Ann. Inst. Statist. Math.
36 (1984), Part A, 11-21

SOME DATA-ANALYTIC MODIFICATIONS TO BAYES-STEIN ESTIMATION

TOM LEONARD

(Received Apr. 10, 1982; revised Nov. 15, 1982)

Summary

The usual Bayes-Stein shrinkages of maximum likelihood estimates towards a common value may be refined by taking fuller account of the locations of the individual observations. Under a Bayesian formulation, the types of shrinkages depend critically upon the nature of the common distribution assumed for the parameters at the second stage of the prior model. In the present paper this distribution is estimated empirically from the data, permitting the data to determine the nature of the shrinkages. For example, when the observations are located in two or more clearly distinct groups, the maximum likelihood estimates are roughly speaking constrained towards common values within each group. The method also detects outliers; an extreme observation will either be regarded as an outlier and not substantially adjusted towards the other observations, or it will be rejected as an outlier, in which case a more radical adjustment takes place. The method is appropriate for a wide range of sampling distributions and may also be viewed as an alternative to standard multiple comparisons, cluster analysis, and nonparametric kernel methods.

1. Introduction

Consider observations $x_1, \cdots, x_m$ which are independent, given respective parameters $\theta_1, \cdots, \theta_m$ and where x_i possesses density, or probability mass function $f_i(x_i; \theta_i)$ for $x_i \in \mathcal{X}$ and $\theta_i \in \Theta$, for $i=1, \cdots, m$. Suppose further that the θ_i are a priori *exchangeable* and that they possess the prior probability structure of a random sample from a distribution with density $g(\theta_i)$.

Most Bayesian simultaneous estimation methods (e.g. Leonard [8],

AMS (MOS) subject classifications: 62G05, 62H15, 62J07.
Key words: James-Stein, Bayes, shrinkage, estimator, outlier, multiple comparions, nonparametric, kernel estimator.

Lindley and Smith [10], and Clevenson and Zidek [1], for binomial, normal, and Poisson situations) take the density g to belong to a parametrized family, and then introduce second stage distributional assumptions about the parameters of g. The choice of g very often involves a unimodal density with thin tails (e.g. normal or Gamma). These choices typically lead to posterior estimates of the θ_i which shrink the x_i towards a common value (e.g. zero, the prior mean, or the average observation) thus providing Bayesian analogues of frequentist procedures (e.g. James and Stein [6], and Efron and Morris [3]).

Whilst the previous choices of prior will be adequate in numerous situations, shrinkages towards a common value may be less appropriate in cases where g does not assume such an idealized form. For example, Dawid [2] investigates prior densities with thicker tails than the normal and shows that it is then unreasonable to shrink in extreme observations as radically as suggested by an analysis based upon a normal prior. Alternatively, g might possess more than one mode in which case fairly complex shrinkages might be involved.

In the present paper we relax previous assumptions involving thin-tailed unimodal densities and indeed proceed to the other extreme by supposing that the statistician possesses absolutely no prior information about the density g. Our motivation is to investigate the shrinkages which are actually suggested by the data, rather than imposed by particular functional forms assumed for g. If there were some partial information about g then this could be introduced via the method proposed by Leonard [9] for smoothing densities; this aspect will not however be considered in this paper.

We will explore the consequences of estimating g empirically from the data. Readily computable estimates will be obtained which avoid problems of specifying the tail-behaviour, modality, and general shape of g.

Laird [7] and Lindsay [11], [12], [13] investigate the theoretical properties of the maximum likelihood estimate of g, obtained by maximizing the log-likelihood functional

$$L(g)=\sum_{i=1}^{m}\log\int_{\theta} f_i(x_i;\,\theta)g(\theta)d\theta\ . \tag{1.1}$$

Lindsay [11] shows, under general conditions, that the maximum likelihood estimate of g is a discrete mixture of Kronecker-delta functions of the form

$$g^*(\theta)=\sum_{j=1}^{p}\phi_j\delta_{b_j}(\theta) \tag{1.2}$$

where $\sum\phi_j=1$, with $p\leqq m$, and $\delta_{b_j}(\theta)$ denotes the Kronecker-delta function at $\theta=b_j$. Laird proposes a fairly complex scheme based on the EM algorithm for estimating $\phi_1,\cdots,\phi_p$ conditional upon a specified p. The optimal p may then be ascertained by comparing the log-likelihoods in (1.1) for different p. This iterative scheme will definitely converge, due to general properties of the EM algorithm. There is, however, no guarantee that convergence will be quick. Indeed, when these are a large number of terms in the mixture, or when the specified p is in contradiction to the values suggested by the data, the iterations could become quite tedious. It is moreover necessary to complete the iterations for each value of p.

In the next section a computational shortcut is described which will be appropriate whenever the optimal p is small compared with m. This shortcut will avoid the possibly tedious iterations on the mixing probabilities, and will also estimate the optimal p during a single set of iterations on some location parameters. The numerical solution will provide the maximum likelihood estimates of p and $\phi_1,\cdots,\phi_p$ but when $\phi_1,\cdots,\phi_p$ are constrained to be integer multiples of m^{-1}. This restriction on the parameter space leads in practice in much more rapid convergence of the maximum likelihood iterations. The general idea is to replace (1.2) by an equiprobable mixture, with m possibly different locations, and then to estimate these locations by maximum likelihood. The estimated locations will in practice cluster into several subsets, with equal estimates within each subset. The number of such subsets will then estimate p and the proportions of estimates in the various subsets will estimate $\phi_1,\cdots,\phi_p$. This idea provides an alternative to a large literature of procedures following the (non-parametric) empirical Bayes philosophy. Previous work is well-catalogued by Laird and includes the pioneering work of H. Robbins, most importantly Robbins [15].

2. The empirical estimation of the prior density

Consider the limiting situation where the sampling variation in each of the $f_i(x_i\,|\,\theta_i)$ distributions approach zero, so that the θ_i become effectively known and equal to their maximum likelihood estimates $\hat{\theta}_i$. In this limiting case the maximum likelihood estimate of $g(\theta)$ is

$$\tilde{g}(\theta)=m^{-1}\sum_{i=1}^{m}\delta_{\hat{\theta}_i}(\theta)=m^{-1}\sum_{i=1}^{m}\delta_{\theta_i}(\theta) \qquad (\theta\in\Theta)\,. \tag{2.1}$$

This motivates us to consider, in general, estimates for g which take the form

$$\hat{g}(\theta)=m^{-1}\sum_{i=1}^{m}\delta_{a_i}(\theta) \qquad (\theta\in\Theta)\,, \tag{2.2}$$

but where $a_1, \cdots, a_m$ are now arbitrary points to be estimated from the data. We anticipate that, when the first-stage sampling variation is reintroduced, this will cause the a_i to adjust the $\hat{\theta}_i$ by reducing their overall spread, and hence cause a sort of Stein-effect on the $\hat{\theta}_i$. Substituting the function in (2.2) for g in (1.1) provides us with the log-likelihood of $a_1, \cdots, a_m$, which is given by

$$L(a)=\sum_{i=1}^{m} \log \sum_{k=1}^{m} f_i(x_i, a_k)-m \log m . \tag{2.3}$$

The a_i will be estimated by maximizing the function in (2.3). The optimizing values could be interpreted as hypothetical observations from the distribution g roughly speaking equal in information content about g to the information about g contained in the log-likelihood functional (1.1). Note that in all the numerical examples we have considered, the optimal values for $a_1, \cdots, a_m$ will become concentrated at a smaller number of estimated points, say $b_1, \cdots, b_p$. The prior probability ϕ_j attached to point b_j should then be estimated by

$$g(b_j)=\#(a_i\,;\, a_i=b_j)/m \qquad (j=1, \cdots, p) . \tag{2.4}$$

This yields a discrete distribution, of the form described in (1.2), which assigns estimated probabilities to p estimated points, where p is also obtained empirically. We anticipate that it will often be close in numerical terms to the unrestricted maximum likelihood estimate proposed by Laird. Differentiating the function in (2.3) with respect to a_l gives us, after some rearrangement

$$\frac{\partial L}{\partial a_l}=\sum_{i=1}^{m} P_{il} \frac{\partial \log f_i(x_i\,;\, a_l)}{\partial a_l} \qquad (l=1, \cdots, m) \tag{2.5}$$

where

$$P_{il}=A_{il}\Big/\sum_{k=1}^{m} A_{ik} \tag{2.6}$$

with

$$A_{il}=f_i(x_i\,;\, a_l) . \tag{2.7}$$

Note that, when $a_1, \cdots, a_m$ are unequal, the expression in (2.6) is just the posterior probability that $\theta_i=a_l$, under the prior distribution in (2.1). Therefore, solving the maximum likelihood equations for the a_l also gives us empirical estimates for the entire posterior distribution for each θ_i for $i=1, \cdots, m$; so that posterior estimates may also be obtained for the θ_i. Equating the derivatives in (2.3) to zero yields a set of equations which may in general be solved by any standard iterative procedure (e.g. Newton-Raphson). However, the computations turn out to be particularly simple in a variety of special cases.

(a) *Exponential family of sampling distributions*

When the sampling densities f_i assume the forms

$$f_i(x_i;\theta_i)=\exp\{B(\theta_i)+t(x_i)C(\theta_i)+D(x_i)\} \tag{2.8}$$

for appropriate choices of the functions B, C, D, and t, then the maximum likelihood equations for the a_l are

$$\frac{-B^{(1)}(a_l)}{C^{(1)}(a_l)}=\sum_{i=1}^m t(x_i)P_{il}\Big/\sum_{i=1}^m P_{il} \qquad (l=1,\cdots,m) \tag{2.9}$$

where the P_{il} are defined in (2.6). Equations (2.9) may be solved by substituting trial values (initially the values $\hat{\theta}_l$) for the a_l in the right-hand sides, transforming the left-hand sides into fresh values for the a_l and then cycling until convergence. For example, when the x_i possess Poisson distributions with respective means θ_i, we have,

$$a_l=\sum_{i=1}^m x_iP_{il}\Big/\sum_{i=1}^m P_{il} \tag{2.10}$$

demonstrating that each a_l takes the form of a weighted average of $x_1,\cdots,x_m$. This provides an alternative to the procedure described by Simar [17] for mixtures of Poisson distributions. The iterations for for $a_1,\cdots,a_m$ described in this section could also be justified via the EM algorithm, under the constraint in (2.2), by regarding $x_1,\cdots,x_m$ as incomplete data and $\theta_1,\cdots,\theta_m$ as missing values. Therefore convergence is guaranteed. Since no iterations are required on the mixing probabilities, convergence is usually very rapid.

(b) *Binomial distributions with unequal sample size*

If the x_i are independent and possess binomial distributions, given the corresponding probabilities θ_i and sample sizes n_i then the maximum likelihood equations for the a_l are given by

$$a_l=\sum_{i=1}^m x_iP_{il}\Big/\sum_{i=1}^m n_iP_{il} \tag{2.11}$$

where we may take the A_{il} in the expression for P_{il} in (2.6) to satisfy

$$A_{il}=a_l^{x_i}(1-a_l)^{n_i-x_i} \tag{2.12}$$

since the functional contributions to the sampling distribution cancel themselves out. Note that $-2\log A_{il}$ takes the form of a distance measure between x_i/n_i and a_l. Hence a_l in (2.11) will depend more heavily upon those x_i/n_i nearby then on outlying x_i/n_i. This creates a mechanism enabling $a_1, \cdots, a_m$ to take full account of the random variability in $x_1, \cdots, x_m$.

(c) *Normal observations with unknown variance*

Suppose now that for $i=1, \cdots, m$ and $j=1, \cdots, n_i$; the observations x_{ij} are independent and normally distributed with respective group means θ_i and common variance σ^2. Then σ^2 may be estimated jointly with the prior values a_l by solving the joint maximum likelihood equations

$$a_l = \sum_{i=1}^{m} n_i \bar{x}_i P_{il} \Big/ \sum_{i=1}^{m} n_i P_{il} \qquad (l=1, \cdots, m) \tag{2.13}$$

and

$$\sigma^2 = N^{-1} S_w^2 + N^{-1} \sum_{i=1}^{m} n_i \sum_{k=1}^{m} (\bar{x}_i - a_k)^2 P_{ik} \tag{2.14}$$

where

$$N = \sum_{i=1}^{m} n_i , \qquad \bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij} , \qquad S_w^2 = \sum_{ij} (x_{ij} - \bar{x}_i)^2 ,$$

and the P_{il} are defined in (2.6), with

$$A_{il} = \exp \left\{ -\frac{1}{2} n_i \sigma^{-2} (\bar{x}_i - a_k)^2 \right\} . \tag{2.15}$$

Equations (2.13) and (2.14) may be solved by combining the iterations recommended in (a), for fixed σ^2, with simple cyclic substitutions on σ^2. The above procedure may be employed in either the Model I or Model II ANOVA situations since our assumptions relate either to an exchangeability model for fixed effects, or a random effects model. Note that the classical F-test for equality of the means may be replaced by an inspection as to whether or not all the estimated a_l are equal; t-tests for individual differences may be avoided by comparing the posterior means discussed in the next section.

3. Posterior estimation of the sampling parameters

Once the iterations have been completed for the a_l and P_{il}, the parameters $\theta_1, \cdots, \theta_m$ may be estimated (e.g. by their empirical posterior means)

$$\tilde{\theta}_k = \sum_{l=1}^{m} a_l P_{kl} \qquad (k=1, \cdots, m) . \tag{3.1}$$

For example, in the normal situation in section (2c) we have

$$\tilde{\theta}_k = \sum_{i=1}^{m} n_i \bar{x}_i \sum_{l=1}^{m} P_{kl} P_{il} \Big/ \sum_{i=1}^{m} n_i P_{il} \tag{3.2}$$

which can be arranged in the form of a weighted average of $\bar{x}_1, \cdots, \bar{x}_m$. Again, as $-2 \log A_{il}$, from (2.15), is a distance measure between $\bar{x}_i$ and a_k, the posterior mean in (3.2) will take more account of $\bar{x}_i$'s which are close to $\bar{x}_l$ rather than those which are some distance away. We suggest that (3.2) will in many practical situations be preferable to the James-Stein estimator, as far as meaningful statistical interpretations are concerned since it does not shrink all the $\bar{x}_i$ irrevocably towards a common value without taking into account the statistical scatter of the data.

4. Numerical examples

The data in Table 1 relate to the males and females on 10 different courses, and were previously analyzed by Leonard [8] using a Bayes-Stein estimation technique for binomial data.

Table 1. Classification of students according to sex and course

Course	Female	Male	% of Females	Bayes-Stein	Empirical
1	42	47	47.2	44.4	44.0
2	32	40	44.4	41.6	44.0
3	45	57	44.1	42.1	44.0
4	10	16	38.5	34.5	43.2
5	7	20	25.9	26.7	21.1
6	3	12	20.0	24.1	18.2
7	3	13	18.8	23.6	17.3
8	5	22	18.5	22.3	15.7
9	12	72	14.3	16.9	15.7
10	11	84	11.6	14.5	15.3

The rows of the table were not originally arranged according to the values of the percentages; the present ordering is intended simply for ease of presentation. The Bayes-Stein estimates in the fifth column shrink each observed proportion towards an average value of 28.0 The amounts of shrinkage vary according to sample size and according to distance from the average value when measured on logistic scale. Application of our empirical method in Section 2b yielded an estimated common prior distribution for the binomial probabilities. This assigned prior probabilities 4/10 and 6/10 to the values 0.440 and 0.153. We see from the last column of Table 1 that our empirical procedure has discerned that the observed percentages lie in two clearly distinct groups. It has moreover decided that the fourth percentage lies in the first group, and therefore pulls the 38.5 value right up to 43.2, in the opposite direction than the radical shrinkage to 34.5 which was suggested by James-Stein. The first three percentages are regarded as equal with the fourth percentage just a small distance away. The second group of six percentages causes shrinkages for the first five which are all opposite in direction to that suggested by Bayes-Stein. Percentage number 5 is slightly unwilling to join the group, because of possible inclinations to either join the first group or to stay on its own. Overall the differences from James-Stein are quite remarkable.

We also reanalyzed the famous baseball batting example introduced by Efron and Morris [5]. Again, the common prior distribution was estimated by a two-point discrete distribution, but this time the two points were close enough together to retain Bayes-Stein type shrinkages towards a common value. Interestingly our posterior means were virtually identical to the estimates proposed by Efron and Moris even though the latter were based upon very different (parametric) assumptions. Therefore our estimates seem to agree with Bayes-Stein when the scatter of the data is well-enough behaved to justify these simple shrinkages.

The data in Table 2 comprise a subset of a well-known 14×14 contingency table introduced by Karl Pearson [14]. The entries in the fourth column give the proportions of sons who follow their father's occupation, for each of fourteen occupations; the categories have again

Table 2. Proportions of sons following their father's occupation

Occupation (i)	x_i	n_i	Observed Proportion	Smoothed Proportion
1	0	26	0.000	0.020
2	6	88	0.068	0.103
3	11	106	0.104	0.103
4	7	54	0.130	0.115
5	6	44	0.137	0.127
6	4	19	0.211	0.221
7	18	69	0.261	0.257
8	9	32	0.281	0.270
9	6	18	0.333	0.334
10	23	51	0.451	0.477
11	54	115	0.470	0.480
12	20	41	0.488	0.480
13	28	50	0.560	0.480
14	51	62	0.823	0.823

been rearranged into a suitable order. In this case our empirical prior distribution assigned respective probabilities 1/14, 4/14, 4/14, 4/14 and 1/14 to the points 0.020, 0.103, 0.257, 0.480, and 0.823, representing a number of interesting features in the scatter of the data. The corresponding posterior means we described in the fifth column of the table. The first two groups illustrate that our method can be used to decide whether or not particular observations are outliers. The second proportion (0.068) has been pulled back into the main group, whilst the first proportion (0.000) has been left virtually alone. Similarly the 14th proportion (0.823) is left alone by the fifth group whilst the ninth proportion is of interest as an internal outlier isolating itself between the third and fifth groups.

Our method provides a type of cluster analysis since it groups the observations into definite clusters. Also, the method seems to be robust under deviations from the assumption of exchangeability of $\theta_1, \cdots, \theta_m$. If there is strong evidence in the data to refute exchangeability for a particular parameter then the latter is simply estimated as an outlier without radically effecting the other estimates. Indeed, our method effectively splits the parameters up into exchangeable subsets thus providing an alternative to the Efron and Morris [4] procedure for deciding whether to combine possibly related estimation problems. Finally, our method could be viewed as an alternative to standard techniques for multiple comparisons since it smooths the data to a form where it is easy to compare subsets of the parameters.

5. Relationship with nonparametric kernel methods

Suppose, for simplicity, that $f_i(x_i; \theta_i)$ belongs to the symmetric location family

$$f_i(x_i; \theta_i) = f(|x_i - \theta_i|) . \tag{5.1}$$

Then our method estimates the marginal density

$$\xi(x) = \int_\theta f(|x-\theta|) g(\theta) d\theta \tag{5.2}$$

by

$$\hat{\xi}(x) = m^{-1} \sum_{i=1}^{m} f(|x - a_i|) \qquad (x \in X) \tag{5.3}$$

where the a_i are calculated via our computational procedure. We see that (5.3) could also be used as an estimate for the density $\xi(\cdot)$ under the assumption that the sampling (rather than marginal) density of $x_1, \cdots, x_m$ is equal to $\xi(x)$. These are close similarities with nonparametric kernel estimators of the form

$$\xi^*(x)=m^{-1}\sum_{i=1}^{m} f(|x-x_i|)\ . \tag{5.4}$$

These are prevalent in the literature; see Silverman [16] for some recent developments. The estimate ξ^* averages the kernels $f(|x-x_i|)$ centered on the data points, rather than centered on $a_1,\cdots,a_m$, as in (5.3).

Kernel estimators are open to criticism on the following grounds

(i) They tend to lead to estimators which are too "flat". The variance corresponding to $\xi^*(x)$ is theoretically always larger than the sample variance of the observations.

(ii) When an equal kernel is placed over each data point, then, according to its spread, the estimator very often tends to be either too flat, or too bumpy in the details.

(iii) When, say, f is a normal density with mean zero and variance σ^2, the value σ^{-1} is referred to as the "band width" and regulates the degree of smoothing. It is notoriously difficult to obtain a reasonable analytic method for estimating σ^2 from the data.

Our procedure promises to answer all three criticisms. Firstly, as the a_i are more compressed than the x_i the estimator ξ in (5.3) will always be less flat. Secondly, by estimating the a_i according to the scatter of the data it will avoid many of the problems in (iii). Thirdly, when f is a normal (or other symmetric) density with scale parameter σ^2 we may estimate σ^2 as well. In the normal case we may use equations (2.12)–(2.14) with single replications $n_i=1$, when the equations still possess enough structure to sensibly estimate σ^2.

The kernel ideas will be pursued in greater detail elsewhere.

Acknowledgements

The author wishes to thank the referee for his helpful comments.

UNIVERSITY OF WISCONSIN-MADISON

REFERENCES

[1] Clevenson, M. and Zidek, J. W. (1975). Simultaneous estimation of the means of independent Poisson laws, *J. Amer. Statist. Ass.*, **70**, 698-705.

[2] Dawid, A. P. (1973). Posterior expectation for large observations, *Biometrika*, **61**, 664-667.

[3] Efron, B. and Morris, C. (1973a). Stein's estimation rule and its competitors—an empirical Bayes approach, *J. Amer. Statist. Ass.*, **68**, 117-130.

[4] Efron, B. and Morris, C. (1973b). Combining possibly related estimation problems (with discussion), *J. R. Statist. Soc.*, B, **36**, 379-421.

[5] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Ass.*, **70**, 311-319.

[6] James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proc. 4th Berkeley Symposium*, **1**, 361-379.

[7] Laird, N. M. (1978). Non-parametric maximum likelihood estimation of a mixing distribution, *J. Amer. Statist. Ass.*, **73**, 805-811.

[8] Leonard, T. (1972). Bayesian methods for binomial data, *Biometrika*, **59**, 581-589.

[9] Leonard, T. (1978). Density estimation, stochastic processes, and prior information (with discussion), *J. R. Statist. Soc.*, B, **40**, 113-146.

[10] Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion), *J. R. Statist. Soc.*, B, **34**, 1-41.

[11] Lindsay, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution, in *Statistical Distributions in Scientific Work* (ed. C. Taillie et. al.), Vol. 5, 95-109.

[12] Lindsay, B. G. (1982a). A geometry of mixture likelihoods: A general theory, *Penn. State Univ. Tech. Report.*

[13] Lindsay, B. G. (1983a). A geometry of mixture likelihoods, Part II: The exponential family, *J. Amer. Statist. Ass.*, **4**, 1200-1209.

[14] Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation, *Drapers Co. Res. Mem. Biometrics Series.*

[15] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems, *Ann. Math. Statist.*, **35**, 1289-1302.

[16] Silverman, B. W. (1978). Choosing the window width when estimating a density, *Biometrika*, **65**, 1-12.

[17] Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process, *Ann. Statist.*, **4**, 1200-1209.