JRSSB, 1990

Parametric models that allow π_{ij} to vary among constituencies may incorporate

(a) fixed effects, via logistic dependence on constituency level variables such as partisanship or population density (Brown and Payne, 1986) and
(b) random effects via beta-binomial distributions for X_{i1}, X_{i2}, which generalize easily to Dirichlet-multinomial distributions when there are more than two parties and abstentions are allowed (Firth, 1982; Brown and Payne, 1986).

An attractive alternative is to smooth the $\{\pi_{ij}\}$ nonparametrically, using either geographical location or other constituency level variables to define neighbours. However, the overparameterization noted earlier makes the EMS method unworkable in this case; a more promising approach might be to 'roughen' the π_{ij} from a suitable, fixed-effects-only, parametric model. The method of local likelihood (Tibshirani and Hastie, 1987; Firth *et al.*, 1989), used in place of the M step, would be a suitable way to achieve this; no separate S step is involved.

Note that this problem differs slightly from those considered in the paper in that it is one of prediction, rather than estimation. In PET, for example, the parameters $\{\lambda_s\}$ are the objects of direct interest, whereas here the corresponding $\{\pi_{ij}\}$ are parameters of a superpopulation model constructed to allow inference about the unobserved counts $\{X_{i1}: i=1, \ldots, n\}$.

I thank the authors for rekindling my interest in this topic.

Andrew Gelman (Harvard University, Cambridge): Section 5 of this stimulating paper points out that the EMS algorithm corresponds to EM_p with a quadratic penalty on the vector $\boldsymbol{\theta}$. This penalty can be shown to correspond to a conditional autoregressive (CAR) normal prior density on $\theta \propto \surd\lambda$. We shall modify the problem by analysing an EMS algorithm with smoothing applied to $\boldsymbol{\theta}$, which, as discussed in Section 5.3, should differ little from smoothing $\boldsymbol{\lambda}$. (If anything, smoothing $\boldsymbol{\theta}$ should be preferred, because the square-root transformation stabilizes the Poisson variances.) In the Bayesian interpretation, EM_p leads to the posterior mode of $\boldsymbol{\theta}$, given data n, and the M_p step must yield the conditional posterior mode of $\boldsymbol{\theta}$, given expected sufficient statistics $\hat{m}$. We shall here answer the following question: given a CAR model, what smoothing matrix $\mathbf{S}$ leads the EMS algorithm to a posterior mode?

Label $\hat{\mu}_s = (\hat{m}_s/a_s q_s)^{1/2}$; the conditional distribution $p(\hat{m}|\lambda)$ is of Poisson form and can be approximated by a normal distribution on the square-root scale:

$$\hat{\mu}_s \surd(a_s q_s) \sim N(\theta_s \surd(a_s q_s), \tfrac{1}{4}) \tag{2}$$
$$\hat{\mu} \sim N(\theta, \Sigma).$$

The variance matrix is $\Sigma = \mathrm{diag}(\sigma_s^2) = \mathrm{diag}(1/4a_s q_s)$.

A CAR model is defined by its matrix of coefficients $\mathbf{C} = (c_{rs})$ and its matrix of conditional variances $\mathbf{T} = \mathrm{diag}(\tau_r^2)$:

$$E(\theta_r|\theta_s, s \neq r) = \sum_{s \neq r} c_{rs}\theta_s, \qquad c_{rr} = 0 \text{ for all } r,$$
$$\mathrm{var}(\theta_r|\theta_s, s \neq r) = \tau_r^2.$$

We require $\Sigma_s c_{rs} = 1$ for all r to force the prior distribution to be location invariant, which creates an intrinsic autoregression of order zero (Künsch, 1987). The precision of this improper multivariate normal distribution is $(\mathbf{I} - \mathbf{C}^T)\mathbf{T}^{-1}$.

The M_p step maximizes the resulting posterior density on $\boldsymbol{\theta}$; up to the accuracy of approximation (2), this is just a normal density with mode

$$E(\theta|m) = \Sigma^{-1}(\Sigma^{-1} + (\mathbf{I} - \mathbf{C}^T)\mathbf{T}^{-1})^{-1}\hat{\mu},$$

which will be obtained by the S step, operating on θ, if

$$\mathbf{S} = \Sigma^{-1}(\Sigma^{-1} + (\mathbf{I} - \mathbf{C}^T)\mathbf{T}^{-1})^{-1},$$

implying:

$$\mathbf{S}^{-1} = \mathbf{I} + (\mathbf{I} - \mathbf{C}^T)\Sigma\mathbf{T}^{-1} \tag{3}$$
$$= (\mathbf{I} - \mathbf{C}')(\mathbf{I} + \Sigma\mathbf{T}^{-1}),$$

where $\mathbf{C}' = (c'_{rs})$ is defined by

$$c'_{st} = \frac{\sigma_s^2}{\sigma_s^2 + \tau_s^2} c_{rs}.$$

Thus, the appropriate smoother $\mathbf{S}$ is just a normalized form of the covariance matrix of a CAR model with smaller autoregression coefficients than those of the prior model. Conversely, the model corresponding to a given smoother $\mathbf{S}$ can be found by solving equation (3) for $\mathbf{C}$ and $\mathbf{T}$, with the constraint that the diagonal elements of $\mathbf{C}$ must be zero.

To see how a prior model leads to the particular smoother suggested in the paper, assume that the variances σ_s^2 are equal, and consider the CAR model on a square lattice with $c_{rs} = \frac{1}{8}$ for all first- and second-order neighbours and equal conditional variances τ_s^2. For τ^2 comparable with or larger than σ^2, equation (3) can be approximately inverted:

$$\mathbf{S} \approx \frac{\tau^2 + \sigma^2/2}{\tau^2 + 3\sigma^2/2}\mathbf{I} + \frac{\sigma^2}{\tau^2 + 3\sigma^2/2}\mathbf{C},$$

corresponding to the linear smoother of Section 4.3.2 with weights $W = 8(\tau^2/\sigma^2 + \frac{1}{2})$. Setting $a_s = 1$, $q_s = 1$ and $\tau_s^2 = 1$ yields $W = 36$.

Professor H. M. Hudson (Macquarie University, North Ryde): Even with the benefits of the excellent ring grid pixel definition introduced here, PET reconstruction retains the characteristics of a highly overparameterized system. The question of which of the many reconstructions maximizing likelihood is obtained has been recognized. By penalizing a solution in a way that encourages smoothness, or choosing a smoother alternative solution that retains similar projections, we gravitate towards a specific reconstruction that better accords with our desire for smoothness.

In a perfect imaging system where the emission counts $\{m_s\}$ from all pixels s were available the traditional variance bias compromise would indicate the benefit of some smoothing to form an appropriate density estimate whenever Poisson count variability is not negligible. Does the EMS approach provide a more effective reconstruction than a smoothed maximum likelihood reconstruction? Have the authors contrasted the adequacy of the final reconstructions μ and λ in the EMS method?

There remains a numerical concern distinct from these statistical considerations. The EMS approach can share the slow convergence of the EM algorithm. If Poisson variability is negligible, as with coarser discretizations and large tube counts, we would expect no benefit from EMS procedures unless smoothing is very slight, when EM and EMS algorithms provide very similar iterations. It is our experience that in these conditions the EM and EMS algorithms remain unacceptably slow in their convergence.

Alternatives to the EM algorithm that improve the slow convergence have been introduced by Barnett *et al.* (1989). The approach suggested there is based on a modified Fisher scoring (MFS) algorithm, which, in its simplest form, is no more computationally demanding than the EM algorithm, but which improves markedly the rate of convergence towards a maximum likelihood solution. Recent studies by our doctoral student, K. Notodiputro, have demonstrated important reductions in iterations to convergence. The iteratively reweighted least squares approach of McCullagh and Nelder (1989) for generalized linear models is modified by approximations of the Fisher information matrix appropriate for projection data. The MFS approach can also be used to improve convergence of some penalized likelihood procedures.

Iain M. Johnstone (Stanford University): I would like to mention some possibly related literature that came to my attention while reading this thought-provoking paper. A Gaussian analog of the authors' Poisson model (2.1) might involve unobservable complete data

$$(y_{st}) \overset{\text{ind}}{=} N(x_{st}\beta_s, n^{-1}\sigma^2)$$

and observed data $y_t = \Sigma_s y_{st}$. The EM updating equations then have the form (setting $\gamma_t = \Sigma_s x_{st}\beta_s$)

$$\beta_s^{\text{new}} = \beta_s^{\text{old}} + (T\sum_t x_{st}^2)^{-1}\sum_t x_{st}(y_t - \gamma_t^{\text{old}}).$$