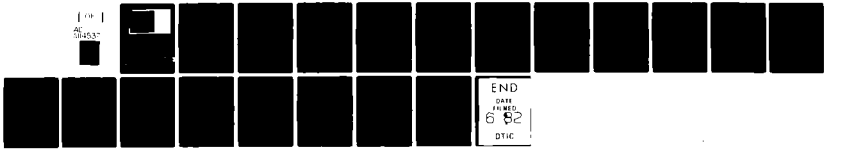
AD-A114 537 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
APPLICATIONS OF THE EM ALGORITHM TO THE ESTIMATION OF BAYESIAN --ETC(U)
MAR 82 T LEONARD DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2344 NL



(2)

AD A114537

MRC Technical Summary Report #2344

APPLICATIONS OF THE EM ALGORITHM TO THE ESTIMATION OF BAYESIAN HYPERPARAMETERS

Tom Leonard

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

March 1982

Received January 8, 1982

DTIC FILE COPY

Approved for public release
Distribution unlimited

DTIC
ELECTE
MAY 18 1982
S D
E

Sponsored by
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

82 05 18 027

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

APPLICATIONS OF THE EM ALGORITHM TO THE ESTIMATION OF BAYESIAN HYPERPARAMETERS

Tom Leonard

Technical Summary Report #2344
March 1982

ABSTRACT

Applications of the EM algorithm to the estimation of Bayesian hyperparameters are discussed and reviewed in the context of the author's philosophy involving the inductive and pragmatic modelling of sampling distributions and prior structures. Frequently the hyperparameters may be estimated from the data, thus avoiding the subjective assessment of these values. The ideas are applied to multiple regression models, histograms and multinomial distributions. A numerical example is described in the context of smoothing the cell probabilities of several multinomial distributions.

AMS(MOS) Subject Classifications: 62F15; 62H12

Key Words: Hyperparameter, EM algorithm, multiple regression, ridge regression, histogram smoothing, exchangeability, shrinkage estimators, multinomial distributions.

Work Unit No. 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

One of the drawbacks of Bayesian methods is that it is usually difficult to subjectively specify the hyperparameters. However in situations involving several parameters it is often possible to assess a sensible structure for the prior distribution and to estimate the hyperparameters empirically from the data. The EM algorithm provides a valuable mechanism for doing this. A general approach is described in the contexts of multiple regression models, histograms, and multinomial distributions and a numerical example from educational testing is used to illustrate the procedures.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

APPLICATIONS OF THE EM ALGORITHM TO THE ESTIMATION OF BAYESIAN HYPERPARAMETERS

Tom Leonard

1. Background and Statistical Ideas.

The resurgence in Bayesian statistics over the past fifteen or so years has been due to its recognition, as a methodology based upon pure probability theory and hence free from theoretical counterexample, and as an approach to scientific investigation which assists the deductive, inductive, and pragmatic reasoning of the statistician in a meaningful and illuminating way. In his writings Professor Bruno De Finetti has highlighted three fundamental concepts. These are Coherence, Scientific Induction, and Exchangeability. All these are related to the very necessary iteration between inference about parameters, conditional on the truth of the sampling model (coherence, deduction) and the process of scientific modelling (inductive and pragmatic reasoning together with more formal theoretical procedures). For accounts of scientific modelling in a Bayesian context see Box (1980) and Leonard (1978, 1981, 1982).

Many traditional Bayesians (e.g. Lindley et al., 1978) have pursued the concept of coherence on its own and have concentrated their energies on trying to extract coherent subjective distributions from the scientist. Less attention has been paid to the statistical problem of gaining insight from data sets in relation to their scientific background. Whilst prior distributions are very useful mathematical and philosophical constructs for generating meaningful statistical procedures, it is unlikely that all the useful knowledge possessed by a scientific expert will be representable in the form of a probability distribution. An expert's information is usually more complex and

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

diverse; it is usually necessary to extract this information via an interation between the statistician, the expert, and the statistical methodology, the data, and the scientific background. Pragmatic and inductive judgements seem more appropriate than trying to extend coherence for inference about parameters, given the model, to coherence of the whole statistical procedure. I see Bayesian statisticians as possessing tremendous advantages in the sense that, when treated pragmatically, their methodologies and philosophies can be used to advise the investigator how to handle his down-to-earth practical problems involving scientific data, in a very realistic and doubtlessly superior manner.

Suppose that the statistician is concerned with a $p \times 1$ vector $\underset{\sim}{\beta}$ of parameters of interest; let $\underset{\sim}{x}$ represent his observations. To be able to make a Bayesian inference about $\underset{\sim}{\beta}$ he is generally faced with three complicated practical problems. These are

(1) Modelling the form of the sampling distribution $p(\underset{\sim}{x}|\underset{\sim}{\beta},\underset{\sim}{\xi})$ of $\underset{\sim}{x}$ given $\underset{\sim}{\beta}$, and (perhaps) further parameters $\underset{\sim}{\xi}$.

(2) Modelling the <u>prior structure</u> i.e. the form ofthe prior distribution $\pi(\underset{\sim}{\beta}|\underset{\sim}{\lambda})$ of $\underset{\sim}{\beta}$ given some hyperparameters $\underset{\sim}{\lambda}$.

(3) Ascertaining suitable values for the hyperparameters $\underset{\sim}{\lambda}$.

Of these, (1) is particularly important when the elements of $\underset{\sim}{x}$ are continuous measurements; for categorical data Poisson or multinomial assumptions often suffice. The Bayesian non-parametric estimation of sampling densities is, for example, discussed by Leonard (1973 and 1978), and Atilgan and Leonard (1981). The Bayesian modelling of sampling distributions seems to me to provide one of the most important furture directions for our subject, considerable more work is needed in this area. This aspect will be discussed further in Section 4.

The modelling (2) of prior structures is equally important, and more difficult because the data are not so immediately relevant. It seems necessary to avoid the obvious retreat to conjugacy, except, perhaps, for the linear statistical model (conjugate priors only exist for a restrictive class of sampling models and in non-normal cases they possess quite restrictive covariance structures). In fact, De Finetti's concept of exchangeability provides us with one way of thinking about prior structures. We could

(a) Seek a suitable $p \times 1$ vector $\underset{\sim}{\gamma}$ of transformations of the elements of $\underset{\sim}{\beta}$ such that we are prepared to take the elements of $\underset{\sim}{\gamma}$ to be a priori exchangeable.

(b) Model a suitable exchangeable distribution for the elements of $\underset{\sim}{\gamma}$ e.g. by employing the two stage structure recommended by De Finetti's theorem; or a first stage prior plus empirical procedures at the second stage.

These ideas will be discussed, for particular examples, in later sections. Both (a) and (b) should in general be based upon such aspects as pragmatic judgement in relation to the scientific background, intuition about how substantively particular assumptions are likely to affect the posterior conclusions, and judgements about how reasonable the posterior estimates are in statistical terms (e.g. there is a conceptual duality between estimation procedures and the undelying prior assumptions). Both the exchangeable distribution and transformations may be taken to depend upon the hyperparameters $\underset{\sim}{\lambda}$.

The statistician is finally faced with problem (3) i.e. ascertaining appropriate values for the hyperparameter $\underset{\sim}{\lambda}$. When the dimension of $\underset{\sim}{\lambda}$ is small compared with $\underset{\sim}{\beta}$ the data will frequently possess considerable information about $\underset{\sim}{\lambda}$. The information about $\underset{\sim}{\xi}$ and $\underset{\sim}{\lambda}$ is summarized by their

<u>marginal likelihood</u>

$$p(\underset{\sim}{x}|\underset{\sim}{\xi},\underset{\sim}{\lambda}) = \int p(\underset{\sim}{x}|\underset{\sim}{\beta},\underset{\sim}{\xi})\pi(\underset{\sim}{\beta}|\underset{\sim}{\lambda})d\underset{\sim}{\beta} \; . \qquad (1.1)$$

We see that, rather than pursuing the unenviable task of extracting the subjective beliefs from a scientist, it is possible to use the marginal likelihood to simply estimate $\underset{\sim}{\lambda}$ from the data. A complicated way of doing this is to assign a hyperprior to $\underset{\sim}{\lambda}$, yielding a hierarchical (two-stage) prior for $\underset{\sim}{\beta}$ (see, for example, Lindley and Smith, (1972) and the obvious hyperposterior for $\underset{\sim}{\lambda}$ by multiplication with the marginal likelihood in (1.1). However, whilst uninformative hyperpriors can be useful (e.g. Leonard, 1976), it would be difficult for the applied worker to model the structure of informative hyperpriors. Also, formal Bayesian procedures for estimating $\underset{\sim}{\beta}$, unconditionally from $\underset{\sim}{\lambda}$, typically become very tedious computationally.

We will instead estimate $\underset{\sim}{\xi}$ and $\underset{\sim}{\lambda}$ by jointly maximizing the marginal likelihood in (1.1). The EM algorithm provides us with an excellent computational scheme for doing this in a very wide range of situations. Estimates for $\underset{\sim}{\beta}$ will be proposed which provide good approximations to formal hierarchical Bayes estimates whenever the marginal likelihood is moderately informative.

In situations where the objective of the analysis is to gain insight from a data-set, the idea of empirically estimating the hyperparameters, rather than specifying them a priori, seems appealing. I think that a coherent Bayesian would simply be attempting an impossible task if he tried to construct an entire multivariate distribution just based upon subjective information. However, if we think in terms of the specification of a meaningful prior structure, with any spare hyperparameters estimated empirically from the data, then the procedure will often make more statistical sense.

2. The EM Algorithm.

It is possible to apply the algorithm developed by Dempster et. al. to the estimation of ξ and $\underset{\sim}{\lambda}$ by the maximum of the marginal likelihood in (1.1). In the present context this involves the following two steps.

Expectation Step (E-Step). Using the latest vectors ξ^* and $\underset{\sim}{\lambda}^*$ for ξ and $\underset{\sim}{\lambda}$ calculate the expectation

$$\varepsilon(\xi,\underset{\sim}{\lambda}|\xi^*,\underset{\sim}{\lambda}^*) = E[\log p(\underset{\sim}{x}|\underset{\sim}{\beta}) + \log\pi(\underset{\sim}{\beta}|\underset{\sim}{\lambda})] \qquad (2.1)$$

where the expectation should be taken with respect to the posterior distribution of $\underset{\sim}{\beta}$, given $\xi = \xi^*$ and $\underset{\sim}{\lambda} = \underset{\sim}{\lambda}^*$

The Maximization Step (M-step). Obtain new values for ξ and $\underset{\sim}{\lambda}$ by maximizing the expectation $\varepsilon(\xi,\underset{\sim}{\lambda}|\xi^*,\underset{\sim}{\lambda}^*)$ obtained at the E-step. Return to the E-step and keep cycling until convergence.

In a University of Wisconsin 1981 technical report, C. F. Wu was the first to give general conditions under which this procedure converges to the maximum of the marginal likelihood (1.1). The above procedure for hyper-parameters has been employed in particular cases by Laird (1978) and Chen (1980).

Example: Multiple Regression Models.

Consider firstly the estimation of the parameters of several multiple regressions (e.g. Lindley and Smith, 1972, Smith 1973). Suppose we observe vectors $\underset{\sim}{y}_1,\cdots,\underset{\sim}{y}_m$, of respective dimensions $n_1,\cdots,n_m$, where

$$\underset{\sim}{y}_i|\underset{\sim}{\beta}_i,\sigma_i^2 \sim IN(\underset{\sim}{x}_i,\underset{\sim}{\beta}_i,\sigma_i^2 I_{n_i}) \quad (i = 1,\cdots,m),$$

and that $\underset{\sim}{\beta}_1,\cdots,\underset{\sim}{\beta}_m$ may be viewed as exchangeable with

$$\underset{\sim}{\beta}_i|\underset{\sim}{\mu},\underset{\sim}{C} \sim IN(\underset{\sim}{\mu},\underset{\sim}{C}).$$

This assumption will frequently be appropriate when we possess a symmetry of prior ignorance about the $\underset{\sim}{\beta}_i$. It greatly assists us in modelling the prior structure, since we need now just concern ourselves with the within-

regression covariance structure represented by $\underset{\sim}{C}$. This however may be simply estimated from the data, so that no further prior modelling is required. When $\underset{\sim}{\mu}$ and $\underset{\sim}{\sigma}^2 = (\sigma_1^2,\cdots,\sigma_n^2)^2$, are known, the posterior distribution of the $\underset{\sim}{\beta}_i$ is

$$\underset{\sim}{\beta}_i | \underset{\sim}{\mu},\underset{\sim}{C},\underset{\sim}{\sigma}^2 \sim IN(\underset{\sim}{\beta}_i^*,\underset{\sim}{D}_i)$$

where

$$\underset{\sim}{\beta}_i^* = (\sigma_i^{-2}\underset{\sim}{X}_i^T\underset{\sim}{X}_i + \underset{\sim}{C}^{-1})^{-1}(\sigma_i^{-2}\underset{\sim}{X}_i^T\underset{\sim}{y}_i + \underset{\sim}{C}^{-1}\underset{\sim}{\mu}) \quad (i = 1,\cdots,m) \tag{2.1}$$

and

$$\underset{\sim}{D}_i^{-1} = \sigma_i^{-2}\underset{\sim}{X}_i^T\underset{\sim}{X}_i + \underset{\sim}{C}^{-1}. \tag{2.2}$$

Rather than bringing in a set of complicated prior assumptions for $\underset{\sim}{\mu},\underset{\sim}{C}$, and $\underset{\sim}{\sigma}^2$ it is much more straightforward to simply empirically estimate this quantities via their marginal likelihood. Combinations of the E-step and M-step in Section 2 tells us that the marginal maximum likelihood estimates satisfy the equations

$$\underset{\sim}{\mu} = \underset{\sim}{\beta}_\cdot^* = m^{-1}\sum_{i=1}^{m}\underset{\sim}{\beta}_i^* \tag{2.3}$$

$$\underset{\sim}{C} = m^{-1}\sum_{i=1}^{m}(\underset{\sim}{\beta}_i^* - \underset{\sim}{\beta}_\cdot^*)(\underset{\sim}{\beta}_i^* - \underset{\sim}{\beta}_\cdot^*)^T + m^{-1}\text{trace}(\underset{\sim}{D}_i) \tag{2.4}$$

and

$$\sigma_i^2 = n_i^{-1}\sum_{i=1}^{m}(\underset{\sim}{y}_i - \underset{\sim}{X}_i\underset{\sim}{\beta}_i^*)^T(\underset{\sim}{y}_i - \underset{\sim}{X}_i\underset{\sim}{\beta}_i^*) + n_i^{-1}\text{ trace}(\underset{\sim}{X}_i\underset{\sim}{D}_i\underset{\sim}{X}_i^T) \tag{2.5}$$

where $\underset{\sim}{\beta}_i$ and $\underset{\sim}{D}_i$ satisfy (2.1) and (2.2).

Moreover, convergence is guaranteed if we substitute trial values for $\underset{\sim}{\mu},\underset{\sim}{C}$ and $\underset{\sim}{\sigma}^2$ into the right hand sides of (2.1) and (2.2), and put the values obtained for $\underset{\sim}{\beta}_i^*$ and $\underset{\sim}{D}_i$ into the right hand sides of (2.3) - (2.5), then return to (2.1) and (2.2) and repeat the procedure until convergence.

In the prior ignorance case, the procedure proposed by Smith reduces to the above equations but with the important second terms in the right hand sides of (2.4) and (2.5) omitted. Hence overshrinkages towards $\underset{\sim}{\beta}^*$ were

observed in his numerical example. With the extra terms included large shrinkages may still occur, but only when the data suggest that this is reasonable.

Let us now turn to the single multiple regression situation

$$\underset{\sim}{Y}|\underset{\sim}{\beta},\sigma^2 \sim IN(\underset{\sim}{X},\underset{\sim}{\beta},\sigma^2\underset{\sim}{I}_n)$$

and suppose that the columns of $\underset{\sim}{X}$ relate to p different sets of observed explanatory variables. We now make the (obvious) claim that no general purpose non-uniform prior structure exists, and that meaningful prior structures can only be developed if we utilize background information concerning the nature of the explanatory variables, or informative prior knowledge about $\underset{\sim}{\beta}$. In the absence of such information we should simply estimate $\underset{\sim}{\beta}$ by least squares.

In particular, the exchangeable prior

$$\underset{\sim}{\beta}|\tau^2 \sim N(\underset{\sim}{0},\tau^2\underset{\sim}{I}_p)$$

has been proposed as a means of justifying the ridge estimator

$$\underset{\sim}{\beta}^* = (\underset{\sim}{X}^T\underset{\sim}{X} + k\underset{\sim}{I}_p)^{-1}\underset{\sim}{X}^T\underset{\sim}{Y} \qquad (2.6)$$

since the posterior mean of $\underset{\sim}{\beta}$ may be obtained by setting

$$k = \sigma^2/\tau^2 \qquad (2.7)$$

in (2.6).

However, neither exchangeability nor ridge regression seem appropriate in prior ignorance situations. Any non-trivial linear transformation on the $\underset{\sim}{X}$ matrix focuses attention on a new set of parameters which are not exchangeable if the elements of $\underset{\sim}{\beta}$ are exchangeable. In prior ignorance situations there is no way of discerning to which form of the model the ridge estimator should be applied. It therefore does not make any sense to adjust any set of parameter estimates towards zero, or towards any other origin; and the least squares vector possesses greater statistical viability.

If some way could be found of justifying the estimator in (2.6) then it would of course be easy to estimate σ^2 and τ^2, and hence k, by the EM algorithm. The equations are

$$\sigma^2 = n^{-1}(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}^*)^T(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta}^*) + n^{-1} \text{ trace } (\underset{\sim}{X}\ \underset{\sim}{D}\ \underset{\sim}{X}^T) \quad (2.8)$$

and

$$\tau^2 = p^{-1}\underset{\sim}{\beta}^{*T}\underset{\sim}{\beta}^* + \text{trace}(\underset{\sim}{D}) \quad (2.9)$$

where

$$\underset{\sim}{D}^{-1} = \sigma^{-2}\underset{\sim}{X}^T\underset{\sim}{X} + \tau^{-2}\underset{\sim}{I}_p. \quad (2.10)$$

It is possible to show that the solutions to these equations are noticeably more conservative in shrinking towards zero than is the ridge trace method of Hoerl and Kennard (1971).

However, it is particularly important, in this single multiple regression situation to base any deviation from the least squares vector upon definite prior knowledge.

3. Smoothing the Probabilities in a Histogram.

Consider a grouped histogram concentrated on a bounded interval (a,b), and with s cells $I_1, I_2, \cdots, I_s$ of equal width. Let $\theta_1, \cdots, \theta_s$ sum to unity and denote the cell probabilities, taken in order, and let $x_1, \cdots, x_s$ summing to n, denote the cell probabilities. We let the x's satisfy the usual multinomial assumptions given the θ's, and provide a number or refinements and improvements to the method proposed by Leonard (1973) for obtaining smooth estimates of the θ's.

It is easier to think in terms of prior structures by making multivariate normal assumptions about the logits $\gamma_1, \cdots, \gamma_s$ satisfying

$$\theta_j = e^{\gamma_j} / \sum_{j=1}^{s} e^{\gamma_s} \qquad (j = 1, \cdots, s) \tag{3.1}$$

In particular we assume that the log-contrasts

$$\xi_j = \gamma_j - \gamma_{j+1} \qquad (j = 1, \cdots, s-1) \tag{3.2}$$

possess prior covariance structure

$$\mathrm{cov}(\xi_j, \xi_k) = \sigma^2 \rho^{|j-k|} \qquad (0 < \sigma^2 < \infty;\ |\rho| < 1). \tag{3.3}$$

We originally assumed the structure in (4.3) for the logits themselves, but when considering the continuous case (see Leonard, 1978), it became apparent that, by taking differences first, a more reasonable smoothing (avoiding too much flattening) of the θ's would result. The hyperparameter ρ measures the degree of smoothness of the hypothetical true density of the raw observations, and σ^2 measures the 'closeness' to the 'null hypothesis' discussed below.

We hence assume that the vector $\underset{\sim}{\gamma}$ possesses a multivariate normal prior distribution

$$\underset{\sim}{\gamma} | \underset{\sim}{\mu}, \sigma^2, \rho \sim N(\underset{\sim}{\mu}, \underset{\sim}{C}) \tag{3.4}$$

where

$$\underset{\sim}{C} = \underset{\sim}{B}\ \underset{\sim}{A}\ \underset{\sim}{B}^T \tag{3.5}$$

with the $(j,k)^{th}$ element of the $(s-1) \times (s-1)$ matrix $\underset{\sim}{A}$ equal to the quantity in (4.3) and

$$\underset{\sim}{\beta} = \begin{pmatrix} 0 & 0 & 0 & - & - & - & - & 0 \\ 1 & 0 & 0 & & & & & \vdots \\ 1 & 1 & 0 & & & & & \vdots \\ 1 & 1 & 1 & 0 & & & & \vdots \\ \vdots & & & 1 & 0 & & & \vdots \\ \vdots & & & & 1 & \ddots & & \vdots \\ \vdots & & & & & \ddots & 0 & \vdots \\ 1 & 1 & 1 & - & - & - & 1 & \end{pmatrix} \qquad (3.6)$$

The choice of the prior vector $\underset{\sim}{\mu}$ may be based upon the statistician's 'null hypothesis' about the probabilities in the histogram. For example, if his hypothesis that the underlying density of the raw observations is $f_0(t)$ for $t \in (a,b)$ then he should set

$$\mu_j = \log[\int_{I_j} f_0(t)dt] \quad \text{for } j = 1,\cdots,s. \qquad (3.7)$$

If any unknown parameters appear in his choice of $f_0(t)$ then we suggest that these should be estimated from the data by conventional techniques. The shrinkage parameter σ^2 plays the role of the significance level in the standard chi-square goodness of fit test.

In this situation we of course do not have exchangeability of the γ_j or the ξ_j. However, the second order lagged differences

$$(\gamma_{j+2} - \mu_{j+2} - \gamma_{j+1} + \mu_{j+1}) - \rho(\gamma_{j+1} - \mu_{j+1} - \gamma_j + \mu_j)$$

are exchangeable for $j = 1,\cdots,s-2$. This ties in with the philosophy outlined in Section 1 of seeking appropriate transformations of the parameters which are exchangeable. The lagged differences create an autoregressive type smoothing on the θ's so that the posterior estimates of θ_j will take account of estimates in adjacent intervals.

When $\underset{\sim}{\mu}, \sigma^2$, and ρ are specified, we have the approximate posterior distribution

$$\underset{\sim}{\gamma} | \underset{\sim}{x}, \underset{\sim}{\mu}, \sigma^2\rho \sim N(\check{\underset{\sim}{\gamma}}, \underset{\sim}{D}) \tag{3.8}$$

where $\check{\underset{\sim}{\gamma}}$ is the posterior mode vector, and satisfies the non-linear system

$$n \check{\underset{\sim}{\theta}} = \underset{\sim}{x} - \underset{\sim}{C}^{-1}(\check{\underset{\sim}{\gamma}} - \underset{\sim}{\mu}) \tag{3.9}$$

where $\check{\underset{\sim}{\theta}} = (\check{\theta}_1, \cdots, \check{\theta}_s)^T$, with $\check{\theta}_j = e^{\check{\gamma}_j} / \sum e^{\check{\gamma}_g}$, and

$$\underset{\sim}{D}^{-1} = n[\text{diag}(\check{\theta}_1, \cdots, \check{\theta}_s) - \check{\underset{\sim}{\theta}}\check{\underset{\sim}{\theta}}^T] + \underset{\sim}{C}^{-1} \tag{3.10}$$

denotes the posterior information matrix.

This system can be solved via Newton-Raphson, and the EM algorithm can be used to (approximately) estimate σ^2 and $\sigma^2\rho$ by

$$\sigma^2 = s^{-1} \sum_{j=1}^{s-1} (\check{\xi}_j - \eta_j)^2 + s^{-1} \sum_{j=1}^{s} r_{jj} \tag{3.11}$$

and

$$\sigma^2\rho = s^{-1} \sum_{j=1}^{s-2} (\check{\xi}_j - \eta_j)(\check{\xi}_j - \eta_{j+1}) + s^{-1} \sum_{j=1}^{s-2} r_{j,j+1} \tag{3.12}$$

where $\check{\xi}_j = \check{\gamma}_j - \check{\gamma}_{j+1}$, $\eta_j = \mu_j - \mu_{j+1}$, and r_{jk} is the $(j,k)^{th}$ element of

$$\underset{\sim}{R} = \underset{\sim}{G}\ \underset{\sim}{D}\ \underset{\sim}{G}^T$$

where

$$\underset{\sim}{R} = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & \ddots & \ddots & \\ & & & & 1 & -1 \end{pmatrix}$$

Cyclic substitutions are again appropriate. See Leonard (1978, p. 115) for a numerical example.

4. Simultaneous Estimation for Several Multinomial Distributions.

Suppose now that, for $i = 1,\cdots,m$, the elements of the $s \times 1$ vector $\underset{\sim}{x}_i$ possess a multinomial distribution with sample size n_i, given the vector of ceall probabilities $\underset{\sim}{\theta}_i = (\theta_{i},\cdots,\theta_{is})^T$, and that these m multinomial distributions are independent, given the parameters. We now have m sets of logit vectors $\underset{\sim}{\gamma}_i = (\gamma_{i1},\cdots,\gamma_{is})^T$ satisfying

$$\theta_{ij} = e^{\gamma_{ij}} / \sum_g e^{\gamma_{ig}} \qquad (i = 1,\cdots,m;\ j = 1,\cdots,s). \tag{4.1}$$

As in the several regression line situation, it will often be appropriate to assume exchangeability between the parameter vectors $\underset{\sim}{\theta}_1,\cdots,\underset{\sim}{\theta}_m$ rather than between the elements within any particular vector. In this case we assume

$$\underset{\sim}{\gamma}_i \mid \underset{\sim}{\mu},\underset{\sim}{C} \sim IN(\underset{\sim}{\mu},\underset{\sim}{C}) \qquad (i = 1,\cdots,m)$$

yielding the approximate posterior distribution

$$\underset{\sim}{\gamma}_i \mid \underset{\sim}{x},\underset{\sim}{\mu},\underset{\sim}{C} \sim IN(\check{\underset{\sim}{\gamma}}_i,\underset{\sim}{D}_i)$$

where

$$n_i \check{\underset{\sim}{\theta}}_i = \underset{\sim}{x}_i - \underset{\sim}{C}^{-1}(\underset{\sim}{\gamma}_i - \underset{\sim}{\mu}) \qquad (i = 1,\cdots,m) \tag{4.2}$$

with $\check{\underset{\sim}{\theta}}_i = (\check{\theta}_{i1},\cdots,\check{\theta}_{is})^T$ where

$$\check{\theta}_{ij} = e^{\gamma_{ij}} / \sum_g e^{\check{\gamma}_{ig}}$$

and

$$\underset{\sim}{D}_i^{-1} = n_i\{\mathrm{diag}(\check{\theta}_{i1},\cdots,\check{\theta}_{ij}) - \check{\underset{\sim}{\theta}}_i \check{\underset{\sim}{\theta}}_i^T\} + \underset{\sim}{C}^{-1}. \tag{4.3}$$

As we have replications on $\underset{\sim}{\mu}$ and $\underset{\sim}{C}$ is is no longer necessary to assume a specific covariance structure like (4.3) within each multinomial, since $\underset{\sim}{\mu}$ and $\underset{\sim}{C}$ can be estimated in their entirety from the data; EM gives

$$\underset{\sim}{\mu} = \check{\underset{\sim}{\gamma}}$$

and

$$\underset{\sim}{C} = m^{-1} \sum_{i=1}^{m} (\check{\underset{\sim}{\gamma}}_i - \check{\underset{\sim}{\gamma}}_\cdot)(\check{\underset{\sim}{\gamma}}_i - \check{\underset{\sim}{\gamma}}_\cdot)^T + m^{-1} \sum_{i=1}^{m} \underset{\sim}{R}_i^{-1}$$

which may be solved iteratively together with (5.2) and (5.3).

In the first six columns of Table 1 we give the percentages of pupils respectively obtaining grades one to six on a test taken at 40 different schools. We assumed exchangeability between the schools and obtained the following estimates for $\underset{\sim}{\mu}$, diag($\underset{\sim}{C}$), and the correlation matrix $\underset{\sim}{B}$ associated with $\underset{\sim}{C}$:

$$\underset{\sim}{\mu} = (-0.51,\ 0.41,\ 0.56,\ 0.59,\ -0.82,\ -0.24)^T$$

corresponding to a common prior estimate of

$$\underset{\sim}{\xi} = (0.087,\ 0.217,\ 0.255,\ 0.262,\ 0.064,\ 0.115)^T$$

for the $\underset{\sim}{\theta}_i$,

$$\text{diag}(\underset{\sim}{C}) = (1.08,\ 0.39,\ 0.25,\ 0.26,\ 0.46,\ 0.92),$$

and

$$\underset{\sim}{B} = \begin{pmatrix} 1 & 0.79 & 0.57 & -0.07 & -0.38 & -0.57 \\ 0.79 & 1 & 0.79 & 0.29 & -0.14 & -0.31 \\ 0.57 & 0.79 & 1 & 0.61 & 0.21 & -0.01 \\ -0.07 & 0.29 & 0.61 & 1 & 0.68 & 0.62 \\ -0.38 & -0.14 & 0.21 & 0.68 & 1 & 0.83 \\ -0.51 & -0.31 & -0.01 & 0.62 & 0.83 & 1 \end{pmatrix}.$$

Note that the matrix $\underset{\sim}{B}$ gives moderately high correlations between adjacent cells within each school and negative correlations between γ_{ij} and $\gamma_{i,j+k}$ for $|k| > 2$. The between school exchangeability has helped us to estimate the prior structure within each school. This is similar in spirit to the autoregressive structure in (4.3) since it enables us to take account to the ordering of the cells.

The smoothed percentages in Table 1 smooth the observed percentages (a) by shrinkages utilizing the collateral information in the common vector $100\underset{\sim}{\xi}$ by smoothing within each school; using the convariance structure estimated via $\underset{\sim}{C}$. Note, for example, that the zeros create no extra problem: their smoothed values are all positive, the amount depending on collateral and within school information.

TABLE 1: OBSERVED AND SMOOTHED PERCENTAGES

	OBSERVED						SMOOTHED						
$i \backslash j$	1	2	3	4	5	6	1	2	3	4	5	6	n_i
1	6.7	17.8	24.4	28.9	6.7	15.6	6.6	18.8	24.2	28.2	7.2	15.0	45
2	0.0	21.6	24.3	18.9	13.5	21.6	3.8	16,9	22.2	26.2	9.9	20.9	37
3	5.3	15.8	42.1	26.3	5.3	5.3	8.3	21.4	29.1	26.5	5.8	8.9	19
4	22.2	25.9	29.6	11.1	7.4	3.7	19.5	26.9	26.5	17.9	4.2	4.9	27
5	16.7	33.3	11.1	16.7	5.6	16.7	12.7	25.4	22.8	22.3	5.7	11.1	18
6	5.9	7.4	26.5	33.8	8.8	17.6	4.8	12.9	23.8	31.5	8.9	18.0	68
7	31.7	17.1	14.6	19.5	9.8	7.3	24.1	22.1	21.7	19.1	5.7	7.4	41
8	4.5	4.5	22.7	31.8	9.1	27.3	3.6	12.4	20.2	29.7	9.6	24.5	22
9	16,1	45.2	19.4	9.7	3.2	6.4	17.5	33.6	23.8	16.8	3.4	5.0	31
10	13.0	31.5	31.5	24.1	0.0	0.0	14.9	30.3	29.1	20.5	2.4	2.9	54
11	23.5	35.3	20.6	20.6	0.0	0,0	22.7	31.7	24.3	16.5	2.2	2.6	34
12	22.8	26.3	21.1	15.8	3.5	10.5	19.8	26.2	23.0	18.7	4.4	7.8	57
13	7.1	14.2	14.2	32.1	10.7	21.4	5.3	15.7	20.5	29.0	9.0	20.4	28
14	13.9	33.3	25.0	19.4	0.0	8.3	14.2	29.2	25.6	20.9	3.7	6.5	36
15	0.0	18.2	13.6	54.5	9.1	4.5	4.2	17.2	23.2	34.7	7.4	13.4	22
16	12.5	31.3	18.8	18.8	6.3	12.5	11.3	24.8	24.6	23.4	5.7	10.1	16
17	0.0	9.4	25.0	50.0	9.4	6.3	3.3	14.3	24.4	36.3	8.0	13.7	32
18	0.0	5.3	15.8	21.1	26.3	31.6	2.0	9.1	16.1	26.6	13.9	32.3	19
19	0.0	3.7	7.4	14.8	3.7	70.4	0.7	4.6	8.4	20.0	9.4	57.1	27
20	2.4	7.3	19.5	36.6	4.8	29.3	2.7	11.3	18.8	31.9	8.4	27.0	41
21	29.2	16.7	8.3	37.5	4.2	4.2	19.5	23.5	22.8	23.3	4.4	6.5	24
22	14.2	21.4	42.9	21.4	0.0	0.0	14.8	26.4	29.0	21.2	3.7	4.8	14
23	0.0	17.4	34.8	17.4	17.4	13.0	4.8	17.2	25.5	26.8	9.7	15.9	23
24	5.6	19.4	30.1	22.2	11.1	11.1	7.1	19.9	26.5	26.1	7.9	12.5	36
25	31.3	18.8	28.1	15.6	3.1	3.1	25.6	25.3	25.5	16.7	3.2	3.7	32
26	9.5	19.0	42.9	14.3	0.0	14.3	10.1	22.9	28.4	23.5	5.2	9.8	21
27	5.6	37.0	22.2	25.9	3.7	5.6	8.8	30.0	25.4	24.5	4.3	7.1	54
28	0.0	30.8	7.7	30.8	0.0	30.8	4.4	17.7	20.2	28.7	7.4	21.6	13
29	4.0	12.0	16.0	32.0	8.0	28.0	3.7	13.6	19.3	29.5	9.1	24.8	25
30	0.0	16.7	27.8	33.3	5.6	16.7	4.5	17.1	23.9	30.1	7.6	16.8	18
31	3.7	11.1	37.0	37.0	0.0	11.1	5.8	18.0	27.6	30.9	5.8	11.9	27
32	0.0	42.9	28.6	21.4	7.1	0.0	9.6	27.6	27.4	23.6	4.8	6.9	14
33	0.0	14.3	28.6	21.4	21.4	14.3	4.5	15.8	23.2	28.1	10.1	18.2	14
34	19.5	39.0	17.1	22.0	0.0	2.4	19.4	33.0	23.4	18.1	2.5	3.5	41
35	37.9	13.8	37.9	6.9	3.4	0.0	31.8	24.6	26.2	12.8	2.4	2.1	29
36	0.0	18.8	6.2	25.0	6.3	43.8	2.5	11.9	15.9	27.2	9.4	33.1	16
37	13.3	20.0	20.0	26.7	6.7	13.3	9.8	21.6	24.5	25.7	6.4	11.8	15
38	16.7	37.5	41.7	4.2	0.0	0.0	20.3	32.4	28.3	14.3	2.2	2.3	24
39	18.3	31.7	21.7	15.0	6.7	6.7	17.5	28.7	24.0	18.4	4.8	6.5	60
40	0.0	11.1	11.1	33.3	22.2	22.2	3.4	13.3	19.7	29.4	10.5	23.7	9

REFERENCES

Atilgan, T. and Leonard, T. (1982) Penalized likelihood methods for density estimation. Mathematics Research Center Technical Summary Report University of Wisconsin - Madison.

Box, G. E. P. (1981) Sampling inference, Bayes inference, and robustness in the advancement of learning (with Discussion) in Bayesian Statistics ed. J. M. Bernardo) University of Valencis Press, 366-384.

Chen, C. (1980) An empirical Bayes method for estimating a covariance matrix, J. Roy. Statist. Soc., B, 42.

Dempster, A. P., Laird, W. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm, (with Discussion) J. Roy. Statist. Soc Ser. B 39 1-38.

Hoerl, A. E. and Kennard, R. W. (1970) Ridge regression in biased estimation for non-orthogonal problems. Technometrics 12 55-67.

Laird, N. M. (1978) Empirical Bayes estimation for two-way contingency tables. Biometrika 65.

Leonard, T. (1973) A Bayesian method for histograms. Biometrika 60, 297-308.

Leonard, T. (1976) Some alternatives to multiparameter estimation. Biometrika 63, 69-75.

Leonard, T. (1978) Density estimation, stochastic processes, and prior information. (with Discussion) J. Roy. Statist. Soc. B 40, 113-146.

Leonard, T. (1981) The roles of inductive, modelling and coherence in Bayesian Statistics (with Discussion) in Bayesian Statistics ed. J. M. Bernardo) University of Valencia Press.

Leonard, T. (1982) Some philosophies of inference and modelling (with Discussion) in Scientific Inference, Data Analysis and Robustness (ed. G. E. P. Box, T. Leonard, and C. F. Wu) Academic Press, 1982, to appear.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979) On the reconciliation of probability assessments (with Discussion) J. Roy. Statist. Sur. A, 142, 146-198.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with Discussion) J. Roy. Statist. Soc. B 34, 1-41.

Smith, A. F. M. (1973) A general Bayesian linear model. J. Roy. Statist. Soc. B, 35, 67-75.

TL/db

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2344	2. GOVT ACCESSION NO. AD-A114 537	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Applications of the EM Algorithm to the Estimation of Bayesian Hyperparameters		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Tom Leonard		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 4 -Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE March 1982 13. NUMBER OF PAGES 16
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Hyperparameter, EM algorithm, multiple regression, ridge regression, histogram smoothing, exchangeability, shrinkage estimators, multinomial distributions

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Applications of the EM algorithm to the estimation of Bayesian hyperparameters are discussed and reviewed in the context of the author's philosophy involving the inductive and pragmatic modelling of sampling distributions and prior structures. Frequently the hyperparameters may be estimated from the data, thus avoiding the subjective assessment of these values. The ideas are applied to multiple regression models, histograms and multinomial distributions. A numerical example is described in the context of smoothing the cell probabilities of several multinomial distributions.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DATE
ILME
8