On Empirical Bayes Procedures for Predicting Simple Exponential Survival

by

Seymour Geisser*

University of Minnesota
Technical Report No. 527
February 1989

* This work was supported by NIGMS grant 25271

On Empirical Bayes Procedures for Predicting Simple Exponential Survival

Seymour Geisser

University of Minnesota

1. Introduction

In hierarchical Bayesian analyses where the prior distribution of a set of parameters say θ depends on a set of hyperparameters τ we are typically faced with the problem of dealing with τ in some manner. There are generally three possibilities to be considered: (1) assume a value for τ, (2) assume a distribution for τ depending on known values which merely shifts the burden one step backwards or (3) provide some estimate of τ from the data at hand. Although the latter procedure is technically incoherent, it may be close enough to coherence on the one hand and eminently sensible on the other. The usual procedure when taking the third course is to calculate the marginal density of the observations given the hyperparameter set τ and then estimating τ from this density by either maximum likelihood or the method of moments. Focusing on prediction, we examine a simple case where the method of maximum likelihood is inadequate to the task. Here the method of moments, while reasonable, still has a particular deficiency. In the light of this we propose a sample reuse method which appears to be superior to the other methods in the case under scrutiny and propose its use in other cases.

2. Prediction From the Exponential Distribution

Let $X_1,\ldots,X_N, X_{N+1}$ be a random sample from

$$f(x|\theta) = \theta e^{-\theta x} \qquad \theta>0 \quad x>0.$$

Suppose we assume a prior density for θ

$$p(\theta|\gamma,\delta) \propto \theta^{\delta-1}e^{-\gamma\theta} \qquad \gamma>0, \quad \delta>0, \tag{1}$$

then for $x^{(N)} = (x_1,\ldots,x_N)$

$$p(\theta|\gamma,\delta,x^{(N)}) \propto \theta^{N+\delta-1}e^{-\theta(\gamma+Nx)} \tag{2}$$

where $N\bar{x} = \sum_1^N x_i$. The predictive density of X_{N+1} then is

$$f(x_{N+1}|\gamma,\delta,x^{(N)}) = \frac{(N+\delta)(N\bar{x}+\gamma)^{N+\delta}}{(N\bar{x}+\gamma+x_{N+1})^{N+\delta+1}}, \tag{3}$$

and

$$E(X_{N+1}|x^{(N)},\delta,\gamma) = \frac{\gamma+N\bar{x}}{N+\delta-1}. \tag{4}$$

which we will make use of subsequently.

Now suppose the values γ and δ are not readily assessable subjectively or otherwise. In such cases a convenient method for using the data, that will

enable the approximation of either the posterior or predictive density is to use data based estimates for γ and δ so that

$$\hat{\varphi}(\theta|x^{(N)}) \propto \theta^{N+\hat{\delta}-1} e^{-\theta(\hat{\gamma}+N\bar{x})}$$

and

$$\hat{f}(x_{N+1}|x^{(N)}) = \frac{(N+\hat{\delta})(N\bar{x}+\hat{\gamma})^{N+\hat{\delta}}}{(N\bar{x}+\hat{\gamma}+x_{N+1})^{N+\hat{\delta}+1}} \quad . \tag{5}$$

This is accomplished by calculating the marginal density

$$f(x_{(N)}|\delta,\gamma) = \int f(x^{(N)}|\theta)p(\theta|\delta,\gamma)d\theta$$

and using this likelihood to obtain the estimates of δ and γ that are to be inserted in $f(x_{N+1}|x^{(N)}\delta,\gamma)$. This is sometimes referred to as an empirical Bayes procedure. Since this can be confused with the originally termed and somewhat different empirical Bayes procedure of Robbins (1956), it would be more appropriate to term this as an hyperparameter estimative Bayes procedure to indicate just where the approximation occurs. This is in the spirit of Good (1965) or the parametric empirical Bayes approach outlined in Cox and Hinkley (1974). In our case

$$f(x^{(N)}|\delta,\gamma) = \frac{\Gamma(N+\delta)\gamma^{\delta}}{\Gamma(\delta)[N\bar{x}+\gamma]^{N+\delta}} \quad . \tag{6}$$

3. Estimation of the Hyperparameters

The usual two approaches available for estimating γ and δ are maximum likelihood and the method of moments. We shall obtain the values for γ and δ by these two methods discuss their shortcomings and propose another method.

Now for maximum likelihood estimation, we find from (6) that

$$\log f = \log \frac{\Gamma(N+\delta)}{\Gamma(\delta)} + \delta \log \gamma - (N\bar{x}+\gamma)$$

and

$$\frac{d \log f}{d\delta} = \sum_{j=1}^{N} \frac{1}{N+\delta-j} + \log \gamma - \log(N\bar{x}+\gamma) = 0 \qquad (7)$$

$$\frac{d \log f}{d\delta} = \frac{\delta}{\gamma} - \frac{N+\delta}{N\bar{x}+\gamma} = 0. \qquad (8)$$

From (8), $\gamma = \delta\bar{x}$, which when substituted into (7) yields

$$\sum_{j=1}^{N} \frac{1}{N+\delta-j} = \log \frac{N+\delta}{\delta} . \qquad (9)$$

Clearly the solution for (9) is $\delta \to \infty$ and then $\hat{\gamma} \to \infty$ since $\bar{x} > 0$ with probability 1. It is then clear $\hat{\gamma} \to \hat{\delta}\bar{x}$ and $\hat{\delta} \to \infty$ maximizes (6). Hence as $\delta \to \infty$

$$\Pr(\theta = \bar{x}^{-1}) \to 1$$

and

$$f(X_{N+1}|x^{(N)}) \to \frac{1}{\bar{x}} e^{-\frac{X_{N+1}}{\bar{x}}} \qquad (10)$$

which is unacceptable since this merely substitutes the maximum likelihood estimator of θ, $\hat{\theta} = \bar{x}^{-1}$ in the sampling density of x_{N+1} for all N. The estimate above would only be adequate for sufficiently large N. Because the predictive density of x_{N+1} is always more diffuse than the posterior density of θ we shall in what follows focus only on the former.

For application of the method of moments we first need to equate the first two sample moments of the exchangeable random variables $X_1,\ldots,X_N$ in $f(x^{(N)}|\delta,\gamma)$ to their expectations. This yields

$$\bar{x} = \frac{\gamma}{\delta-1}$$

and

$$N^{-1}\sum_{1}^{N} X_i^2 = \frac{2\gamma^2}{(\delta-1)(\delta-2)} \qquad (11)$$

with the restriction that $\delta>2$. Note that originally δ was only restricted to be positive. The solution from (11) is for $N\geq 2$

$$\delta_M = \max(2, \frac{2(N-1)}{N-1-t^2}) \qquad (12)$$

where $t^2 = N\bar{x}^2/s^2$ and $(N-1)s^2 = \sum_{1}^{N}(x_i-\bar{x})^2$, and $\gamma_M = \bar{x}(\delta_M-1) \geq \bar{x}$.

A drawback is the restriction $\delta_M \geq 2$ which is minor when compared to $\hat{\delta} \to \infty$

for the maximum likelihood estimation. Hence the estimated predictive density is

$$f_M(X_{N+1}|X^{(N)}) = \frac{\Gamma(N+\delta_M)(N\bar{x}+\delta)^{N+\delta_M}}{\Gamma(\delta_M)[N\bar{x}+\gamma_M+X_{N+1})^{N+\delta_M}} . \qquad (13)$$

We noted previously that for unbounded values of (γ_M, δ_M) which resulted from the maximum likelihood values for example we obtained

$$\frac{1}{\bar{x}} e^{-\frac{X_{N+1}}{\bar{x}}}$$

and as N grows this approached $\theta e^{-\theta X_{N+1}}$. Hence, whether or not we obtain finite or unbounded values for (δ_M, γ_M), as N grows,

$$f_M(X_{N+1}|X^{(N)}) \to \theta e^{-\theta X_{N+1}} . \qquad (14)$$

However it can be shown that for all N, γ_M and δ_M are finite with probability 1. At any rate, it would appear that the method of moments is a definite improvement over the maximum likelihood approach for this problem.

4. Predictive Sample Reuse Procedure

We now propose yet another method for estimating the predictive density based on the predictive sample reuse (PSR) approach Geisser (1975). In this case we do not use the marginal density $f(x^{(N)}|\delta,\gamma)$ but

$$E(X_{N+1}|X^{(N)},\delta,\gamma) = \frac{\gamma+N\bar{x}}{N+\delta-1} ,$$

from the actual predictive density of X_{N+1}. We can form a predictor for x_j from $x_{(j)}$, the other N-1 observations with x_j deleted, along with γ and δ based on the above expectation, namely

$$\tilde{x}_j = \frac{\gamma+(N-1)\bar{x}_{(j)}}{N-1+\delta-1} \tag{15}$$

where $(N-1)\bar{x}_{(j)} = \sum_{i\neq j}^{N} x_i$. We then form a discrepancy measure

$$D = \sum_j \left[x_j - \frac{\gamma+(N-1)\bar{x}_{(j)}}{N-1+\delta-1} \right]^2 \tag{16}$$

and minimize this with respect to γ and δ. Taking derivatives of D with respect to γ and δ and setting them equal to zero yields as solutions

$$\hat{\delta} = \max \left[1, \frac{t^2 + \frac{N-1}{N-2}}{t^2 - \frac{N-1}{N-2}} \right] \tag{17}$$

$$\hat{\gamma} = (\bar{\delta}-1)\bar{x}. \tag{18}$$

Note that as t^2 grows $\tilde{\delta} \to 1$ from above and $\tilde{\gamma} \to 0$ from above. Hence the estimate of δ is at least as large as 1, which is an improvement over the method of moments approach which required the estimate of δ to be at least as large as 2. The estimated predictive density now is

$$\tilde{f}(X_{N+1}|X^{(N)}) = \frac{\Gamma(N+\tilde{\delta})(N\bar{x}+\tilde{\gamma})^{N+\tilde{\delta}}}{\Gamma(\tilde{\delta})(N\bar{x}+\gamma+X_{N+1})^{N+\tilde{\delta}+1}} \qquad (19)$$

and we note that as N grows $\tilde{f}(X_{N+1}|X^{(N)})$ approaches the sampling density of X_{N+1} since $(\tilde{\delta},\tilde{\gamma})$ will be finite with probability 1 for all N. Using the relationship

$$\gamma = \bar{x}(\delta-1)$$

we express the method of moments and sample reuse estimated predictive density as

$$\frac{\Gamma(N+\hat{\delta})}{\Gamma(\hat{\delta})(N+\hat{\delta}-1)\bar{x}} \left[1 + \frac{X_{N+1}}{(N+\hat{\delta}-1)\bar{x}} \right]^{-(N+\hat{\delta}+1)} \qquad (20)$$

with $\hat{\delta} = \tilde{\delta}$ or δ_M. Note also that for the non-informative improper prior density on θ, namely,

$$p(\theta) \propto \frac{1}{\theta}$$

we merely substitute $\hat{\delta} = \hat{\gamma} = 0$ to attain the appropriate result from (20). All of these values for γ,δ serve only to change the effective sample size from N to $N+\hat{\delta}-1$ while preserving the mean $\bar{x}$.

5. Comparison of δ_M and $\tilde{\delta}$.

We now further compare δ_M with $\tilde{\delta}$ to indicate a preference for the sample

reuse procedure. For N=2, $\delta_M = 2 > \tilde{\delta} = 1$. For N=3, $\delta_M > \tilde{\delta}$ for all t^2.

For $N \geq 4$

$$\delta_M > \tilde{\delta} \quad \text{for } 0 \leq t^2 < \frac{N-1}{N-2}$$

$$\delta_M < \tilde{\delta} \quad \text{for } \frac{N-1}{N-2} \leq t^2 \leq a(N)$$

$$\delta_M > \tilde{\delta} \quad \text{for } t^2 \geq a(N)$$

where

$$a(N) = \frac{[12(N-1)^2(N-2)+(N-1)^4]^{\frac{1}{2}} - (N-1)^2}{2(N-2)}$$

Further a(N), bounded below by 2, increases monotonically to an upper bound of 3. It can also be shown that $N^{-1}t^2 \to 1$ as N grows. This implies that $\tilde{\delta} \to 1$ as N increases. We also note that δ_M has a singularity at $t^2 = N-1$, a value in which neighborhood t^2 is expected to be with a non-negligible frequency. Hence δ_M will behave somewhat erratically with an appreciable frequency. To illustrate this, a monte carlo experiment was performed calculating δ_M, $\tilde{\delta}$ and $R=\delta_M/\tilde{\delta}$ for n=5,10,20,30. Although frequency diagrams bear out the previous remarks we shall only report the means and standard deviation involved to 3 significant figures.

Table 1: Means and (Standard Deviations) of δ_M, $\tilde{\delta}$ and R based on 1000 replications

N	δ_M	$\tilde{\delta}$	R
5	6.01	1.80	3.23
	(24.9)	(1.62)	(12.3)
10	12.6	1.24	9.75
	(120)	(.154)	(93.0)
20	17.2	1.11	15.3
	(122)	(.050)	(109)
30	14.6	1.07	13.5
	(103)	(.025)	(95.7)

The only time $\tilde{\delta}$ can be very large is when t^2 is to the right of, but very close to, (N-1)/(N-2), an interval which will be of exceedingly low probability for t^2 and hence for $\tilde{\delta}$. Actually if $\tilde{\delta}$ takes on a large value one might be suspicious of the exponential assumption for the sampling distribution.

In summary, we conclude that $\tilde{\delta}$ is greatly preferred to δ_M because of its stability and its drastically reduced relative influence on the effective sample size appearing in the predictive distribution.

6. Remarks

The PSR method while useful in this problem can have a drawback in other problems since it requires that the predictive function contain all the unknown hyperparameters. When this is not the case the method obviously fails. However in these cases it can still be used to estimate those parameters that are

included in the predictor. These values can then be plugged back in so that

$$f(x^{(N)}|\tilde{\tau}_1,\tau_2)$$

where $\tau = (\tau_1,\tau_2)$ and τ_2, the set not included now can be estimated by either of the two other procedures from the above estimated likelihood. Another approach would be to use further predictive functions which would involve predicting x_j^2, for example, by the predictive expectation of EX_j^2, say, if it included τ_2 and use the PSR method on it to obtain $\tilde{\tau}_2$.

7. Acknowledgment. This work was supported by NIGMS grant 25271.

References

1. Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics, London, Chapman and Hall.

2. Good, I.J. (1965). The estimation of probabilities, MASS. M.I.T. PRESS

3. Geisser, S. (1975). The predictive sample reuse method with applications. JASA, 70, 350, 320-328.

4. Robbins, H. (1965). An empirical Bayes approach to statistics, Proc. 3rd Berkeley Symp., 131-148.