, 326 (1991).
8. S. Geman and D. Geman, *IEEE Trans. Pattern Anal. Machine Intell.* **6**, 721 (1984).
9. J. Besag, *J. R. Statist. Soc. B* **48**, 259 (1986).
10. U. Grenander *et al.*, *HANDS: a Pattern Theoretic Study of Biological Shapes* (Springer, 1990).
11. U. Grenander and M.I. Miller, *J. R. Statist. Soc. B* **56**, 549 (1994).
12. A.P. Dempster *et al.*, *J. R. Statist. Soc. B* **39**, 1 (1977).
13. J. Zhang, J., *IEEE Trans. Signal Proces.* **40**, 2570 (1992).
14. J. Zhang, J., *IEEE Trans. Image Proces.* **2**, 27 (1993).
15. C.J. Geyer and E.A. Thompson, *J. R. Statist. Soc. B* **54**, 657 (1992).
16. J. Besag, *Statistician* **24**, 179 (1975).
17. W. Qian, and D.M. Titterington, *Phil. Trans. R. Soc. Lond. A* **337**, 407 (1991).
18. W. Qian and D.M. Titterington, *IEEE Trans. Pattern Anal. Machine Intell.* **15**, 748 (1993).
19. W. Qian *et al.*, *J. Am. Statist. Assoc.* **91**, 944 (1996).
20. M.I. Jordan *et al.*, in *Learning in Graphical Models* ed. M.I. Jordan, 105 (MIT Press, Cambridge, MA, 1999).
21. J. Heikkinen and H. Högmänder, *Appl. Statist.* **43**, 569 (1994).
22. T. Rydén and D.M. Titterington, *J. Comp. Graph. Statist.* **7**, 194 (1998).
23. B. Wang and D.M. Titterington, *Bayesian*

- Anal.*, to appear (2006).
24. M.I. Jordan, *Statist. Sci.* **19**, 140 (2004).
 25. D.M. Titterington, *Statist. Sci.* **19**, 128 (2004).
 26. Y. Vardi and D. Lee, *J. R. Statist. Soc. B* **55**, (1993).
 27. D.M. Titterington and C. Rossi, *Signal Proces.* **9**, 101 (1985).
 28. V. Di Gesu and M.C. Maccarone, *Signal Proces.* **6**, 201 (1984).
 29. B.W. Silverman *et al.*, *J. R. Statist. Soc. B* **52**, 271 (1990).
 30. P.J. Green, *IEEE Trans. Med. Imaging* **9**, 84 (1990).
 31. H.M. Hudson and R.S. Larkin, *IEEE Trans. Med. Imaging* **13**, 601 (1994).
 32. M.E. Daube-Witherspoon and G. Muehllehner, *G. IEEE Trans. Med. Imaging* **5**, 61 (1986).
 33. A.R. De Pierro, *IEEE Trans. Med. Imaging* **6**, 174 (1987).
 34. D.M. Titterington, *IEEE Trans. Med. Imaging* **6**, 52 (1987).
 35. G.E.B. Archer and D.M. Titterington, *Statist. Sinica* **5**, 77 (1995).
 36. C.L. Byrne, *IEEE Trans. Image Proces.* **2**, 96 (1993).
 37. P.P.B. Eggermont, *Lin. Algeb. Applics.* **130**, 25 (1990).
 38. P. Hall and J. Yin, *J. R. Statist. Soc. B* **65**, 869 (2003).