Product Partition Models for Normal Means

by

Evelyn M. Crowley
Purdue University

Technical Report #93–37

Department of Statistics
Purdue University

July 1993

△

Abstract

We consider probability models for the estimation of normal means that allow for some of the means to be equal. These probability models, which are called product partition models, specify prior probabilities for a random partition. The posterior probability of the partition given the observations has the same form. The resulting estimate of the means, the product estimate, is obtained by conditioning on the partition and summing over all possible partitions. The large number of computations involved leads to the use of Markov sampling to compute the product estimate. We compare the product estimate to other estimates of normal means both in a simulation study and in the prediction of batting averages.

KEY WORDS: Partition; Product estimate; Clustered means; Markov sampling.

1. INTRODUCTION

In this article, we will consider the normal means problem: $X_i|\mu_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$. We are interested in estimating the means. We will do this using product partition models. These partition the objects $\{1, \ldots, n\}$ into several sets; within each set, the μ_i's are equal. We specify a prior probability distribution for a random partition ρ and update this distribution into a posterior distribution of the same form. Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. We obtain the product estimate by conditioning on the partition and summing over all possible partitions. We will use Markov sampling to compute the product estimate.

Other approaches to estimating normal means include Stein estimation and its generalisations, parametric and nonparametric empirical Bayes estimation and maximum likelihood estimation (Efron and Morris 1972, 1973, 1975; Escobar 1992; George 1986a,b; James and Stein 1961; Laird 1978; Stein 1956).

In Section 2, product partition models are described in general. The need for Markov sampling is discussed in Section 3. In Section 4, we describe product partition models for the specific case of the normal means problem. In Section 5, we discuss the method we use to compute the product estimate - Markov sampling. We describe the general technique of Markov sampling as well as the computations involved in applying the technique to our problem. In Section 6, the product estimate is compared to the maximum likelihood estimate, an empirical Bayes estimate and a nonparametric mixture estimate in a simulation study. We consider a range of different settings of parameters. The product estimate has mean square error comparable to the best of the other estimates in many of the settings and is superior for a wide variety of settings. Finally, in Section 7, we use the above estimates to predict batting

averages.

2. PRODUCT PARTITION MODELS

Hartigan (1990) has developed a method for constructing probability models by means of random partitions. Given a set of objects $S_0 = \{1, 2, \ldots, n\}$, a partition $\rho = \{S_1, S_2, \ldots, S_k\}$ is defined by the property that $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_i S_i = S_0$. The partition ρ has probability

$$P(\rho = \{S_1, S_2, \ldots, S_k\}) = K \prod_{i=1}^{k} c(S_i) \tag{1}$$

where $c(S) \geq 0$ is a *cohesion* defined for each $S \subseteq S_0$ and K is chosen so that the probabilities sum to one over all possible partitions ρ. Notice that the cohesions are not uniquely determined: the probabilities

$$P(\rho = \{S_1, S_2, \ldots, S_k\}) = K^* \prod_{i=1}^{k} c^*(S_i), \tag{2}$$

where $c^*(S) = (\prod_{i \in S} \alpha_i) c(S)$ for arbitrary positive α_i and K^* is chosen so that the probabilities sum to one over all possible partitions ρ, define the same random partitions. We will use this later to simplify some calculations.

For each object i, there is an observation X_i. Let X_S denote the vector of observations for $i \in S$ and let $p_S(X_S)$ be the conditional density for the observations in a set S, given that $S \in \rho$. Given the random partition $\rho = \{S_1, S_2, \ldots, S_k\}$, the observations $X_{S_1}, X_{S_2}, \ldots, X_{S_k}$ are independent with density

$$p(\mathbf{X} | \rho = \{S_1, S_2, \ldots, S_k\}) = \prod_{i=1}^{k} p_{S_i}(X_{S_i}). \tag{3}$$

Definitions (1) and (3) uniquely determine the joint distribution of $\mathbf{X}$ and ρ, and the marginal density of $\mathbf{X}$.

The posterior probability of ρ given the observations is

$$P(\rho = \{S_1, S_2, \ldots, S_k\}|\mathbf{X}) = (K/\nu(\mathbf{X})) \prod_{i=1}^{k} c(S_i) p_{S_i}(X_{S_i}),$$

where $\nu(\mathbf{X})$ is the marginal density of $\mathbf{X}$. This is also a product partition model with cohesions $c(S)p_S(X_S)$, which will be called posterior cohesions. Thus, product partition models have computational advantages similar to those for conjugate priors in traditional Bayes theory. For some applications of product partition models, see Hartigan (1990) and Barry and Hartigan (1992, 1993).

3. ENUMERATION OF PARTITIONS

For a fixed subset S of S_0, the probability that it is a set in the random partition ρ is called the *relevance* of S, denoted by $r(S)$. The relevance of S may be computed from the cohesions using the function

$$\begin{aligned} \lambda(\emptyset) &= 1, \\ \lambda(S) &= \sum c(S_1) \ldots c(S_m), \end{aligned}$$

where the summation is over all partitions of S into subsets $S_1, \ldots, S_m$. To compute λ, the following recursion may be used: choose a particular object $i \in S$ and sum over all $T \subseteq S$ that contain i

$$\lambda(S) = \sum_{\{T|i \in T\}} c(T)\, \lambda(S - T).$$

The relevance of S can be written as

$$r(S) = \frac{c(S)\, \lambda(S_0 - S)}{\lambda(S_0)}. \tag{4}$$

The relevances are useful because when computing the product estimate, we can condition on the set and sum over all possible sets instead of conditioning on the partition and summing over all possible partitions. So we will

need the posterior relevances of all subsets of S_0 instead of the posterior probabilities of all possible partitions. The number of partitions of n objects into k sets is a Stirling number of the second kind:

$$\left[k^n - \binom{k}{1}(k-1)^n + \binom{k}{2}(k-2)^n + \ldots (-1)^{k-1}\binom{k}{k-1}\right]/k!.$$

The number of possible partitions of n objects increases at a rate faster than exponential and slower than factorial. For example, when $n = 10$, the number of possible partitions is equal to 112,519. The number of possible sets of n objects is equal to $2^n - 1$, which is much smaller than the number of possible partitions.

Although we have reduced the calculations considerably, we still need to compute the relevances of the $2^n - 1$ sets to compute the product estimate. This is not feasible in practice (except for small sample sizes). Instead, we will use Markov sampling to compute an approximation to the product estimate (see Section 5). We will, however, use the relevances to check the accuracy of the approximation for small sample sizes.

4. CLUSTERED MEANS

We will focus on the following normal means problem using product partition models: $X_i|\mu_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$. For notational convenience (when $n < 10$), we will denote, for example, the partition $\{\{2\}, \{1,4\}, \{3\}\}$, by (2)(14)(3). Let μ^S be the common mean for the μ_i's with $i \in S$: $\mu_i = \mu^S$, $i \in S$. So if $\rho = (6)(253)(14)$, we have $\mu_6 = \mu^{(6)}$, $\mu_2 = \mu_5 = \mu_3 = \mu^{(253)}$ and $\mu_1 = \mu_4 = \mu^{(14)}$. Let $f_S(\mu^S)$ be the conditional density of μ^S, given that $S \in \rho$. The conditional density of X_S, given that $S \in \rho$ can be written

$$p_S(X_S) = \int p_S(X_S|\mu^S, S \in \rho)\, f_S(\mu^S)\, d\mu^S. \tag{5}$$

From conditional independence, $p_S(X_S|\mu^S, S \in \rho)$ is a product of normal densities. Let $I\!P$ denote expectation. The posterior distribution of μ_i given $\mathbf{X}$ depends only on values for the member of the partition to which it belongs:

$$\hat{\mu}_i = I\!P(\mu_i|\mathbf{X}) = \sum_{\{S|i\in S\}} I\!P(\mu^S|S \in \rho, X_S)\, r(S|\mathbf{X}),$$

where $r(S|\mathbf{X})$ is the posterior relevance of S. That is, the estimate of μ_i is a weighted average of the estimates based on the sets S that include i, the weights being given by the posterior probabilities of these sets given the observations. The estimate of $\boldsymbol{\mu}$, $(\hat{\mu}_1, \ldots, \hat{\mu}_n)$, will be called the *product estimate.* There are three components in the problem:

(a) Prior cohesions.

Let $c(S) = (n_S - 1)!/m^{(n_S-1)}$ where n_S is the number of objects in the set S and m a parameter which we will need to estimate. The relevance of S is equal to $m\, B(n_S, n + m - n_S)$. Large values of m encourage small sets.

(b) Distribution of the data.

We have $X_i|\mu_i \sim N(\mu_i, 1), i = 1, \ldots, n$. Therefore $p_S(X_S|\mu^S, S \in \rho) = \prod_{i\in S} (2\pi)^{-1/2} \exp(-(X_i - \mu^S)^2/2)$.

(c) Prior for parameters.

Let $\mu^S \sim N(\mu_0, {\sigma_0}^2/n_S)$, where μ_0 and σ_0^2 are parameters which we will need to estimate. Then $f_S(\mu^S) = (n_S/2\pi\sigma_0^2)^{1/2} \exp(-n_S\,(\mu^S - \mu_0)^2/2\sigma_0^2)$. The prior distribution for μ^S is chosen so that the parameter value for small sets varies more from the overall mean than the parameter value for large sets; this is to discourage small sets unless the corresponding μ^S is quite different from the other μ_i's.

Together, (a), (b) and (c) determine all the other distributions in the prod-

uct partition model. Substituting in equation (5) gives

$$p_S(X_S) = (2\pi)^{-n_S/2}(1+\sigma_0^2)^{-1/2} \times \exp\left(-\frac{1}{2}\left(\sum_{i\in S}(X_i-\overline{X}_S)^2 + \frac{n_S}{1+\sigma_0^2}(\overline{X}_S-\mu_0)^2\right)\right), \quad (6)$$

where $\overline{X}_S = \sum_{i\in S} X_i/n_S$. The posterior density of μ^S conditional on $S \in \rho$ is normal with mean $(\sigma_0^2\overline{X}_S + \mu_0)/(\sigma_0^2+1)$ and variance $\sigma_0^2/n_S(\sigma_0^2+1)$.

We now compute the joint distribution of ρ and $\mathbf{X}$, treating the parameters μ_0, σ_0^2 and m as fixed constants. We will need this distribution, $P(\rho = \{S_1, S_2, \ldots, S_k\}, \mathbf{X})$, in the Markov sampling. First, replace the prior cohesions, using (2), by

$$\begin{aligned} c(S) &= ((n_S-1)!/m^{(n_S-1)}) \prod_{i\in S} m\,(2\pi)^{1/2} \exp\left(\frac{1}{2}(X_i-\overline{X})^2\right) \\ &= m\,(n_S-1)!\,(2\pi)^{n_S/2} \exp\left(\frac{1}{2}\sum_{i\in S}(X_i-\overline{X})^2\right). \end{aligned} \quad (7)$$

This gives a model equivalent to that given by the previous choice of cohesions, but it will simplify the expression for $c(S)\,p_S(X_S)$. These prior cohesions give $K = d(\mathbf{X})\,\Gamma(m)/\Gamma(n+m)$, where $d(\mathbf{X})^{-1} = (2\pi)^{n/2}\exp(\sum_{i=1}^n (X_i-\overline{X})^2/2)$. We have

$$P(\rho = \{S_1, S_2, \ldots, S_k\}, \mathbf{X}) = K \prod_{r=1}^{k} c(S_r)\,p_{S_r}(X_{S_r}). \quad (8)$$

From equations (6) and (7) and using the fact that

$$\sum_{i\in S}(X_i-\overline{X})^2 = \sum_{i\in S}(X_i-\overline{X}_S)^2 + n_S\,(\overline{X}_S-\overline{X})^2,$$

we obtain $c(S)\,p_S(X_S)$ to be equal to

$$m\,(n_S-1)!\,(1+\sigma_0^2)^{-1/2} \exp\left(\frac{n_S}{2}(\overline{X}_S-\overline{X})^2\right) \exp\left(-\frac{1}{2}\frac{n_S}{1+\sigma_0^2}(\overline{X}_S-\mu_0)^2\right).$$

Substituting for K and $c(S)\,p_S(X_S)$ in equation (8) and using the fact that

$$\sum_{r=1}^{k} n_{S_r}(\overline{X}_{S_r}-\mu_0)^2 = \sum_{r=1}^{k} n_{S_r}(\overline{X}_{S_r}-\overline{X})^2 + n\,(\overline{X}-\mu_0)^2.$$

it follows that

$$P(\rho = \{S_1, S_2, \ldots, S_k\}, \mathbf{X}) = d(\mathbf{X}) \frac{\Gamma(m)}{\Gamma(n+m)} m^k \left(\prod_{r=1}^{k} (n_{S_r} - 1)! \right)$$
$$\times (1+\sigma_0^2)^{-k/2} \exp\left(-\frac{1}{2} \frac{n}{1+\sigma_0^2} (\overline{X} - \mu_0)^2 \right) .$$
$$\times \exp\left(\frac{1}{2} \frac{\sigma_0^2}{1+\sigma_0^2} \sum_{r=1}^{k} n_{S_r} (\overline{X}_{S_r} - \overline{X})^2 \right) \qquad (9)$$

5. MARKOV SAMPLING

5.1 The Method

There has been a renewal of interest in the technique of Markov sampling recently (see Gelfand and Smith 1990; Gelman and Rubin 1992; Geman and Geman 1984; and Geyer 1992). For earlier work on the technique, see Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and Hastings (1970). Suppose we wish to estimate $I\!P(h(\mathbf{Y}))$, where $\mathbf{Y}$ has probability distribution $I\!P$. The usual method is to generate N random samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ from $I\!P$ and estimate $I\!P(h(\mathbf{Y}))$ by $\sum_{i=1}^{N} h(\mathbf{Y}_i)/N$. Note that $\sum_{i=1}^{N} h(\mathbf{Y}_i)/N$ converges in probability to $I\!P(h(\mathbf{Y}))$. When $\mathbf{Y}$ is complicated and it is difficult to generate $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ in this way, an alternative is to generate $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ from a Markov chain whose unique stationary distribution is $I\!P$. The resulting $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ will be correlated. However, $\sum_{i=1}^{N} h(\mathbf{Y}_i)/N$ computed using this method, will still converge in probability to $I\!P(h(\mathbf{Y}))$.

5.2 Application to the Normal Means Problem

In this subsection, all distributions, probabilities and expectations will be conditional on the data $\mathbf{X}$, although we will not write this out explicitly, since

$\mathbf{X}$ does not change during the simulation. For the problem we are considering, $\mathbf{Y}$ corresponds to the random partition ρ and $I\!P$ corresponds to the distribution of ρ. We have n functions $h_j(\mathbf{Y}) = I\!P(\mu_j|\rho)$, $j = 1, \ldots, n$. Note that $I\!P(h_j(\mathbf{Y})) = I\!P(\mu_j)$. The advantage of Markov sampling is that $I\!P(\mu_j|\rho)$ is easy to compute.

We start with a partition $\rho_1 = (1, 2, \ldots, n)$. One step of the Markov chain consists of moving each object $1, 2, \ldots, n$ in turn to arrive at a new partition ρ_2. We repeat this $N - 2$ times and estimate $I\!P(\mu_j)$ by $\sum_{i=1}^{N} I\!P(\mu_j|\rho_i)/N$. In implementation, the first few samples are often deleted from the average, to remove the effect of the starting value ρ_1. An example of a possible step in the Markov chain starting with $\rho_i = (12)(34)$ is:

move 1: (1)(2)(34)

move 2: (1)(2)(34)

move 3: (1)(23)(4)

move 4: (14)(23)

to get $\rho_{i+1} = (14)(23)$.

The transition probabilities for movement of the i^{th} object are computed as follows. Let the current partition be $\rho' = \{S_1, \ldots, S_k\}$; suppose the i^{th} object lies in S_j. Form a new partition $\{S_1^*, \ldots, S_b^*\}$ with the i^{th} object removed:

$$
\begin{aligned}
n_{S_j} > 1 : S_r^* &= S_r, r = 1, \ldots, j-1, j+1, \ldots, k \\
S_j^* &= S_j - \{i\} \\
b &= k \\
n_{S_j} = 1 : S_r^* &= S_r, \quad r = 1, \ldots, j-1, j+1, \ldots, k-1 \\
S_j^* &= S_k \\
b &= k - 1.
\end{aligned}
$$

Now form a new partition ρ^* including the i^{th} object where

(1) $\rho^* = \{S_1^*, \ldots, \{S_t^*, i\}, \ldots, S_b^*\}$, $t = 1, \ldots, b$ or

(2) a new component $S_{b+1}^* = \{i\}$ is created and $\rho^* = \{S_1^*, \ldots, S_b^*, S_{b+1}^*\}$.

The new partition will be equal to ρ^* with probability proportional to $P(\rho = \rho^*)$. The constant of proportionality is chosen so that the transition probabilities sum to one and is equal to

$$\left(\sum_{r=1}^{b} P(\rho = \{S_1^*, \ldots, \{S_r^*, i\}, \ldots, S_b^*\}) + P(\rho = \{S_1^*, \ldots, S_b^*, S_{b+1}^*\})\right)^{-1}.$$

To simplify the expressions for the probabilities, we divide the numerator and the denominator by $P(\rho = \{S_1^*, \ldots, S_b^*, S_{b+1}^*\})$. Let

$$\begin{aligned} Q(t) &= \frac{P(\rho = \{S_1^*, \ldots, \{S_t^*, i\}, \ldots, S_b^*\})}{P(\rho = \{S_1^*, \ldots, S_b^*, S_{b+1}^*\})}, \; t = 1, \ldots, b \qquad (10) \\ Q(b+1) &= 1. \end{aligned}$$

Then, $\rho' \rightarrow \rho^* = \{S_1^*, \ldots, \{S_t^*, i\}, \ldots, S_b^*\}$ with probability

$$Q(t) \Big/ \sum_{r=1}^{b+1} Q(r) \,, t = 1, \ldots, b$$

or $\rho' \rightarrow \rho^* = \{S_1^*, \ldots, S_b^*, S_{b+1}^*\}$ with probability

$$Q(b+1) \Big/ \sum_{r=1}^{b+1} Q(r) \,.$$

Another way to write this, which we will use below, is

$$\rho' \rightarrow \rho^* \text{ with probability } \frac{I\!P(\rho^*)}{\sum_{\rho_0 | \rho_0 \in R_{\rho'}} I\!P(\rho_0)}, \qquad (11)$$

where $R_{\rho'}$ is the set of partitions which can be obtained from ρ' after moving the i^{th} object. Note that $\rho' \in R_{\rho'}$ and that the sets $R_{\rho'}$ partition the family of all partitions. Therefore $\rho' \in R_{\rho^*}$ implies that $R_{\rho'} = R_{\rho^*}$. The sets $R_{\rho'}$, when

$n = 4$ and $i = 2$, are

$$\begin{aligned}
R_{(1)(23)(4)} &= \{(12)(3)(4), (1)(23)(4), (1)(3)(24), (1)(2)(3)(4)\} \\
R_{(12)(34)} &= \{(12)(34), (1)(234), (1)(2)(34)\} \\
R_{(1234)} &= \{(1234), (2)(134)\} \\
R_{(123)(4)} &= \{(123)(4), (13)(24), (13)(2)(4)\} \\
R_{(14)(23)} &= \{(14)(23), (3)(124), (2)(3)(14)\}.
\end{aligned}$$

The family of all partitions of size four is partitioned into

$$\{R_{(1)(23)(4)}, R_{(12)(34)}, R_{(1234)}, R_{(123)(4)}, R_{(14)(23)}\}.$$

We will now show that the transition probabilities in (10) give us a Markov chain with unique stationary distribution $I\!P$. Assume that ρ^* has probability distribution $I\!P_1$, where ρ^* is the partition obtained from ρ' after moving the i^{th} object. Then

$$\begin{aligned}
I\!P_1(\rho^*) &= \sum_{\rho'} I\!P_1(\rho^*|\rho')\, I\!P(\rho') \\
&= \sum_{\rho'|\rho' \in R_{\rho^*}} I\!P_1(\rho^*|\rho')\, I\!P(\rho') \\
&= \sum_{\rho'|\rho' \in R_{\rho^*}} I\!P(\rho') \frac{I\!P(\rho^*)}{\sum_{\rho_0|\rho_0 \in R_{\rho'}} I\!P(\rho_0)} \\
&= \sum_{\rho'|\rho' \in R_{\rho^*}} I\!P(\rho') \frac{I\!P(\rho^*)}{\sum_{\rho_0|\rho_0 \in R_{\rho^*}} I\!P(\rho_0)} \\
&= I\!P(\rho^*)
\end{aligned}$$

The second equality holds because if $\rho' \notin R_{\rho^*}$, then $I\!P_1(\rho^*|\rho') = 0$), the third is got from (11) and the fourth follows as $\rho' \in R_{\rho^*}$ implies that $R_{\rho'} = R_{\rho^*}$. We have shown that if ρ' has probability distribution $I\!P$, then so does ρ^*. The

transition matrix for moving each object varies with i; if objects $1, 2, \ldots, n$ are moved in turn and if $I\!P$ is the initial probability distribution of ρ', then $I\!P$ will be the probability distribution after the movement of each object, and therefore $I\!P$ will be the probability distribution after movement of all n objects. Therefore $I\!P$ is a stationary distribution of the Markov chain. Note that the Markov chain has a finite number of states. It is irreducible as it is possible to go from any particular partition to any other partition. Also, as it is possible to be in the same partition after one Markov step, the Markov chain is aperiodic. Therefore, the limiting distribution of the Markov chain is its unique stationary distribution $I\!P$.

5.3 Prior Parameters

To implement Markov sampling we will need to compute the ratios $Q(t)$ in equation (10) and the functions $h_j(\mathbf{Y}) = I\!P(\mu_j|\rho, \mathbf{X})$. Let $v = 1/(1+\sigma_0^2)$.

From equation (9) and letting

$$n_{S_r^*} = \begin{cases} n_{S_r} + 1, & \text{if } r = t \\ n_{S_r}, & \text{otherwise,} \end{cases}$$

and

$$\overline{X}_{S_r^*} = \begin{cases} (n_{S_r}\overline{X}_{S_r} + X_i)/(n_{S_r}+1), & \text{if } r = t \\ \overline{X}_{S_r}, & \text{otherwise,} \end{cases}$$

we obtain

$$Q(t) = \frac{n_{S_t}}{m\, v^{1/2}} \exp\left(\frac{1-v}{2}\left(\sum_{r=1}^{b} n_{S_r^*}\left(\overline{X}_{S_r^*} - \overline{X}\right)^2 - \sum_{r=1}^{b+1} n_{S_r}\left(\overline{X}_{S_r} - \overline{X}\right)^2\right)\right).$$

The posterior density of μ^S has mean $(\sigma_0^2\overline{X}_S + \mu_0)/(\sigma_0^2+1)$. Hence

$$\begin{aligned} I\!P(\mu_j|j \in S \in \rho, \mathbf{X}) &= I\!P(\mu^S|S \in \rho, \mathbf{X}) \\ &= \overline{X}_S + v\,(\mu_0 - \overline{X}_S). \end{aligned}$$

The quantities Q and h depend on the prior parameters - μ_0, $v = 1/(1+\sigma_0^2)$ and m. We estimated μ_0 by $\overline{X}$ and σ_0^2 by $\sum_{i=1}^{n} (X_i - \overline{X})^2$. Note that we do not divide by n or $n-1$: this is because overestimating σ_0^2 when groups are close together is not as serious as underestimating σ_0^2 when groups are further apart. To estimate m, first let $X_{(i)}$, $i = 1, \ldots, n$ be the ordered data. Start with one set i.e. nsets=1. If the interval between $X_{(i)}$ and $X_{(i+1)}$, $i = 1, \ldots, n-1$, is larger than one, form a new set i.e. nsets=nsets+1. We estimate m by $2^{\text{nsets}-2}$.

6. SIMULATIONS

6.1 Other Estimates of Normal Means

The product estimate $(\hat{\mu}_1, \ldots, \hat{\mu}_n)$ is compared to:
(1) the mle $(\hat{\mu}_1^1, \ldots, \hat{\mu}_n^1)$, where

$$\hat{\mu}_i^1 = X_i;$$

(2) an empirical Bayes (EB) estimate $(\hat{\mu}_1^2, \ldots, \hat{\mu}_n^2)$, which is computed as follows. We have $X_i|\mu_i \sim N(\mu_i, 1)$. Let $\mu_i \sim N(\mu_e, \sigma_e^2)$. Then $X_i \sim N(\mu_e, 1+\sigma_e^2)$. We estimate μ_e by $\overline{X}$ and $1+\sigma_e^2$ by $s^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$. Therefore, the hyperparameters μ_e and σ_e^2 are estimated by

$$\hat{\mu}_e = \overline{X}$$

and

$$\hat{\sigma}_e^2 = \begin{cases} s^2 - 1, & \text{if } s^2 > 1 \\ 0, & \text{f } s^2 \leq 1. \end{cases}$$

We then let $\mu_i \sim N(\hat{\mu}_e, \hat{\sigma}_e^2)$ to obtain

$$\mu_i|\mathbf{X} \sim N\left(\overline{X} + \frac{\hat{\sigma}_e^2}{\hat{\sigma}_e^2+1}(X_i - \overline{X}), \frac{\hat{\sigma}_e^2}{\hat{\sigma}_e^2+1}\right).$$

Therefore,

$$\hat{\mu}_i^2 = \begin{cases} \overline{X} + (X_i - \overline{X})\left(1 - \frac{1}{s^2}\right), & \text{if } s^2 > 1 \\ \overline{X}, & \text{if } s^2 \leq 1. \end{cases}$$

(3) a mixture estimate $(\hat{\mu}_1^3, \ldots, \hat{\mu}_n^3)$, where $\hat{\mu}_i^3$ is computed by assuming that $X_i|\mu_i \sim N(\mu_i, 1)$ and $\mu_i \sim G$ where G is arbitrary. The maximum likelihood estimate of G, $\hat{G}$, has less than or equal to n steps. We will assume that $\hat{G}$ has n steps of size $1/n$ at $a_1, \ldots, a_n$. We can represent this problem as an incomplete data problem and use the EM algorithm (Dempster, Laird and Rubin 1977) to find the maximum likelihood estimates of $a_1, \ldots, a_n$. Then

$$\begin{aligned} \hat{\mu}_i^3 = I\!P_{\hat{G}}(\mu_i|X_i) &= \frac{\int y\,(2\pi)^{-1/2} \exp(-(X_i - y)^2/2)\, d\hat{G}(y)}{\int (2\pi)^{-1/2} \exp(-(X_i - y)^2/2)\, d\hat{G}(y)} \\ &= \frac{\sum_{j=1}^n \frac{1}{n} \hat{a}_j\,(2\pi)^{-1/2} \exp(-(X_i - \hat{a}_j)^2/2)}{\sum_{j=1}^n \frac{1}{n} (2\pi)^{-1/2} \exp(-(X_i - \hat{a}_j)^2/2)} \end{aligned}$$

6.2 Results

In Table 1, we compare the product estimate to the mle, the EB estimate and the mixture estimate using the average difference in estimated mean square error. There are 20 observations and 50 samples for each set of means $\boldsymbol{\mu}$. We use, for example, $19^0\,1^3$ to represent $\mu_1 = \ldots = \mu_{19} = 0, \mu_{20} = 3$. In the Markov sampling, we found that $N = 1000$ samples was sufficient for convergence of the product estimate. We also did the exact calculation to compute the product estimate for a few different $\boldsymbol{\mu}$'s, when $n = 10$. We compared the estimate obtained using Markov sampling to the exact estimate and found that they were very close. We deleted the first 100 samples to remove any effect of the starting value ρ_1 (this was probably more than was necessary).

We find that the mle is better than the product estimate only when there are a few groups two or three SD's apart (22,25). For all the other $\boldsymbol{\mu}$'s considered, the product estimate is superior.

The EB estimate is better than the product estimate when there are three or more similar sized groups one, two or three SD's apart (16–18,20–22,24,25) and for two groups two SD's apart (8). It does not detect small groups of outliers; see 2–6 and 12–15. It doesn't perform well if the groups are more than three SD's apart due to the fact that it shrinks towards the mean of the data.

The mixture estimate is superior to the product estimate for less than six groups three or four SD's apart (5,9,15,18,19,23). It does well when there are outliers four SD's from the rest of the data (5) but not so well for outliers 10 SD's away (6). In general, it doesn't do well for groups less than four SD's apart.

The product estimate is superior to the other three estimates for one group (1), for small groups of outliers less than four SD's or very far for the rest of the data (2–4,6), for two equal sized groups one SD or more than four SD's apart (7,10,11) and for three groups, where two of the groups are small, less than four SD's apart (12–14). To summarise, the product estimate is usually only worse than one of the mixture estimate (which works best for groups more than three SD's but not too far apart) and the EB estimate (which works best for groups three or less SD's apart) and is often superior to both (and almost always superior to the mle).

7. PREDICTING BATTING AVERAGES

We compute the product estimate and the other estimates considered in

Section 6 for the batting averages data of George (1986b). He looks at batting averages of the 26 major league baseball teams for the 1984 season and uses the averages after 300 at bats, $(b^1, \ldots, b^{26})$, to predict the averages for the remainder of the season, $(p^1, \ldots, p^{26})$. Assuming that the number of team hits has a binomial distribution, that is, $300\, b^i \sim \text{Bin}(300, p^i)$, he uses the variance stabilizing transformation $f(b) = 300^{1/2} \arcsin(2b - 1)$ to obtain normality. Explicitly, letting $X_i = f(b^i)$ and $\mu_i = f(p^i)$, the asymptotic normality of b^i and the continuity of f then gives $X_i|\mu_i \sim N(\mu_i, 1)$.

The product estimate, the EB estimate and the mixture estimate are re-transformed to get estimates of $p^i = f^{-1}(\mu_i)$. The batting averages and re-transformed estimates are given in Figure 1. The mle estimate is very variable following no particular pattern. The EB estimate is almost constant for all of the teams. The product and mixture estimates do not vary much but are influenced slightly more by the data i.e. the batting averages after 300 at bats. Very large or small data values tend to lead to larger or smaller estimates.

We also computed the squared error losses for the different estimates and compared these to each other and to the estimates given in George (1986b). The loss for the mle (26.64) is very large. The losses for the product (4.30), EB (4.28) and mixture (4.33) estimates are very similar. The product estimate does almost as well as the best of George's general Stein estimates (4.24) and has smaller loss than any of his multiple shrinkage estimates (5.37, 4.73, 4.43).

REFERENCES

Barry, Daniel and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *The Annals of Statistics*, 20, 260–279.

——— (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association,* 88, 309–319.

Crowley, E. M. (1992), "Estimation of Clustered Parameters," unpublished Ph.D. dissertation, Yale University, Dept. of Statistics.

——— (1993), "Estimation of Clustered Parameters," Technical Report 93–36.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Efron, B. and Morris, C. (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case," *Journal of the American Statistical Association,* 67, 130–139.

——— (1973), "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach," *Journal of the American Statistical Association,* 68, 117–130.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalisations," *Journal of the American Statistical Association,* 70, 311–319.

Escobar, M.D. (1992), "Estimating Normal Means With a Dirichlet Process Prior," Technical Report #512, Carnegie Mellon University, Dept. of Statistics.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Gelman, A and Rubin, D. B. (1992), "Inference From Iterative Simulation

Using Multiple Sequences," *Statistical Science,* 7, 457–511.

George E. I. (1986a), "Minimax Multiple Shrinkage Estimation," *The Annals of Statistics,* 14, 188–205.

——— (1986b), "Combining Minimax Shrinkage Estimators," *Journal of the American Statistical Association,* 81, 437–445.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science,* 7, 473–511.

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics, Part A – Theory and Methods,* 19, 2745–2756.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* 87, 97–109.

James, W. and Stein, C. (1961), "Estimation With Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* (vol. 1), Berkeley, CA: University of California Press, pp. 361–379.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association,* 73, 805–811.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087–1092.

Stein, C. (1956), "Inadmissibility of the Usual Estimate for the Mean of a Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* (vol. 1), Berkeley, CA:

University of California Press, pp. 197–206.

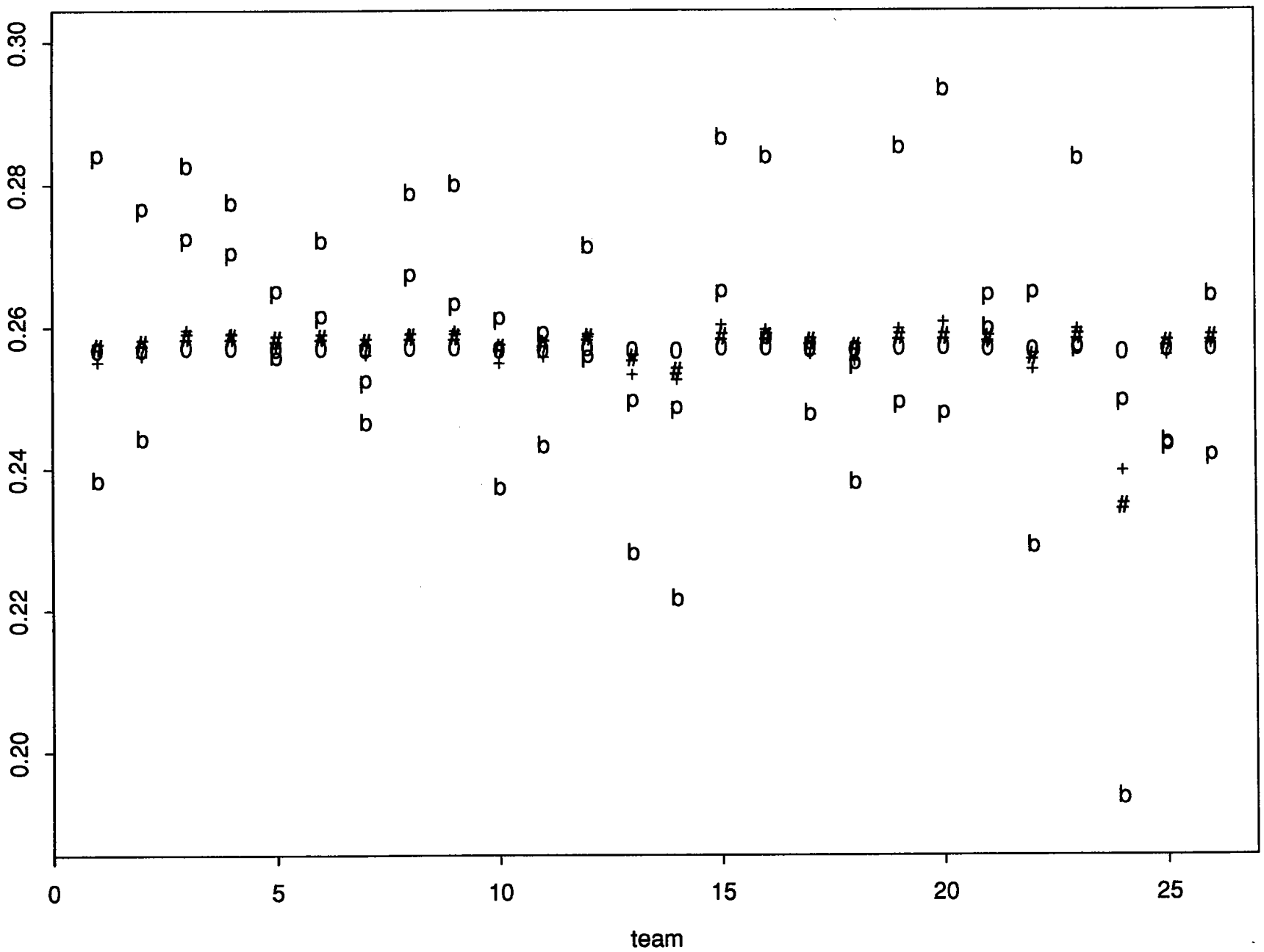
Table 1: Comparison of the Product Estimate to the Other Estimates.

		Average Difference in Estimated Mean Square Error		
	μ	Mle	EB	Mixture
1.	20^0	−18.2 (.7)	0.0 (.1)	−0.6 (.2)
2.	$19^0\,1^1$	−17.4 (.7)	−0.1 (.1)	−0.7 (.2)
3.	$19^0\,1^2$	−15.5 (.7)	−0.6 (.1)	−0.8 (.2)
4.	$19^0\,1^3$	−14.0 (.7)	−2.2 (.2)	−0.6 (.2)
5.	$19^0\,1^4$	−14.6 (.7)	−5.4 (.3)	0.5 (.3)
6.	$19^0\,1^{10}$	−17.5 (.7)	−14.6 (.5)	−0.6 (.2)
7.	$10^0\,10^1$	−14.0 (.7)	−0.1 (.1)	−1.1 (.2)
8.	$10^0\,10^2$	−6.7 (.9)	0.8 (.3)	−0.5 (.3)
9.	$10^0\,10^3$	−6.1 (1.0)	−1.9 (.5)	0.1 (.4)
10.	$10^0\,10^4$	−10.7 (.8)	−8.0 (.4)	−0.2 (.3)
11.	$10^0\,10^{10}$	−16.8 (.6)	−16.3 (.5)	−1.3 (.2)
12.	$14^0\,3^1\,3^2$	−10.7 (.8)	−0.1 (.2)	−0.8 (.2)
13.	$14^0\,3^2\,3^4$	−7.4 (.7)	−2.5 (.3)	−0.9 (.3)
14.	$14^0\,3^3\,3^6$	−6.3 (.7)	−3.6 (.5)	−0.4 (.3)
15.	$14^0\,3^4\,3^8$	−8.4 (.5)	−6.7 (.5)	1.0 (.4)
16.	$6^0\,7^1\,7^2$	−9.1 (.8)	0.6 (.2)	−0.7 (.3)
17.	$6^0\,7^2\,7^4$	−2.5 (.8)	1.3 (.4)	−1.1 (.3)
18.	$6^0\,7^3\,7^6$	−0.4 (.8)	1.6 (.5)	0.7 (.5)
19.	$6^0\,7^4\,7^8$	−5.5 (.7)	−4.3 (.5)	1.1 (.4)
20.	$5^0\,5^1\,5^2\,5^3$	−5.2 (.9)	1.4 (.3)	−1.1 (.3)
21.	$5^0\,5^2\,5^4\,5^6$	−0.4 (.7)	2.0 (.4)	−1.1 (.4)
22.	$5^0\,5^3\,5^6\,5^9$	0.7 (.5)	1.9 (.4)	−1.4 (.5)
23.	$5^0\,5^4\,5^8\,5^{12}$	−2.0 (.4)	−1.2 (.4)	2.3 (.6)
24.	$3^0\,3^1\,3^2\,3^3\,3^4\,3^5\,2^6$	−0.7 (.7)	2.5 (.5)	−1.9 (.4)
25.	$3^0\,3^2\,3^4\,3^6\,3^8\,3^{10}\,2^{12}$	1.8 (.5)	2.8 (.4)	−3.3 (.4)

NOTE: The difference is between the product estimate and the other estimate. Values in parentheses are the standard errors of the difference in estimated mean square error.

Figure Captions

Figure 1. Batting Averages and Estimates. b, average after 300 at bats = mle; p, average for remainder of season; +, product estimate; 0, EB estimate; #, mixture estimate.


0.30
0.28
0.26
0.24
0.22
0.20
0
5
10
15
20
25
team