An EM algorithm for Gaussian Markov Random Fields

Will Penny,
Wellcome Department of Imaging Neuroscience,
University College, London WC1N 3BG.
wpenny@fil.ion.ucl.ac.uk

October 28, 2002

Abstract

Lavine [7] has shown how 1-dimensional Linear Dynamical Systems (LDS's) can be used for exact inference in 2- (or higher) dimensional Gaussian Markov Random Fields (GMRFs). His trick is to relate the row of an image to the state of the LDS and introduce pseudo-observations (which turn out to be zero) such that the evolution equation implements vertical neighbour contstraints and the observation equation implements horizontal neighbour constraints. Thus, exact inference can take place using Kalman smoothing. In this report we show how nonstationary smoothness parameters can be estimated using the M-step of the EM algorithm for LDS's.

1 Introduction

We consider the problem of 'restoring' 2D images that have been corrupted by additive Gaussian noise

$$v_{ij} = h_{ij} + e_{ij} \tag{1}$$

where v_{ij} are the observed image values, h_{ij} are the true uncorrupted image values which we wish to estimate and e_{ij} is additive zero mean Gaussian noise. The images are of dimension $i = 1..N_y$ by $j = 1..N_x$ and have a total

of $N = N_x N_y$ pixels. We use the column vectors V_i and H_i to denote the ith observed and hidden rows of the images. Whole images are denoted by matrices V and H.

In this paper the noise is considered to be independent from voxel to voxel. The noise variance is considered to be either stationary or nonstationary, the latter case being of most interest.

The above problem is known in image processing as image restoration. It is solved by assuming that the images are smooth, in some as yet undefined sense, so that information from local neighbourhoods can be used to estimate h_{ij}. In what follows $\mathsf{N}(m, \Sigma)$ denotes a univariate/multivariate Gaussian with mean m and variance/covariance Σ.

2 Gaussian Markov Random Fields

We consider the Conditional Autoregressive (CAR) model introduced by Besag [1]. An alternative choice would have been the Simultaneious Autoregressive model described in [8] (page 88) and [6],[3].

CARs are defined as follows. Firstly, for each location i, j we define a set of neighbouring locations. The probability density for h_{ij} is then specified in terms of these neighboring values. This set of conditional distributions is then sufficient to specify the joint distribution $p(H)$.

The fact that the probability densities are initially specified by conditional densities based on neighbouring values is the Markov property. For this reason CARs are also known as a Gaussian Markov Random Fields (GMRFs) [2].

In this paper we consider a neighbourhood system based on cardinal points ie. $\mathcal{N}_{ij} = \{h_{i-1,j}, h_{i+1,j}, h_{i,j-1}, h_{i,j+1}\}$. The conditional densities are

$$p(h_{ij}|\mathcal{N}_{ij}) = \mathsf{N}\left(\mu_{ij}, \frac{1}{(r_j + s_i)}\right) \tag{2}$$

where $r_j = 1/\alpha_j$ and $s_i = 1/\beta_i$. The mean is given by

$$\mu_{ij} = \frac{1}{2(r_j + s_i)}[r_j(h_{i-1,j} + h_{i+1,j})) + s_i(h_{i,j-1} + h_{i,j+1}))] \tag{3}$$

The terms α_j and β_i specify prior variances for data in the jth column and ith row of the image. These are also written as column vectors α and β. Together, these densities specify the joint prior $p(H)$.

The GMRF prior for $p(H)$ is not a proper prior, however, because it specifies only differences between neighbouring values rather than absolute values. We do get a multivariate Gaussian for $p(H)$ but all the rows and columns of the precision matrix (eq 4 of [7]) add up to zero. It is therefore singular.

The observed values are related to the unobserved values through

$$p(v_{ij}|h_{ij}) = \mathsf{N}(h_{ij}, \lambda_{ij}) \tag{4}$$

where λ_{ij} is the variance of the ijth observation. We also write Λ_i as the vector of variances for the ith row.

From the above specification of the prior and the likelihood it is possible to compute the posterior using Bayesian inference in the usual way. A computational problem appears however, as to compute the posterior one must invert a very large (N by N) matrix. One solution is to take advantage of the block tridiagonal structure in this matrix using, for example, Block Cyclic Reduction (BCR) (see [5], page 177). But in this paper we use a (fabulous !) trick introduced by Lavine [7] who shows that inference in 2D GMRFs can take place using the machinery of one dimensional Linear Dynamical Systems (LDS). Whilst the underlying algorithm is not necessarily faster than BCR it does, however, provide access to the already established theory of Bayesian estimation and inference for LDS.

The standard LDS framework is described in the appendix, along with an associated EM algorithm. GMRFs can be mapped onto LDS's by associating rows in images with state vectors at different time points. Specifically, we write

$$\begin{aligned} x_t &= H_i \\ A &= I_J \end{aligned} \tag{5}$$

We then form the augmented observation vector

$$y_t = \begin{bmatrix} V_i \\ 0_{J-1} \end{bmatrix} \tag{6}$$

where 0_{J-1} is a $J-1$-element column vector of zeros. We also form an augmented observation matrix

$$C = \begin{bmatrix} I_J \\ F \end{bmatrix} \tag{7}$$

where F is the $(J-1)$-by-J matrix

$$F = \begin{bmatrix} 1 & -1 & 0 & .. & 0 \\ 0 & 1 & -1 & .. & 0 \\ .. & .. & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{8}$$

The state noise covariance matrix is then set to

$$Q_t = \mathsf{diag}(\alpha) \tag{9}$$

where 1_J is a row of 1's. The observation noise covariance matrix is set to

$$R_t = \begin{bmatrix} \mathsf{diag}(\Lambda_i) & 0 \\ 0 & \mathsf{diag}(\beta) \end{bmatrix} \tag{10}$$

We note that non-rectangular grids can be accomodated by making A and F row-specific, ie. use A_i and F_i. Zero-valued rows in the matrix A_i would indicate the lack of a vertical dependence and zero-valued rows in F_i the lack of a horizontal dependence. This allows one to smooth discontiguous fields.

We also use the initial values $\mu_1 = V_i$ and $\Sigma_1 = I_J$. Using these values it is now possible to run the Kalman smoothing algorithm shown in the Appendix. The constitutes the E-Step of an EM algorithm for inference in LDS's described by Ghahramani [4]. It is then possible for us to re-estimate α and β using the M-Step updates for Q and R also given in the appendix. From equation 33 it can be shown that

$$\alpha_j = \frac{1}{T-1}\left[\left(\sum_{t=2}^{T} P_t(j,j)\right) - 2\left(\sum_{t=2}^{T} P_{t,t-1}(j,j)\right) + \left(\sum_{t=2}^{T} P_{t-1}(j,j)\right)\right] \tag{11}$$

This is equivalent to

$$\alpha_j = \frac{1}{I-1}\sum_{i=2}^{I} E[(h_{ij} - h_{i-1,j})^2] \tag{12}$$

where $E[]$ is the expected value under the posterior distribution. This makes intuitive sense. For homogeneous priors we have

$$\alpha = \frac{\sum_j \alpha_j}{J} \tag{13}$$

From equation 35 and our definitions of y_t and C it can be shown that

$$\beta_i = \frac{1}{T}\sum_{t=1}^{T}(FP_tF^T)(i,i) \tag{14}$$

This is equivalent to

$$\beta_i = \frac{1}{J-1}\sum_{j=1}^{J-1} E[(h_{ij} - h_{i,j+1})^2] \tag{15}$$

For homogeneous priors we have

$$\beta = \frac{\sum_{i=1}^{I}\beta_i}{I} \tag{16}$$

3 Discussion

The figure shows the restoration of an artificial image corrupted with Gaussian noise, the nonstationary GMRF algorithm being superior.

The positive features of the algorithm are that images can be nonstationarily smoothed. Importantly, the smoothing can take into account different noise variance values at each voxel. Further, the full posterior distribution over voxel values is readily computed (it falls out of the Kalman smoothing step).

A subtlety with the EM estimation of the smoothness (variance) parameters is that the β values cannot be computed directly. This is because the pseudo-observations, being zero, will result in β estimates that are zero. Therefore to estimate them, we use a two-stage process where in the first stage the β's are fixed and the α's are estimated. We then re-run the algorithm with transposed images. These two stages can be interleaved so that we do not get excessive smoothing in one (or the other) direction.

A Linear Dynamical System

Our model is a linear dynamical system for T p-variate observations and a latent-space of dimension k. The state-space equations are

$$\begin{aligned} x_t &= Ax_{t-1} + w_t \\ y_t &= Cx_t + v_t \end{aligned} \tag{17}$$

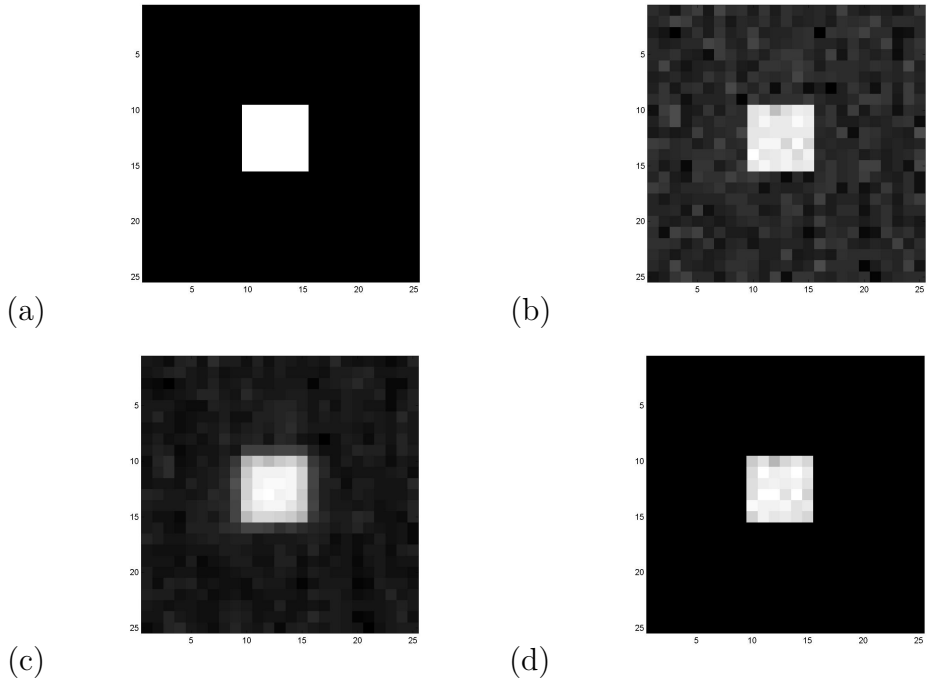

Figure 1: *(a) Original image, (b) Corrupted image, Error=2.92, (c) Image restored with GMRF stationary prior, Error=2.66, (d) GMRF nonstationary prior, Error=0.26.*

where $p(w_t) = \mathsf{N}(0, Q)$, $p(v_t) = \mathsf{N}(0, R_t)$ and Q and R_t are covariance matrices. The variables x_t are state variables and y_t the observations. They are vectors of dimension $k \times 1$ and $p \times 1$ respectively. We have

$$p(y_t|x_t) = \mathsf{N}(Cx_t, R_t) \tag{18}$$
$$p(x_t|x_{t-1}) = \mathsf{N}(Ax_{t-1}, Q) \tag{19}$$
$$p(x_1) = \mathsf{N}(\mu_1, \Sigma_1) \tag{20}$$

which define the observation model, state transition model and initial state distribution. If we know A, C, Q and R then the hidden state variables can be inferred using Kalman smoothing.

By the Markov property

$$p(X, Y) = p(x_1) \prod_{t=1}^{T} p(x_t|x_{t-1}) \prod_{t=1}^{T} p(y_t|x_t) \tag{21}$$

Therefore the joint log-likelihood is a sum of quadratic terms

$$L = \log p(X,Y) = -\sum_{t=1}^{T} \frac{1}{2}\left[(y_t - Cx_t)' R_t^{-1}(y_t - Cx_t) - \log|R_t|\right] \quad (22)$$

$$- \sum_{t=2}^{T} \frac{1}{2}\left[(x_t - Ax_{t-1})' Q^{-1}(x_t - Ax_{t-1})\right] - \frac{T-1}{2}\log|Q|$$

$$- \frac{1}{2}\left[(x_1 - \mu_1)'\Sigma^{-1}(x_1 - \mu_1)\right] - \frac{1}{2}\log|\Sigma_1| - \frac{T(p+k)}{2}\log 2\pi \quad (23)$$

A.1 EM algorithm

The EM algorithm requires us to maximise the auxiliary function

$$F = \int p(X|Y)\log p(X,Y)dX \quad (24)$$

This quantity depends on three expectations which we denote as follows

$$\begin{aligned} m_t &= \int p(x_t|Y)x_t dx_t \quad (25)\\ P_t &= \int p(x_t|Y)x_t x_t^{'} dx_t \\ P_{t,t-1} &= \int p(x_t, x_{t-1}|Y)x_t x_{t-1}^{'} dx_t \end{aligned}$$

A.2 E-Step: Kalman smoothing

Following [4], we write the expected value of x_t conditioned on all data up to time t as $x_t^t \equiv E[x_t|y_1^t]$. Similarly, the corresponding covariance is given by $\Sigma_t^t \equiv \mathrm{Var}[x_t|y_1^t]$.

This step implements the recursive computation of x_t^t and Σ_t^t from x_{t-1}^{t-1} and Σ_{t-1}^{t-1}.

$$\begin{aligned} x_t^{t-1} &= Ax_{t-1}^{t-1} \quad (26)\\ \Sigma_t^{t-1} &= A\Sigma_{t-1}^{t-1}A^{'} + Q \\ K_t &= \Sigma_t^{t-1}C'\left(C\Sigma_t^{t-1}C' + R_t\right) \\ x_t^t &= x_t^{t-1} + K_t(y_t - Cx_t^{t-1}) \\ \Sigma_t^t &= \Sigma_t^{t-1} - K_t C\Sigma_t^{t-1}. \end{aligned}$$

The procedure is initialised using $x_1^0 = \mu_1$ and $\Sigma_1^0 = \Sigma_1$.

The backward recursions compute x_t^t and Σ_t^t from x_{t-1}^{t-1} and Σ_{t-1}^{t-1}.

$$\begin{aligned} J_{t-1} &= \Sigma_{t-1}^{t-1} A^T (\Sigma_t^{t-1})^{-1} \\ x_{t-1}^T &= x_{t-1}^{t-1} + J_{t-1}(x_t^{'} - A x_{t-1}^{t-1}) \\ \Sigma_{t-1}^T &= \Sigma_{t-1}^{t-1} + J_{t-1}(\Sigma_t^{'} - \Sigma_t^{t-1}) {J_{t-1}}^{'}. \end{aligned} \tag{27}$$

The procedure is initialised using $\Sigma_T^T = \Sigma_T$ and $x_T^T = x_T$ where the right hand side quantities are from the final forward recursion step.

A.3 Posterior Density

The forward and backward steps together allow us to compute x_t^T and Σ_t^T which are the first two moments of x_t conditioned on the *whole* data set. The posterior density is therefore given by $p(x_t|Y) = \mathsf{N}(m_t, V_t)$ where

$$\begin{aligned} m_t &\equiv x_t^T \\ V_t &\equiv \Sigma_t^T. \end{aligned} \tag{28}$$

For the M-step we also need the quantities

$$P_t \equiv \Sigma_t^T + x_t^T (x_t^T)^{'} \tag{29}$$

and $P_{t,t-1} \equiv \Sigma_{t,t-1}^T + x_t^T (x_{t-1}^T)^{'}$ where backward recursions are used for

$$\Sigma_{t-1,t-2}^T = \Sigma_{t-1}^{t-1} {J_{t-2}}^{'} + J_{t-1}(\Sigma_{t,t-1}^T - A\Sigma_{t-1}^{t-1}) {J_{t-2}}^{'} \tag{30}$$

The last recursion is initialised using

$$\Sigma_{T,T-1}^T = (I - K_T C) A \Sigma_{T-1}^{T-1}. \tag{31}$$

A.4 M-Step

The update for a full-covariance Q is (assuming $A = I$), from [4]

$$Q = \frac{1}{T-1} \left[\left(\sum_{t=2}^T P_t \right) - \left(\sum_{t=2}^T P_{t,t-1} \right) - \left(\sum_{t=2}^T P_{t-1,t} \right) + \left(\sum_{t=2}^T P_{t-1} \right) \right] \tag{32}$$

For a diagonal Q we have

$$Q_D = \mathsf{diag}(Q) \tag{33}$$

and for isotropic Q

$$Q_I = \frac{1}{k}\mathsf{Tr}(Q) \tag{34}$$

The update for R is, from [4]

$$R = \frac{1}{T}\left(\sum_{t=1}^{T} y_t y_t^{'} - 2Cm_t y_t^{'} + CP_t C^{'}\right) \tag{35}$$

References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[2] J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, (82):733–746, 1995.

[3] N. Cressie. *Statistics for spatial data*. John Wiley, 1991.

[4] Z. Ghahramani and G.E. Hinton. Parameter Estimation for Linear Dynamical Systems. Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto, 1996. Also available from http://www.gatsby.ucl.ac.uk/ zoubin/papers.html.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, 3rd edition, 1996.

[6] J.P. LePage. Lecture notes in spatial econometrics. Technical report, Department of Economics, University of Toledo, 2002. Available from http://www.econ.utoledo.edu/faculty/lepage/lepage.html.

[7] M.Lavine. Another look at conditionally Gaussian Markov random fields. In J.M. Bernardo, J.O. Berger, A. P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 371–387, 1999.

[8] B.D. Ripley. *Spatial statistics*. John Wiley, 1981.