RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

APRIL 1997

A BAYESIAN NONPARAMETRIC COMPARISON OF TWO TREATMENTS

WORKING PAPER #9705-12

BY

STEPHEN WALKER
IMPERIAL COLLEGE, LONDON, UK

AND

PAUL DAMIEN
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

A Bayesian Nonparametric Comparison of Two Treatments

Stephen Walker[1] and Paul Damien[2]

[1] Department of Mathematics, Imperial College, 180 Queen's Gate, London SW7 2BZ.

[2] Business School, University of Michigan, Ann Arbor, 48109-1234, USA.

Summary

In this paper we present a full Bayesian nonparametric analysis of survival time data, involving information from two types of treatment. The goal of the analysis is to determine whether there is a difference between the two treatments, according to some well defined criteria, which would justify the use of one in preference to the other. In this paper we present an easy to compute posterior distribution which provides direct insight into the difference of interest.

Keywords: Censored data Bayesian bootstrap, exchangeability, partial exchangeability, beta process, survival data.

1 Introduction

A common dataset which arises in a medical context consists of information gathered on individuals, from one population, exposed to one of two possible types of treatment, usually allocated at random: the goal — to determine which is the 'better' treatment. The specific observables we will be concentrating on is when the responses are survival times; that is, $X_1, \cdots, X_n$ are iid from F_1 and $Y_1, \cdots, Y_m$ are iid from F_2, with both the Fs having support on $(0, \infty)$, and the observations are subject to random right censoring.

Modelling differences

The key issue is how to model 'differences' between the two treatments; equivalently, the difference between F_1 and F_2. A recent approach, Hsieh (1996a), is to take F_1 as a 'baseline' distribution and assume that F_2 is a combination location-scale shift of F_1; that is,

$$F_2(x) = F_1\left(\frac{x-\mu}{\sigma}\right),$$

for some (μ, σ) which are to be estimated, and F_1 may then be regarded as a nuisance parameter. Here x denotes the log-survival time.

A popular approach is to assume a proportional hazards model (Cox, 1972) in which differences are modelled through the hazard functions:

$$\Lambda_2(x) = \Lambda_1(x) \exp \mu.$$

Here Λ_1 is often regarded as a nuisance parameter and interest focuses on estimating μ. Recently, Hsieh (1996b) generalised the proportional hazards model to

$$\Lambda_2(x) = (\Lambda_1(x))^\sigma \exp \mu.$$

In this paper we eschew these scale and location shift models as explanations of the difference between treatments. We look for a more robust procedure in which the data itself determines the form of the difference. This is done in a novel way in which we rely on the notions of exchangeability and partial exchangeability.

Exchangeability/partial exchangeability

Let us consider two extreme cases: firstly, under no treatment difference we would regard the entire set of survival times as arising from a single. but unknown, distribution function. In a Bayesian context this would be equivalent to regarding all the individuals as being *exchangeable.* Secondly, at the other extreme, we would not expect information from one group to help in the understanding of the other group, and vice versa. In a Bayesian context, this would be equivalent to the assumption of *partial exchangeability*; that is, individuals are regarded as being exchangeable within groups, and independent between groups. We refer to the exchangeable structure as S_e and the partial exchangeable structure as S_p.

We assume that the truth lies somewhere in between these two extremes. That is, the correct structure S_c is given by

$$S_c = \pi S_e + (1 - \pi) S_p,$$

for some $\pi \in [0, 1]$. Here π will be taken to be the posterior probability that S_e is the 'true' structure, given by

$$\pi = \frac{\pi_0 p(\text{data}|S_e)}{\pi_0 p(\text{data}|S_e) + (1 - \pi_0) p(\text{data}|S_p)},$$

where π_0 is the prior choice for the probability that S_e is the 'true' structure. The above equations mean that to work with structure S_c we would carry out analyses S_e and S_p and then combine these according to the weight π. For example, suppose that interest is in estimating the parameter θ and let $\hat{\theta}_e$ be the estimate obtained under the exchangeable model, $\hat{\theta}_p$ the corresponding estimate under the partial exchangeable model; so

$$\hat{\theta} = \pi \hat{\theta}_e + (1 - \pi) \hat{\theta}_p$$

is the required estimate.

For our proposed method of analysis we are required to assign a (nonparametric) prior to the appropriate Fs, leading to simple calculations of $p(\text{data}|S)$, and hence π. Throughout the rest of the paper we assume that F_1 is the probability distribution of survival times for treatment A and F_2 the corresponding distribution for treatment B. If interest is in obtaining a posterior distribution for a parameter of interest, θ_A, associated with treatment A, then

$$p(\theta_A|\text{data}) = \pi p(\theta_A|\text{all data}) + (1 - \pi) p(\theta_A|\text{data from A}),$$

where $p(\theta_A|\text{data from A})$ is the posterior based solely on the information provided by those observations from treatment A, and $p(\theta_A|\text{all data})$ is the posterior under the exchangeable assumption which assumes treatment A and B are identical and hence is based on all observations.

In Section 2 we discuss the nonparametric prior for the Fs and in Section 3 we calculate π. Section 4 discusses a selection criterion for the 'better' treatment, based on the ideas appearing in Spiegelhalter et al. (1994). Section 5 contains two illustrative examples.

2 Beta process: prior and posterior

In a Bayesian nonparametric framework we will be assigning a prior distribution on the space of probability distributions on $(0, \infty)$. We will be working with a generalisation of the Dirichlet process (Ferguson, 1973), which we need in order to adequately express prior opinion about the unknown distributions. We consider a discretisation of the continuous space $(0, \infty)$, say $\{ds, 2ds, \cdots\}$ for some appropriate ds, and will therefore be assigning a nonparametric prior on $\{1, 2, \cdots\}$: this is required to calculate $p(\text{data})$. Even if we used a continuous prior (process) an approximate discrete version of the process would be required to calculate $p(\text{data})$. Typically, data arising from survival analyses, do so in a discrete form — as information obtained each day, week, or during some other unit of time — and so no generality is lost in developing the prior to posterior calculations in a discretized framework, at the outset. Let λ_t, for $t = 1, 2, \cdots$, be independent and distributed according to beta distributions with parameters (α_t, β_t). Then a random survival distribution $S(t)$ can be obtained from the discrete version of the beta process (Hjort, 1990; Walker and Muliere, 1997) via

$$S(t) = \prod_{s \leq t} (1 - \lambda_s) .$$

If $X_1, \cdots, X_n$ are iid from $S(t)$, possibly with arbitrary right censoring, then the posterior parameters are given by $\alpha_t + n_t$ and $\beta_t + m_t$, where n_t is the number of deaths at time t and m_t is the number of survivors just before time t. A Bayesian bootstrap procedure (Lo, 1993) would be to obtain the posterior parameters and then set the prior parameters to zero; that is, the posterior parameters are (n_t, m_t).

Within this discrete framework we can calculate $p(\text{data})$ straightforwardly,

$$p\big(X_1,\cdots,X_n\big) = \prod_t \left\{ \frac{\alpha_t^{[n_t]}\beta_t^{[m_t]}}{(\alpha_t+\beta_t)^{[n_t+m_t]}} \right\},$$

where $a^{[n]} = a(a+1)\cdots(a+n-1)$ and $a^{[0]} = 1$. Whereas we could consider the zero prior parameters to obtain the posterior distribution, we can not do this to calculate π, which is clear from the above expression for $p(\text{data})$. In the next section we suggest a way to obtain the prior parameters in a meaningful way.

3 Calculating π

We have seen that in order to calculate π we have to evaluate $p(\text{data}|S_e)$ and $p(\text{data}|S_p)$. Let n_t^A and m_t^A be the death process and at-risk process, respectively, for group A and n_t^B and m_t^B the corresponding processes for group B. Also let $n_t = n_t^A + n_t^B$ and $m_t = m_t^A + m_t^B$. Straightforwardly, using the result for $p(X_1,\cdots,X_n)$, we obtain

$$p(\text{data}|S_e) = \prod_t \left\{ \frac{\alpha_t^{[n_t]}\beta_t^{[m_t]}}{(\alpha_t+\beta_t)^{[n_t+m_t]}} \right\}$$

and

$$p(\text{data}|S_p) = \prod_t \left\{ \frac{\alpha_t^{[n_t^A]}\beta_t^{[m_t^A]}}{(\alpha_t+\beta_t)^{[n_t^A+m_t^A]}} \frac{\alpha_t^{[n_t^B]}\beta_t^{[m_t^B]}}{(\alpha_t+\beta_t)^{[n_t^B+m_t^B]}} \right\}.$$

Hence an expression for π can now be constructed.

We could be noninformative in our approach, by taking $\alpha_t = \beta_t = 0$ for all t, in order to estimate $p(X > t)$, for example, but this is not possible when calculating $p(\text{data}|S_e)$ and $p(\text{data}|S_p)$. Assigning non zero values to α_t and β_t, would, of course, imply an informative prior. Here then we discuss a way to select these prior parameters.

If we let $S(t)$ denote the random probability $p(X > t)$, we obtain

$$E[S(t)] = \prod_{s\leq t} \frac{\beta_s}{\alpha_s+\beta_s} \quad \text{and} \quad E[S^2(t)] = E[S(t)] \prod_{s\leq t} \frac{\beta_s+1}{\alpha_s+\beta_s+1}.$$

Our aim now is to match these first two moments from the nonparameteric model with those from a parametric model. We choose the geometric distribution for its simplicity and ease of interpretation. Our (Bayes) parametric model is given by

$$p(X = t|\lambda) = \lambda(1-\lambda)^{t-1}, \;\; t = 1, 2, \cdots,$$

and $\lambda \sim \text{beta}(a, b)$. Therefore,

$$E_\lambda[S(t)] = \frac{b^{[t-1]}}{(a+b)^{[t-1]}} \text{ and } E_\lambda[S^2(t)] = \frac{b^{[2t-2]}}{(a+b)^{[2t-2]}},$$

leading, after some algebra, to

$$\beta_t = \frac{\xi_t(1-\delta_t)}{\delta_t - \xi_t} \text{ and } \alpha_t = \beta_t(1/\xi_t - 1),$$

where $\xi_t = E_\lambda[S(t)]/E_\lambda[S(t-1)]$ and $\delta_t = \xi_t^{-1}E_\lambda[S^2(t)]/E_\lambda[S^2(t-1)]$. A noninformative approach, which would presumably take $a = b = 1$, simplifies to

$$\beta_t = t\alpha_t = t/(2t^2 - 1).$$

These are the values for α_t and β_t which will be used to calculate π. Note, then, that it is not possible for us to consider the Dirichlet process, since this not only requires $\sum_t \alpha_t < \infty$ (which follows from above), but also $\beta_t = \sum_{s>t} \alpha_s$, which does not follow from the development above.

4 Selecting treatments

Our strategy for comparing and selecting the 'best' treatment is based on the work by Spiegelhalter et al. (1994, Section 2) and the reader is referred to that paper for further insights. Here we briefly discuss the main aspects of the decision making process.

We assume that the selection of the preferred treatment will be based on the parameter $\delta = \theta_A - \theta_B$, where each θ corresponds to a parameter of interest for each of the treatments; for example, $\theta = p(X > t_0)$ for some time point t_0. We think of treatment B as being the current choice of treatment and A a new treatment under investigation. Then according to the arguments put forward by Spiegelhalter et al. (1994) we would select treatment A in preference to treatment B if $\delta > \delta_S$, indicating clinical superiority of the new treatment. The full significance of δ_S and its selection criteria is discussed by Spiegelhalter et al. (1994), and references cited therein.

Of course the θs, and hence δ, are unknown. Spiegelhalter et al. (1994, Sections 3 and 4) develop *parametric* posterior distributions for δ. In our Bayesian nonparametric framework we will use the posterior beta processes to construct the posterior distribution for θ_A and θ_B, and hence for δ. The use of the beta process is equivalent to using the censored data Bayesian bootstrap (Lo, 1993); in the limiting case $\alpha_t, \beta_t \to 0$. We will use this bootstrap method to generate samples from the posterior distribution of δ. Here we briefly detail the method.

For $s = 1, \cdots, t_0$, we generate λ_s^A from $\beta(n_s^A, m_s^A)$ (independently) and set $\theta_A = \prod_{s \le t_0}(1 - \lambda_s^A)$, with probability $1 - \pi$, or generate λ_s from $\beta(n_s, m_s)$ (independently) and set $\theta_A = \prod_{s \le t_0}(1 - \lambda_s)$, with probability π. We do a similar procedure for θ_B, based on (n_s^B, m_s^B), and define θ_B accordingly, and set $\delta = \theta_A - \theta_B$. If this is δ_1, we repeat this simulation N times, and use $(\delta_1, \cdots, \delta_N)$ to construct the posterior for δ.

Although there are a number of choices for θ we will consider the case when $\theta = p(X > t_0)$ for some t_0. A more general case is to define $\theta = \sum_{t=1}^{t_0} u_t p(X > t)$, where $\sum_{t=1}^{t_0} u_t = 1$, and the u_t are chosen to reflect the relative importance of surviving beyond time t; i.e., we would take $u_t < u_{t+1}$ for all t.

5 Numerical examples

Example 1. Our first example involves leukemia remission times (in weeks) and the two treatments are an active drug and placebo. The data was discussed and analysed by Cox (1972), Kalbfleisch (1978) and more recently by Laud et al. (1996). The times for the active drug are

$$6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*,$$

where a $*$ denotes a censored observation; the times for the placebo are

$$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.$$

Taking $\pi_0 = 1/2$ we compute π to be 2×10^{-4}. We take this to mean that the data from the two groups should be analysed separately.

For illustration purposes, we construct the posterior distribution for δ based on $\theta = p(X > 10)$, using the censored data bootstrap method. The posterior is presented in Figure 1, and clearly demonstrates the superiority, in survival beyond ten weeks, for treatment A; this is quite similar to the result obtained, using a Cox regression, by Laud et al. (1996).

Example 2. Our second example is taken from Lawless (1982, Example 7.2.1) and also involves remission times, in weeks. Each treatment, A and B, has 20 patients: the observations from treatment A are

$$1, 3, 3, 6, 7, 7, 10, 12, 14, 15, 18, 19, 22, 26, 28^*, 29, 34, 40, 48^*, 49^*$$

and from treatment B are

$$1, 1, 2, 2, 3, 4, 5, 8, 8, 9, 11, 12, 14, 16, 18, 21, 27^*, 31, 38^*, 44.$$

Again, taking π_0 to be $1/2$, we compute the posterior probability π to be 0.77, providing support for the exchangeable structure and no difference in the treatments. For these data, based on a number of (frequentist) tests, Lawless also concludes that there is "no evidence of a difference in distribution."

As in Example 1, we construct the posterior distribution for δ based on $\theta = p(X > 10)$, using the censored data bootstrap method. The posterior is presented in Figure 2, and this time clearly demonstrates equality of treatments, in terms of survival beyond ten weeks.

Discussion

In this paper we develop a simple method, via a nonparametric prior distribution, to selecting the "better" treatment. The novelty of our approach is four-fold: firstly, it offers the practitioner to use contextual prior information in a natural way to model the uncertainty in the underlying survival functions corresponding to the two treatments; secondly, treatment differences are identified based on whether or not the data is exchangeable or partially exchangeable; thirdly, the calculation of a formal decision rule enables the selection of the preferred treatment; and fourthly, our solution does not require Markov chain Monte Carlo simulations. A very easy-to-implement bootstrap is the only simulation aspect to the solution, the rest being closed form analytical expressions.

References

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-202.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 615-629.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* **18**, 1259-1294.

Hsieh, F. (1996a). Empirical process approach in a two-sample location-scale model with censored data. *The Annals of Statistics* **24**, 2705-2719.

Hsieh, F. (1996b). A transformation model for two survival curves: An empirical process approach. *Biometrika* **83**, 519-528.

Kalbfleisch, J.B. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B* **40**, 214-221.

Laud, P.W., Damien, P. and Smith, A.F.M. (1996). Bayesian nonparametric and covariate analysis of failure time data.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons.

Lo, A.Y. (1993). A Bayesian bootstrap for censored data. *The Annals of Statistics* **21**, 100-123.

Spiegelhalter, D.J.,Freedman, L.S. and Parmar, M.K.B. (1994). Bayesian approaches to randomised trials (with discussion). *Journal of the Royal Statistical Society, Series A* **157**, 357-416.

Walker, S.G. and Muliere, P. (1997). Beta-Stacy processes and a generalisation of the Polya-urn scheme. To appear in *The Annals of Statistics.*

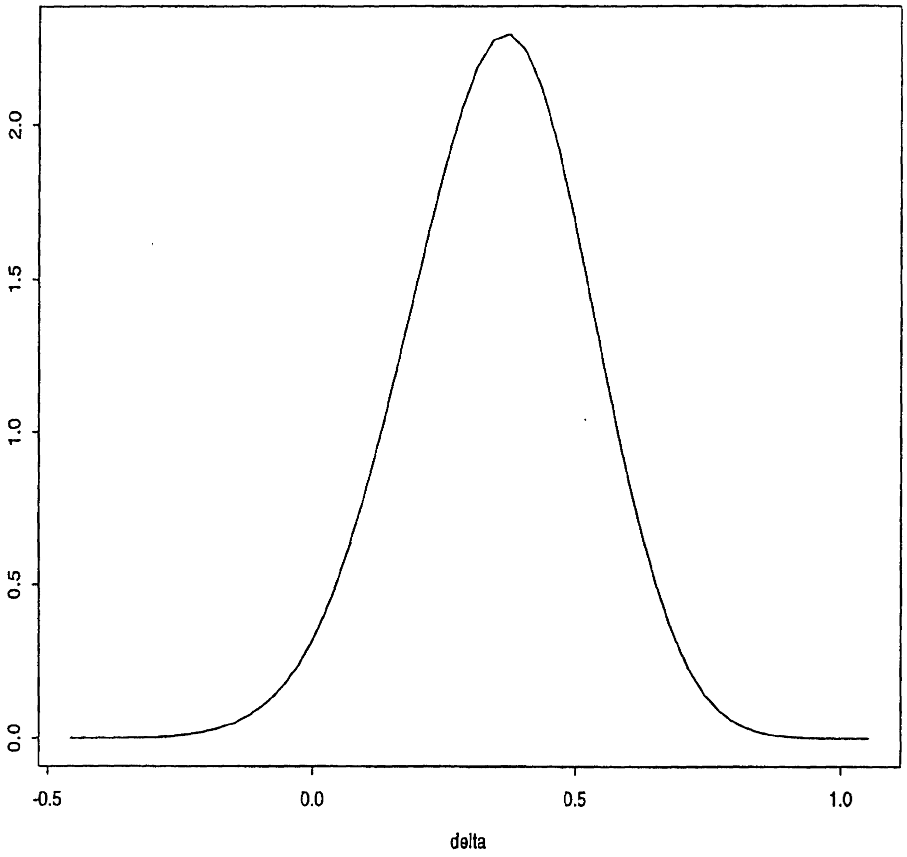

Figure 1: Posterior distribution for δ based on $\theta = p(X > 10)$, Example 1

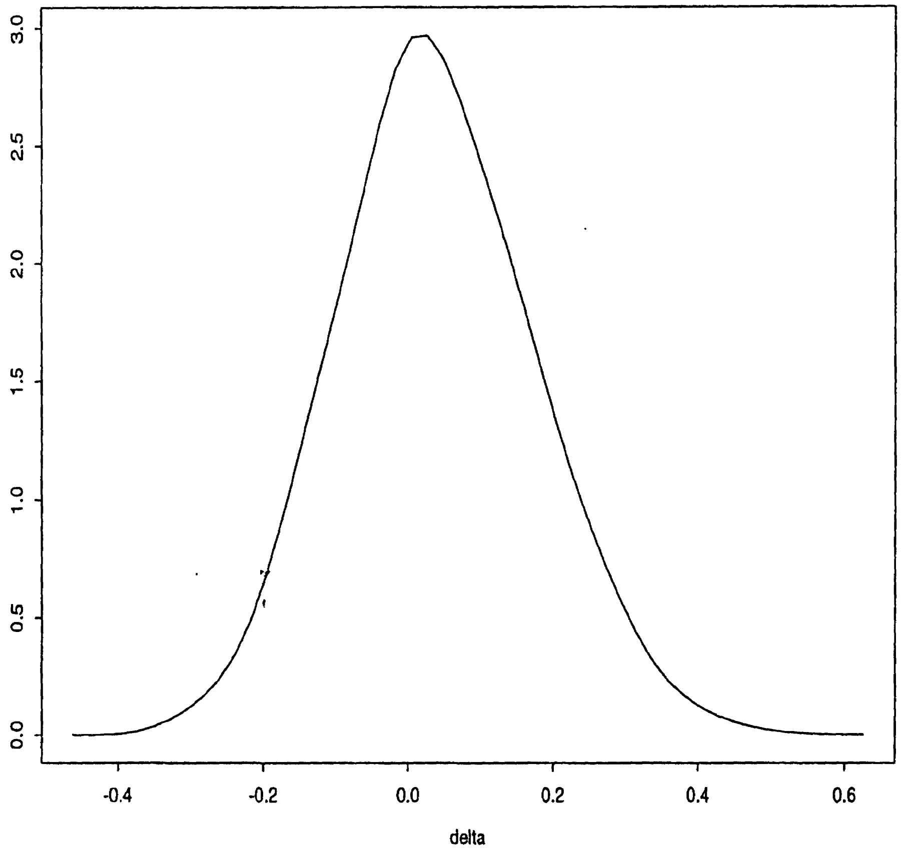

Figure 2: Posterior distribution for δ based on $\theta = p(X > 10)$, Example 2