THE UNIVERSITY OF BRITISH COLUMBIA

DEPARTMENT OF STATISTICS

TECHNICAL REPORT# 273

Bayesian nonparametric subset selection procedures with Weibull components

BY

YUMI KONDO & JAMES V ZIDEK

MAY 2013

Bayesian Analysis (2004) **1**, Number 1

Bayesian nonparametric subset selection procedures with Weibull components.

Yumi Kondo
Department of Statistics, University of British Columbia
James V. Zidek
University of British Columbia

Abstract. This paper presents two solutions within a Bayesian framework to the problem of selecting a subset of populations of breaking strengths from a specified finite set of them that will contain the one with the smallest or (equivalently from a mathematical perspective) largest α^{th} quantile. Independent random samples from each are assumed. Estimates of these quantiles are then used to specify conservative design values for engineering applications. As ASTM, Inc standards require in some cases a nonparametric approach, we propose a semiparametric and nonparametric alternatives in the ensuing sections. The paper wraps up with an empirical comparison of the methods followed by conclusions.

Keywords: Bayesian nonparametric procedures; subset selection; quantiles; Weibull distribution; Dirichlet process; stick-breaking prior

AMS 2000 classifications: 62F07, 62G99

1 Introduction

This paper, which presents approaches to a subset selection problem within a Bayesian framework, is a companion to Van Eeden and Zidek (2012). Both papers have their genesis in a problem encountered by the second author in work done on proprietary data under a nondisclosure agreement. While we are not able under that agreement to give specific details about that problem, we use it to motivate a methodology for subset selection presented in this paper. The general context is that of a manufactured item with multiple sources of supply that is sold under a single label. The population of all items $\mathcal{P}$ can be represented as $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{P}_k$ where K is the number of sources of supply. Each item has a measurable index T of its reliability and the focus of the paper is the conservative specification of a reliability coefficient (RC) for the label such that the index T_j for a randomly selected item j, exceeds RC with high probability say at least 95% for definiteness in this exposition. Since the item could come from any one of the sources of supply and the items available for distribution at any one time will not generally be a random mix of items from all subpopulations, the criterion would lead ideally to identifying the population $\tau \in \{1, \ldots, K\}$ with the smallest fifth percentile $\eta_{k,0.05}$ and then reducing that value by a safety factor to account for various uncertainties. However, generally it will not be feasible to collect samples of sufficient size from each of the subpopulations, to identify τ with certainty. A more practical approach, the one taken in lumber production, selects a subset $S \subset \{1, \ldots, K\}$ that includes τ with high probability say exceeding P^*.

Our problem thus involves manufactured lumber produced by different manufacturers in different geographical regions at different times from different, but similar species, depending on their availability. Subpopulation $k \in \{1, \ldots, K\}$ refers to one of those subdivisions of the population of manufactured lumber. For use in structural engineering, the *RC* of lumber must ensure survival under unforeseeable future loads due to such things as snow, strong winds, and seismic activity. This requirement has been implemented through a quality control classification process called "grading". The grade G assigned to a piece of lumber, which is stamped on its surface, is assigned in accordance with prescribed grading rules based on characteristics of that specimen. Most lumber used in the construction of houses for example is of grade $G = \#2$ or better. Each such grade corresponds to a cross section of the global supply of lumber, $\mathcal{P}$, which we denote by $\mathcal{P}_G = \bigcup_{k=1}^{K} \mathcal{P}_{Gk}$.

Each grade G has a specified design value (called its allowable property) meaning to the consumer, that the breaking strength (hereafter strength) of a randomly selected wood specimen will exceed that value with high probability. But wood is a highly variable material. So the strength distributions for the $\{\mathcal{P}_{iG},\ i = 1, \cdots, K\}$ will not be identical. But it is neither practical or desirable to attempt to establish separate design values for each (k, G) pair. Instead the lumber industry has chosen to specify a single conservatively selected design value for G across the global population.

How should that single value be found? An obvious method, the so called "in–grade" approach, takes a multistage cluster sample of pieces of lumber from $\mathcal{P}_G$. Items in that sample are destructively tested in a lab to obtain an estimate of the relevant population percentile, which after a reduction based on engineering considerations gives the design value for G.

The first method will not generally be feasible since it requires an administrative infrastructure to organize and manage the sampling program. Thus an alternative method has been developed for lumber and this paper concerns that second method. There each of the $\{\mathcal{P}_{Gk},\ k = 1, \cdots, K\}$ is sampled and tested separately, although these samples in combination cannot be considered as representative of $\mathcal{P}_G$. Yet commercial considerations dictate that nevertheless, all lumber in these K subgroups be marketed as a single grade group with a single design value. For historical reasons, this practice is called "species grouping" and the K subgroups are called "species" although today, they may refer to other factors such as region. Species grouping is sufficiently common that prescriptive protocols for finding design values for species groups are published as document ASTM D1990 (ASTM Standard D1990 (2007); hereafter D1990).

The ASTM protocols ensure a conservatively low value while providing stability under change, such as when a particular subpopulation k is withdrawn or is no longer available. The latter is achieved by D1990's stipulation that the design value be calculated by combining the samples from members of a conservatively selected subgroup of the K groups called the "subgroup of controlling species" (hereafter CS). We interpret the CS in our approach, as a subgroup that contains species $\tau \in \{1, \ldots, K\}$, the one with the smallest fifth percentile, at least with reasonable certainty. The estimated fifth percentile based on the CS (or rather in keeping with the ASTM protocol, one-sided lower 5 percent tolerance limit (TL) with 75 percent confidence), divided by a safety factor (as in the ASTM protocol), can then be adopted as the the design values for the species group.

As noted above, we view the lumber application as a paradigm in reliability for supply chains involving multiple sources of supply leading to a commercial product marketed under a single label with a specified reliability coefficient. The same issues as those confronted in our application, will arise due to the heterogeneity of subgroups created by such factors as the source of raw material as well as differences in the workforce, manufacturing technology, measurement technique, or inspection methodology. The controlling species approach developed for manufactured lumber may provide a practical way of setting a conservative, stable reliability coefficient for the combined population of the manufactured items. With that we turn to statistical aspects of the problem.

The statistical methods prescribed by D1990 for finding the subset of controlling species, which we will call the ASTM approach, avoid parametric models for the strength distributions of the $\{\mathcal{P}_{iG},\ i = 1, \cdots, K\}$. But these non–parametric methods, while avoiding strong modelling assumption, are quite complex and their theoretical properties are very difficult to assess. We focus here on the protocol that concerns the 5 percent TL.

That protocol involves a multiple testing procedure based on the chi–squared test and lower TL. First the TL is computed for all K samples combined. For each species sample $k \in \{1, \ldots, K\}$ the number $\hat{n}_k$ below that TL is found and then these numbers are compared by the chi–squared test. Non–rejection (at the 0.01 level of significance) puts all K species into the CS. Rejection leads to a sequential multiple testing procedure beginning with a comparison of the two subpopulations with the two largest $\hat{n}_k$s by the chi–squared test. The procedure continues in a stepwise fashion until rejection at which point the CS is taken to consist of subpopulations considered up to but not included in that test.

The ASTM procedures have recently been found to produce unexpected results, leading to a search for an alternative. That search led to the method in Van Eeden and Zidek (2012) (RS for short), the nonparametric empirical Bayes method referred to hereafter as the Bayesian nonparametric procedure in Taylor et al. (2008) and the new procedure in this paper called the Bayes semiparametric method.

The procedures presented in this paper are developed in a Bayesian framework that allows more flexibility than the one in RS allows, albeit at the expense of greater complexity. Most notably the new framework allows for the incorporation of expert knowledge, about particularly the lower tails of the population distribution, the region of critical concern. At the same time, the two principal methods described in this paper do retain to a great extent, the nonparametric character of the ASTM prescription.

Work on the Bayesian semiparametric procedure presented in this paper, like those in Van Eeden and Zidek (2012) and Taylor et al. (2008), began by reinterpreting the basic problem as none of subset selection rather than multiple testing as in D1990. That new paradigm and more generally the topic of ranking and selection has a long history within the domain of repeated sampling (Van Eeden and Zidek 2012). Much less has been done in Bayesian analysis, a notable contribution being that of Berger and Deely (1988). There the goal is the selection of the normally distributed population with the largest mean, once the null hypothesis of the equality of their means has been rejected. Fong and Berger (1993) provide a history. However, we know of no work on the problem of subset selection within a nonparametric Bayes framework.

We now describe the contents of this paper. Section 2 presents a novel general approach to subset selection within a Bayesian framework. Ways of optimizing the procedure are proposed there, including how costs may be incorporated. To implement the general methodology in reliability analysis, we develop it for the two parameter Weibull population distribution.

Section 3 presents the Bayesian nonparametric approach in Taylor et al. (2008), which builds on work in a different context (Johnson et al. 1999b) involving the strength properties of lumber. The Weibull distribution with estimated parameters serves as the baseline distribution for the Dirichlet process (DP) prior. This approach has the advantage of simplicity and computational speed. However it has the disadvantage that the prior makes the sampling distribution discrete — it takes its jumps at a countable number of random points with probability one.

That disadvantage leads for the fully Bayesian semiparametric approach in Section 4 at the expense of greater technical complexity. This method uses a fairly standard approach (Escobar and West 1995) that assumes a parametric sampling distribution, which conditional on a random shape and scale parametric is Weibull. These parameters, which are random effects that vary from specimen–to–specimen, are sampled from a DP. The result is in effect is an infinite mixture of Weibull distributions that learns the failure modes of which there may be more than one, reflected in a bumpy lower left hand tail of the strength distribution. These modes correspond to clusters of the random effects referred to above. Section 5 illustrates our proposed Bayesian nonparametric and semiparametric procedures based on three real datasets of species, collected in the Forest Products Stochastic Modelling Group. The program is funded by NSERC and is a collaborative effort between UBC, SFU, and FPInnovations. We compare the approaches in Section 6 through a simulation study involving data from a known distribution, which is an adaptation of one constructed by fitting models to a real proprietary dataset. These results provide evidence in favour of the methods proposed in this paper over the ASTM approach. The Bayes nonparametric approach is found to have the advantage of simplicity, but the Bayes semiparametric approach tend overall to be better at least with samples of reasonable size. Conclusions follow in Section 7.

2 Bayesian subset selection

Suppose we have K species and an independent sample $\{\{t_{kj}\}_{j=1}^{m_k}\}_{k=1}^{K}$ of measurements from each $k = 1, \cdots, K$, and measurements across species are also independent. Denote by $t_{(k1)} < \cdots < t_{(km_k)}$ the order statistics for the specie k and $data$=$\{t_{kj}; k = 1, \cdots, K, j = 1, \cdots, m_k\}$ the combined sample from all the species. Furthermore let

$\eta_{k\alpha}$ denote the α^{th}-quantile for species $k = 1, \cdots, K$. Our interest focuses on the smallest of these and we let τ denote the unknown label of the species which possesses it. In other words, $\eta_{\tau\alpha} < \eta_{k\alpha},\ k \neq \tau$ where here and in the sequel we assume no ties.

Practical limitations on the sample sizes mean that we will not be able to identify τ with an acceptably high level of certainty say with a posterior probability of at least P^*. Instead we seek the smallest subset $S \subset \{1, \cdots, K\}$ for which $P(\tau \in S|data) \geq P^*$. To represent this optimization problem in a more explicit form, let $\pi_k \doteq P(\tau = k|data)$ for all $k = 1, \ldots, K$ and $\pi_{(k)}$ be the k^{th} largest π after ranking them from smallest to largest. The optimal Bayesian choice of S would be the subset corresponding to the smallest number d of species with $P^* < \pi_{(K-d+1)} + \cdots + \pi_{(K)}$. Notice that

$$P(\tau = k|data) = \int_0^\infty \Pi_{i \neq k} P(u < \eta_{i\alpha} \mid data) dP(\eta_{k\alpha} \leq u|data). \tag{2.1}$$

Thus solving our problem reduces to characterizing $P(\eta_{i\alpha} \leq u|data),\ 0 < u$ in a suitable form for all $i = 1, \cdots, K$. In general implementation of that solution will depend on the context in which it is to be applied. Guided by our interest in design values and hence in material strength properties, we implement the result above for the Weibull family, which is commonly used in reliability. Its members characterize extreme values (from which design values may be calculated), while including the exponential and Gaussian distributions (approximately in the latter case). In fact a companion to D1990 (ASTM Standard D2915 2011) designates the Weibull along with the lognormal distribution for use in a design context.

The integral in Equation 2.1 can in some cases be found by numerical integration and this was done initially for the approach in Section 3. However, in practice the number of species will usually be small and in that case, the needed probability can be found by Monte Carlo (MC) sampling. Variations of that approach are used in the sections that follow.

Optimizing the procedure

It may not be possible in some applications to specify directly the P^* required for the procedure described above. In that case, the objective function for optimizing the choice of the subset S in Van Eeden and Zidek (2012) is a compromise between the probability of correctly selecting the population τ with the smallest α^{th} quantile, and the need to minimize the size $|S|$ of the subset. From the perspective of multicriteria decision analysis the objective would become

$$\gamma\ P(\tau \in S|data) - (1-\gamma)\ |S|/K \tag{2.2}$$

for some $\gamma \in [0, 1]$, depending on which of the two objectives were seen to be more important. In the absence of any clear ordering of the two, $\gamma = 0.5$ would be a seemingly natural choice. Then the optimal S would be

$$S^{opt} = \underset{S}{\text{argmax}}\{\gamma\ P(\tau \in S|data) - (1-\gamma)\ |S|/K\}.$$

The result: the subset of populations corresponding to the $\{\pi_{(K-d^{\rm opt}+1)}, \ \cdots, \ \pi_{(K)}\}$ where the integer $d^{\rm opt}$ is

$$d^{opt} = \operatorname*{argmax}_d \{\gamma \ (\pi_{(K-d+1)} + \cdots + \pi_{(K)}) - (1-\gamma) \ d/K\}.$$

An alternative, based on a suggestion of Dr Larry Phillips communicated to the second author in a different context, would be the S given by the "bang for the buck" criterion where now $d^{\rm opt}$ is

$$d^{opt} = \operatorname*{argmax}_d \frac{\pi_{(K-d+1)} + \cdots + \pi_{(K)}}{d}.$$

While not a normative criterion like the previous one, this has a natural appeal. As $d = |S|$ increases, at some point, the gains in the probability of correct selection will tend to be outweighed by the increasing size of the S required to attain them.

For brevity, we will not in our empirical assessments carried out in this paper illustrate use of the criteria, leaving that instead to the comparative assessments to be made in a future paper that compares all the methods including the one in D1990 that spawned the work reported here.

Finally, we may optimize the sample sizes as in Van Eeden and Zidek (2012). This can be done in the usual way for Bayesian experimental design, through by pre-posterior analysis. The objective functions would be one of those above evaluated at their associated optimal subset selection rules, conditional on the samples. But now the expectations over the optimized objective function would need to be taken with respect to the marginal distributions of the samples and that, minimized over the $\{n_s\}$ subject to cost constraints. These costs would well differ from population–to–population, and may represent the monetary equivalent of the difficulty of obtaining the samples. However, major computational issues now arise in characterizatizing the optimal sample sizes and this also remains for future work.

3 A nonparametric empirical Bayes procedure

In this section, we implement Equation 2.1 using a nonparametric Bayesian approach. Unlike parametric Bayesian approaches, which restrict the functional form of the sampling distribution's cumulative distribution function (CDF), the nonparametric Bayesian approach allows it to have a flexible form by placing the prior distribution directly on the CDF, hence side-stepping the need to specify a class of parametric models. Ferguson (1973) describes a mathematical structure for doing this by using the Dirichlet process (DP) prior, an infinite dimensional generalization of the Dirichlet distribution. A probability measure H is said to be a realization of the DP with precision parameter v and base measure H_0, denoted as $H \sim DP(v, H_0)$, if any finite partition $A_1, \cdots, A_r$ of the sample space of H has the property that:

$$(H(A_1), \cdots, H(A_r)) \sim Dir(vH_0(A_1), \cdots, vH_0(A_r)),$$

where $Dir(a_1, \cdots, a_r)$ represents the Dirichlet distribution with parameters $a_1, \cdots, a_r$.

Our nonparametric Bayesian approach is based on Johnson et al. (1999a): Denote T_k be the nonnegative strength measurement of the k^{th} species then assume that the probability measure of T_k is from the DP with base measure G_{0k} and precision parameter v. Since we are interested in events of the form $\{T_k \leq t\}$, the definition of DP can be written in terms of the CDF of T_k. More precisely, let G_k be the CDF of T_k, G_{0k} be the base CDF, and $0 < t^1 < t^2 < \cdots < t^r < \infty$. Then $G_k \sim DP(v, G_{0k})$ if for any finite partition of the form $[0, t^1], (t^1, t^2], \cdots, (t^{r-1}, t^r], (t^r, \infty)$ has the joint distribution

$$\begin{aligned}&(G_k(t^1), G_k(t^2) - G_k(t^1), \cdots, G_k(t^r) - G_k(t^{r-1}), 1 - G_k(t^r))\\ &\sim Dir(vG_{0k}(t^1), vG_{0k}(t^2) - vG_{0k}(t^1), \cdots, vG_{0k}(t^r) - vG_{0k}(t^{r-1}), v - vG_{0k}(t^r)).\end{aligned} \tag{3.3}$$

The distribution of α^{th} quantile η_α of T can be derived from (3.3) in terms of a beta distribution. More explicitly for the partition $[0, t]$, (t, ∞), (3.3) becomes

$$\begin{aligned}P(\eta_{k\alpha} < t) =&1 - P(G_k(t) \leq \alpha)\\ =&1 - Beta(\alpha; vG_{0k}(t), v - vG_{0k}(t)),\end{aligned}$$

where $Beta(\cdot; a, b)$ is a CDF of the beta distribution with mean $a/(a + b)$. The second equality follows because of (3.3) and the fact that the Dirichlet distribution is a multivariate generalization of the beta distribution.

The DP has the attractive feature that its posterior distribution is also DP. Under our model, the posterior distribution of the α^{th} quantile $\eta_{k\alpha}$ of k^{th} species is:

$$P(\eta_{k\alpha} \leq t|data) = 1 - Beta(\alpha; \nu_{m_k}(t), v + m_k - \nu_{m_k}(t)), \tag{3.4}$$

where $\nu_{m_k}(t) = vG_{0k}(t) + m_k\hat{F}_k(t)$, and $\hat{F}_k(t)$ is the empirical distribution function. The posterior distributions of the $\{\eta_{k\alpha}\}$ now have a discrete component, and the CDF has jumps at each of the (ordered) sample points t_{kj}, $j = 1, \cdots, m_k$. Thus if in the integrand of Equation (2.1), we let

$$H_s(t) = \Pi_{k \neq s} P(t < \eta_{k\alpha}|data)$$

we can represent that integral explicitly as

$$\int_0^\infty H_s(t)dP(\eta_{s\alpha} \leq t|data) = \sum_{r=1}^{n_{k+1}} [\int_{x_{k(r-1)}}^{x_{kr}} H_s(t)f_{s(r-k)}(t)dt + p_{s(r-1)}H_s(x_{k(r-1)})].$$

In principle this representation could be used to evaluate the integral.

However the integration in Equation (2.1) is done below by sampling independent copies $\eta_{s\alpha}^{(j)}$, $j = 1, \cdots, L$ by first generating samples from the uniform distribution on (0,1) or more succinctly $U(0, 1)$, inverting Equation (3.4), the posterior distribution of $\eta_{s\alpha}$, and then approximating the integral by

$$P(\tau = s|data) \simeq \frac{\sum_{j=1}^L \Pi_{k \neq s} P(\eta_{s\alpha}^{(j)} < \eta_{k\alpha}|data)}{L} \tag{3.5}$$

for an L sufficiently large as to attain approximate convergence.

Implementation

To implement the theory above we must specify the precision parameter v. The form of the posterior distribution of $\eta_{k\alpha}$ in (3.4) suggests that $v/(v+m) = p$ be interpreted as the weight on the prior belief of G_{0k} relative to the empirical distribution function. Therefore we experimentally increase $v = vp/(1-p)$, $p = 0.1, 0.2, \cdots$ to the point at which the results become sensitive to the the $\{G_{0k}\}$. Preliminary data analysis points to the use of a two parameter Weibull CDF for it fits the data the best. More precisely, let $Weibull(\cdot; a, b)$ denote the CDF of the standard Weibull distribution with mean $b\Gamma(1+1/a)$ and let

$$G_{0k}(t) = Weibull(t; \hat{\beta}_k, \hat{\lambda}_k),\ t > 0,$$

$\hat{\lambda}_k$ and $\hat{\beta}_k$ being, respectively, the maximum likelihood estimates of the Weibull's scale and shape parameter. Thus with this definition,

$$\nu_{m_k}(t) = vWeibull(t; \hat{\beta}_k, \hat{\lambda}_k) + m_k \hat{F}_k(t).$$

Overall the method described in this section builds–in a posterior base that provides smoothly increasing posterior cumulative probabilities between the empirical jumps at sample points.

4 A Bayesian semiparametric approach

The main drawback of the nonparametric approach is the somewhat unrealistic assumption of the discrete (singular) distribution on the α^{th} quantiles. This section seeks a compromise between a parametric and a purely nonparametric approach. That goal is accomplished through the use of a by now standard hierarchical DP mixture approach, within the field of nonparametric Bayesian analysis. This field has grown rapidly following the seminal paper of Ferguson (1973, 1974) and is an active area of current research (Broderick et al. 2011).

To characterize $P(\eta_{k\alpha} \leq u|data)$ in Equation (2.1), we first model the distribution of each species T_k separately using the Weibull DP mixture. DP mixture was originally introduced by Antoniak (1974) as a flexible alternative to nonmixture distribution. For notational simplicity, we use T to denote T_k in this Section. Denoting the marginal CDF of T by F and joint CDF for the shape β and scale λ parameters of the Weibull as G, the Weibull DP mixture model is explicitly written as

$$F(t) = \int \int Weibull(t|\beta, \lambda) G(d\beta, d\lambda), \tag{4.6}$$

where $G \sim DP(v, G_0)$. This representation differs from the nonparametric method of Section 3, in that the base CDF G_0 here is our prior choice of the joint CDF of β and λ rather than the sampling distribution itself. Sethuraman (1994) showed that a realization G from the DP has the form

$$G(\beta, \lambda) = \sum_{h=1}^{\infty} \pi_h I(\beta_h^* \leq \beta, \lambda_h^* \leq \lambda),\ (\beta_h^*, \lambda_h^*) \overset{i.i.d.}{\sim} G_0. \tag{4.7}$$

Here $\pi_1 = V_1$ while the $\pi_h = V_h \prod_{l<h}(1 - V_h), h = 2, 3, \cdots$ are probability weights that are formed from a stick-breaking process with $V_h \overset{i.i.d.}{\sim} Beta(\cdot; 1, v)$, for $h = 1, 2, \cdots$ and $I(A)$ being 1 or 0 according to whether or not A is true. The representation (4.7) is often called the stick-breaking representation of DP. The stick-breaking representation shows that this model essentially assumes that the T_k for each species is sampled from a mixture of infinitely many Weibull distributions

$$F(t) = \sum_{h=1}^{\infty} \pi_h Weibull(t; \beta_h, \lambda_h), \ (\beta_h^*, \lambda_h^*) \overset{i.i.d.}{\sim} G_0. \tag{4.8}$$

Given a single realization of F from its posterior distribution, say $F|data$, we could invert it to return the posterior α^{th} quantile of interest. As we want to approximate the posterior distribution of $\eta_{k\alpha}$, we must obtain multiple copies of $F|data$. For this purpose, we apply the density estimation scheme introduced by Ishwaran and Zarepour (2000) in the context of DP mixture of normal distributions. We will design our MCMC algorithm to return B copies of the posterior CDF of the Weibull parameters, $G^{(b)}$ $(b = 1, \cdots, B)$, using the stick breaking representation of DP (4.7). For each approximated posterior sample $G^{(b)}$, we approximate the posterior of (4.8). We note that sampled posterior CDFs could return the interval estimates of the posterior density of T as a by–product of the semiparametric approach by evaluating each $G^{(b)}$ for a grid of values over the values that T may take.

Various authors have proposed methods of fitting Dirichlet process mixed models, which yields approximate inference for densities of the random mixtures (see for example, Kottas and Gelfand (2002) and Escobar and West (1995)). However, the method of Ishwaran and Zarepour (2000) described above has an advantage of its simplicity. We turn now to a detailed description of the hierarchy of the Weibull DP mixture.

The Weibull DP mixture model

The semiparametric model above assumes the measurable determinant of reliability T is from an infinite mixture of Weibull distributions, where the shape β and scale λ parameters of each Weibull distribution is from the base distribution G_0. We now take the joint CDF G_0 to be

$$G_0(\beta, \lambda) = Unif(\beta; 0.01, \phi) Unif(\lambda; 0.01, \gamma). \tag{4.9}$$

where $Unif(\cdot; a, b)$ is the CDF of the uniform distribution with mean $(b+a)/2$ $(a < b)$. We allow the parameters ϕ and γ to be random by adding additional layers to the hierarchy: $\phi \sim Pareto(\phi; a_\phi, l_\phi)$, and $\gamma \sim Pareto(\gamma; a_\gamma, l_\gamma)$, where $Pareto(\cdot; a, b)$ is the CDF of the Pareto distribution with mean $ab/(b-1)$ (if $a > 1$). As l_ϕ (or l_γ) restricts the support of ϕ (or γ) to be greater than l_ϕ (or l_γ), the hyperparameter l_ϕ (or l_γ) can be interpreted as lowest upper bounds for β (or λ). If one has an idea of the largest values of the β (or λ), the mixture components of the distribution of T, they should be selected as l_ϕ (or l_γ). We choose $l_\phi = l_\gamma = 10$. Our sensitivity analysis found that the resulting 95 % credible interval of $\eta_{0.05}$ is robust to the selection of l_ϕ and l_γ (Section 5).

The hyperparameter a_ϕ (or a_γ) is negatively associated with the variance of ϕ (or γ) if a_ϕ (or a_γ) > 2, and their variance become infinity if a_ϕ (or a_γ) ≤ 2. With little specific information available to us, we have conservatively chosen $a_\gamma = a_\phi = 2$, although users would have the option of increasing this value.

Kottas (2006), who proposed the Weibull DP mixture model in the context of survival analysis, reparametrized the scale parameter of the Weibull distribution as $\lambda^* = \lambda^\beta$ so that λ^* has an inverse gamma as a conjugate prior. He then took G_0 to be the product of the CDFs for the uniform and inverse gamma distributions respectively, on β and λ^*. An additional layer in the hierarchy was added for the scale parameter of the inverse gamma distribution as it has a gamma conjugate prior. Although this hierarchical model gains a computational advantage from the use of a conjugate prior, we elected not to select this hierarchical model for the following reasons. (1) As the MOR can range between 0 to 20 (1000 psi), the sampler tends to return fairly large values of λ^*, which can lead to computational instability. (2) We found specifying the hyperparameters in Kottas (2006) challenging in our application and the resulting density estimates of T, sensitive to those specifications. The need for interpretability of the hyperparameters guided our search for a hierarchical model and we hope we have been somewhat successful. Our empirical studies indicate that results obtained by applying our model are fairly robust against misspecification of those hyperparameters.

We also allow the precision parameter v to be random with $Unif(v; 0.01, 5)$ as a prior distribution. Although a conjugate gamma prior is available, we have chosen instead a uniform distribution. This is because we found that the shape of the estimated density of T was rather sensible to the selection of the hyperparameters of the gamma distribution.

Dropping the species subscript k for notational simplicity the full hierarchical model of our Weibull DP mixture is, in summary,

$$\begin{aligned}
T_j|\beta_j, \lambda_j &\overset{ind}{\sim} Weibull(\cdot; \beta_j, \lambda_j), j = 1, \cdots, m \\
\beta_j, \lambda_j|G &\overset{i.i.d.}{\sim} G, j = 1, \cdots, m \\
G|v, \phi, \gamma &\sim DP(v, G_0) \text{ where } G_0 \text{ is defined in (4.9)} \\
v &\sim Unif(0.01, 5) \\
\phi &\sim Pareto(\cdot; a_\phi, l_\phi) \\
\gamma &\sim Pareto(\cdot; a_\gamma, l_\gamma).
\end{aligned}$$

A posterior computation

For posterior inference, Kottas (2006) used a collapsed Gibbs sampler based on the Polya urn representation of DP (Blackwell and MacQueen (1973)), which avoids updating the infinitely many parameters characterizing the CDF of T. However, this sampler does not return a posterior sample of the joint distribution G. This is because it marginalizes out G and samples directly from the posterior distribution of (β_j, λ_j). As we need a posterior sample of Gs, we will instead use a blocked Gibbs sampler (Ishwaran and James (2001)), which approximates G at every iteration. The algorithm is based on the stick-breaking representation of DP (4.7) and it relies on approximating G through a truncation of the stick-breaking representation in (4.7) at some large value M. Since the probability weight π_h decreases as the index h increases, it is reasonable to truncate the infinite sum after the first M terms, by letting $V_M = 1$. The pragmatic choice of M and hence approximation of DP that is *close enough*, would be one that assigns an amount of probability to the final mass point $\pi_M = 1 - \sum_{h=1}^{M-1} \pi_h$ that is less than some small number ϵ. As the stick breaking representation leads to $E(\pi_M|v) = (v/(v+1))^{M-1}$, setting it equal to ϵ returns the choice of

$$M = 1 + \frac{\log \epsilon}{\log (v/(1+v))}. \tag{4.10}$$

Since we assume that v is random with an uniform prior, we choose the M that satisfies (4.10) with the largest possible value of v and $\epsilon = 0.01$, which leads to $M = 26$.

For computational purposes, we introduce latent variables $S_j, j = 1, \cdots, m$, such that $S_j = h$ when $(\beta_j, \lambda_j) = (\beta_h^*, \lambda_h^*)$. Then $P(S_j = h) = \pi_h$ and the joint probability function of the target distribution is expressed as

$$\begin{aligned} & P(\{\beta_h^*\}_{h=1}^M, \{\lambda_h^*\}_{h=1}^M, \{S_j\}_{j=1}^m, \{V_h\}_{h=1}^M, \phi, \gamma, v|data) \\ & \propto P(\phi)P(\gamma)P(v) \prod_{h=1}^M \Big\{ P(V_h|D)P(\lambda_h^*|\gamma)P(\beta_h^*;\phi) \Big\} \\ & \times \prod_{j=1}^m \Big\{ P(S_j;\{V_h\}_{h=1}^M)P(t_j;\beta_{s_j}^*,\lambda_{s_j}^*) \Big\}. \end{aligned} \tag{4.11}$$

Figure 1 shows the directed acyclical graph (DAG) of our hierarchical model. The Appendix gives a detailed description of our MCMC sampling algorithm. In the algorithm, each iteration of MCMC sampling generates an approximation of G which is a function of $\{\beta_h^*\}_{h=1}^M$,$\{\lambda_h^*\}_{h=1}^M$, and $\{V_h\}_{h=1}^M$. Given these posterior samples, we obtain B samples of posterior predictive distribution of T

$$F^{(b)}(t|data) \approx \sum_{h=1}^M \pi_h^{(b)} Weibull(t|\beta_h^{(b)}, \lambda_h^{(b)}), \ \ b = 1, \cdots, B$$

Once the MCMC has run on the datasets for each species separately, we obtain the posterior samples of (4.7) for each k, $\{\{F_k^{(b)}\}_{b=1}^B\}_{k=1}^K$, and hence $\{\{\eta_{k,\alpha}^{(b)}\}_{b=1}^B\}_{k=1}^K$. As the $\{T_{kj}\}$ are independent across species, Equation (2.1) can be estimated as

$$P(\tau = k|data) \simeq \frac{\sum_{b_1,\cdots,b_K=1}^B \prod_{i\neq k} I(\eta_{k,\alpha}^{(b_k)} < \eta_{i,\alpha}^{(b_i)})}{B^K}.$$

5 An illustrative application

This section demonstrates use of the nonparametric and semiparametric methods developed in previous sections on three real datasets of size 282, 98 and 174,collected by the

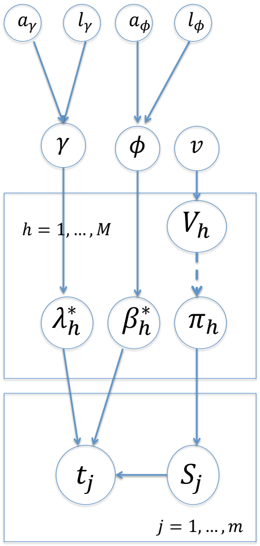
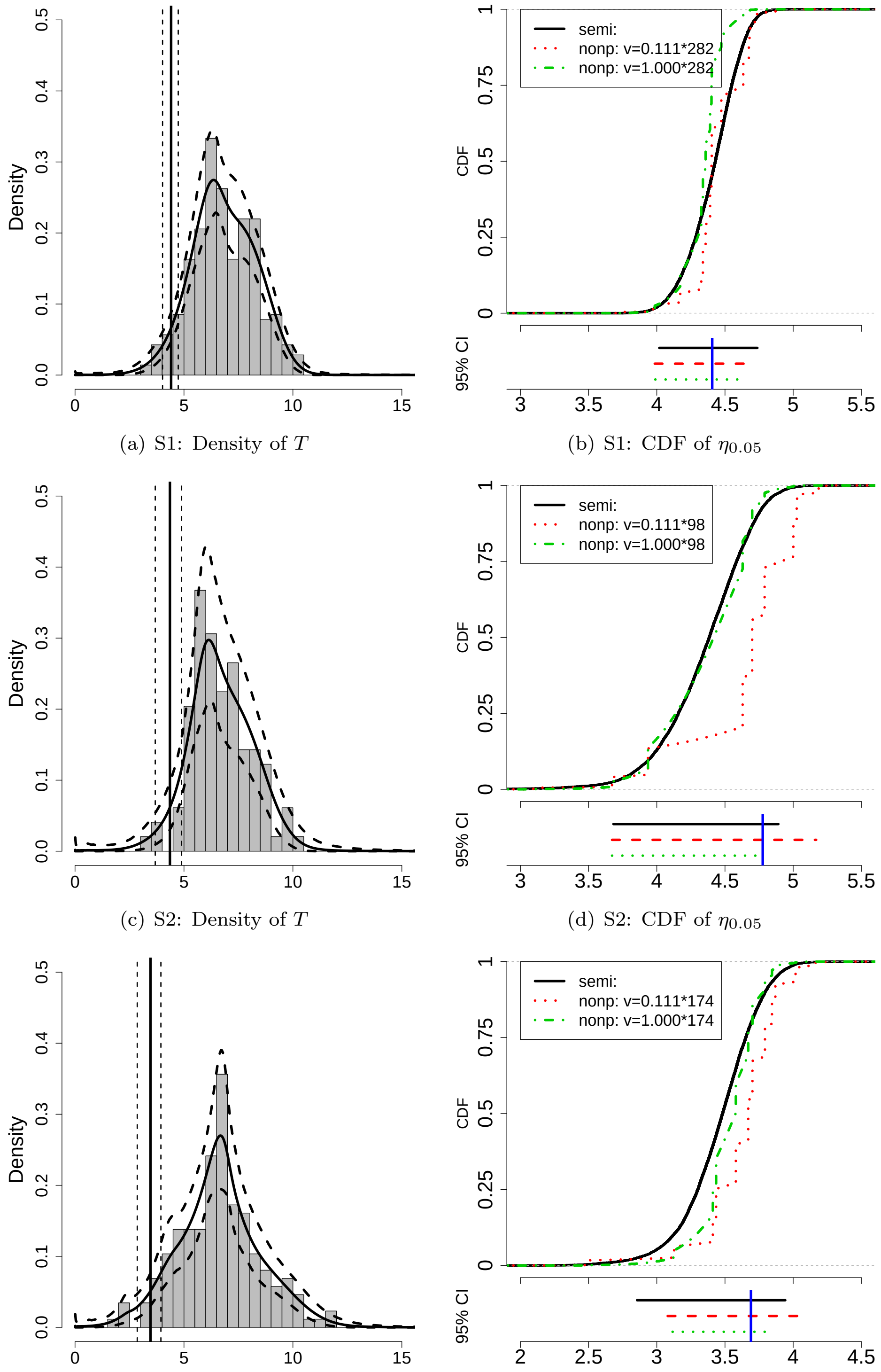


Figure 1: The directed acyclical graph of the semiparametric Weibull DP mixture model. Unbroken arrows represent stochastic relationships, while broken arrows represent deterministic relationships.

Forest Products the Forest Products Stochastic Modelling Group where the measured strength is the *modulus of rupture (MOR)* for lumber from a single species. In the laboratory, the MOR of a randomly selected lumber specimen is measured by gradually increasing the stress on the board and recording the stress level at which the board breaks. For convenience, we refer to the three datasets as S1, S2 and S3. For the semiparametric method, the MCMC sample size is set to 10,000 after discarding the first 5,000 as burn-in and thinning at every 5^{th} iteration. For the the nonparametric method, MC approximation of (3.5) was done with a sample of size 2000 ($= L$). To assess sensitivity to the prior information, the parameter v of the nonparametric model was set to $v = 0.111m$, and m was chosen so that 10%, and 50 % of weights are put on the prior, respectively. We implemented our semiparametric and nonparametric methods in an R package DPw, which is publicly available at the CRAN repository: `http://cran.r-project.org`. For computational efficiency, the MCMC of semiparametric model is coded in C language.

The left hand panels of Figure 4 show the estimated density of MOR given by the semiparametric method for each dataset. The histograms of sampled MORs are superimposed. Although all three datasets contain MORs, the shape of the estimated density from the semiparametric method look quite different; those for S1 and S2 indicate the possibility of two modes whereas that for S3 is unimodal and resembles a Laplace distri-



(a) S1: Density of T

(b) S1: CDF of $\eta_{0.05}$

(c) S2: Density of T

(d) S2: CDF of $\eta_{0.05}$

(e) S3: Density of T

(f) S3: CDF of $\eta_{0.05}$

Figure 2: The left hand panels show the estimated density of MOR from the semi-parametric method. The histograms of sampled MORs are superimposed. The right hand panels show the estimated posterior CDF of the fifth percentile, $\eta_{0.05}$ from the semi and non parametric methods (Top), and the empirical 95 % credible intervals of $\eta_{0.05}$ (Bottom). The blue horizontal line represents the sample 5^{th} percentile.

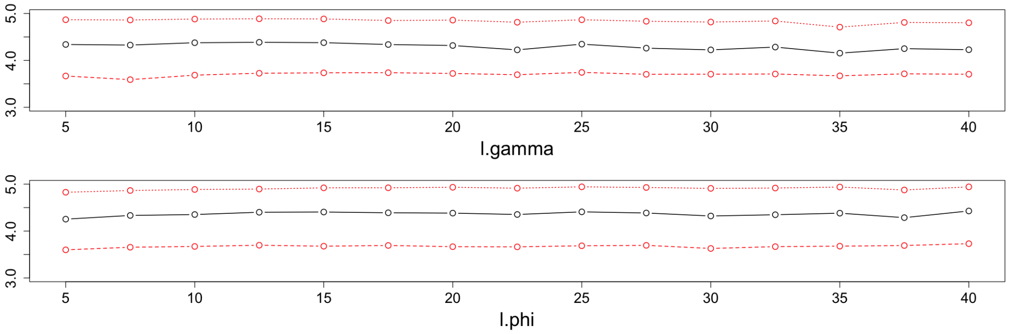


Figure 3: The sensitivity plots show the change in 2.5th percentile (dotted line), mean (solid line) and the 97.5th percentile (dotted line) of the estimated distribution of $\eta_{0.05}$ with different values of l_γ and l_ϕ.

bution. This lends support for nonparametric and semiparametric approaches like those taken in D1990. The right hand panels show the estimated posterior CDF of the fifth percentile, $\eta_{0.05}$ from the semi and non parametric methods (Top), and the empirical 95 % credible intervals of $\eta_{0.05}$ (Bottom). The credible intervals from the nonparametric model with small v $(= 0.111m)$ tends to be larger than the other methods, and the interval tends to lie around the sample's fifth percentile. As well the estimated CDF of the semiparametric method tends resemble more closely, the nonparametric method with $v = m$. These observations are reasonable as v is the weight put on the prior for the Weibull distribution's parameters, relative to the empirical distribution.

S2 exhibits the largest discrepancies in the estimated CDF of $\eta_{0.05}$ among the competing methods. This may be because of the small sample size (98). When the sample size is small, the estimated CDF from the nonparametric method with small v show clear jumps at every observed t_js, which could be unrealistic as a distribution of $\eta_{0.05}$.

Lastly, we performed the sensitivity analysis of the estimated distribution of η_α with respect to the change in l_γ and l_ϕ using S2, which has the smallest sample size. Figure 3 shows that the change in 2.5th percentile (dotted line), mean (solid line) and the 97.5th percentile (dotted line) of the estimated distribution of $\eta_{0.05}$ with different values of l_γ and l_ϕ $(= 5, 7.5, \cdots 50)$ while fixing the other value to 10. We observe that all the estimated quantities are unaffected by the change in the value of l_γ and l_ϕ in the reasonable range. From this experiment, we conclude that the estimated distribution of η_α is robust to the specifications of the parameter l_γ and l_ϕ.

6 Performance assessment

This section presents the results of simulation studies designed to compare the performance of the three subset selection methods: the current ASTM industrial standard described in Introduction; the nonparametric method (Section 3); the proposed semiparametric method (Section 4).

Our proprietary datasets for a single grade consist of less than ten species, and the gaps in percent between successively larger fifth percentiles of the samples of the first seven species are: 7, 3, 8, 1, 17, 3. We will design our simulation model based on this information, the three datasets of MOR analyzed in Section 5 and available information of the sampling distribution of MOR from literature. We generate the quality measures T_k $(k = 1, \cdots, 7)$ under three settings;

- Setting 1: The samples of species 1 - 7 are generated from:
    - 1) $Weibull(\cdot; 3.85, 7.32)$ with $\eta_{0.05} = 3.38$;
    - 2) $Lnorm(\cdot; 1.85, 0.31)$ with $\eta_{0.05} = 3.82$;
    - 3) $Weibull(\cdot; 5.19, 7.27)$ with $\eta_{0.05} = 4.10$;
    - 4) $0.67Lnorm(\cdot; 1.98, 0.17)$+$0.33Lnorm(\cdot; 1.74, 0.23)$ with $\eta_{0.05} = 4.47$;
    - 5) $0.79Weibull(\cdot; 5.43, 7.64)$+$0.21Weibull(\cdot; 12.01, 6.19)$ with $\eta_{0.05} = 4.53$;
    - 6) $0.74Weibull(\cdot; 5.49, 7.60)$+$0.26Weibull(\cdot; 15.81, 5.98)$ with $\eta_{0.05} = 4.61$;
    - 7) $0.98Lnorm(\cdot; 1.90, 0.19)$ +$0.02Lnorm(\cdot; 1.25, 0.10)$ with $\eta_{0.05} = 4.69$,

  where $Lnorm(\cdot; a, b)$ represents the log-normal CDF with mean $e^{a+b^2/2}$. The simulation models for species 1 and 2 use MLEs based on S1 with corresponding kernel, models for species 3, 6 and 7 use MLEs based on S2 and the models for species 4 and 5 use MLEs based on S3. These distributions are selected based on the work of Liu (2010) which analyzed datasets S1 and S2 in the context of quantile estimation and fitted various parametric models using the ML method. Based on his estimated distributions, and additional analysis of S3, we selected the estimated distributions as simulation models so that the percentage gaps between the estimated quantiles reflect the ones in our proprietary data.

- Setting 2: The species are from the Weibull distribution with shape = 4.726 and scale parameters varying 10.803, 11.56, 11.906, 12.858, 12.988, 15.194, 15.651 so that the percentage gaps between the estimated quantiles reflect the ones in our proprietary data. These values are selected to be close to the MLEs of the MOR datasets reported in Johnson et al. (1999a).

- Setting 3: The species are from the standard two–parameter Weibull distribution with shape = 4.726 and scale parameters of 10.803 for two species and 11.906 for the rest. This is the scenario when there are two species that have the smallest fifth percentile.

For each setting samples of size 100 as well as 360 (= m_k for all k) are generated 300 times. Then the ASTM procedure, the semiparametric and the nonparametric models are fit to each set of data to estimate the probability that the k^{th} species is the weakest for each $k = 1, \cdots, 7$. To assess sensitivity of the prior specifications, the parameter

		Subset size		% capture the weakest		% Set stays the same	
m		100	360	100	360	100	360
	ASTM	5.92	2.85	100	100	12	74
P^*=95	NP1	3.13	1.71	100	100	67	98
	NP2	2.41	1.93	98	97	86	100
	SP	2.91	1.50	100	100	82	100
P^*=99	NP1	4.79	2.35	100	100	29	76
	NP2	3.91	2.46	100	100	34	83
	SP	4.61	1.88	100	100	50	99
P^*=99.5	NP1	5.35	2.99	100	100	18	57
	NP2	4.55	3.22	100	100	25	61
	SP	5.39	2.07	100	100	27	99

Table 1: Setting 1

v of the nonparametric model is set to $v = 0.111m_k$, and m_k. All the rest of the hyperparameter values of the non– and semiparametric models are selected as in the previous section.

One desideratum for a method of specifying design values through the species grouping approach pertains to the withdrawal of a species that is not in the subset of controlling species (CS). Ideally this should not change the design value that was originally computed solely on the basis of the CS in order to ensure continuity in published standards. However a poor method could mean that when it was reapplied to the remaining, K -1 species, the CS could change along with the design value. In fact, counterintuitively, it could actually increase. Thus stability in the CS itself would be desirable and that performance property is one of the ones we now explore along with the size of the CS and the success rate in capturing the species with the smallest fifth percentile in the CS. For our analysis, we remove the sample from the species with the highest true fifth percentile, reapply the subset selection procedures and examine the result. For the setting 3, where five species have the same strongest true fifth percentile, the seventh species is removed as the strongest species.

6.1 Result

Tables 1,2 and 3 show the average subset size, proportion of times that weakest species are captured and the porportion of times the subset selected using seven species is the same as the subset selected using six species without the strongest one over the 300 simulations. In setting 3, where two species are the weakest, we counted the times when both of the two species are contained in CS. NP1 and NP2 represent the nonparametric Bayesian procedure with $v = 0.111m$ and m respectively. For both of the sample sizes considered, all the competing subset selection procedures include τ almost all the time in their CSs when τ is unique among the 7 species. The main findings follow:

- **Comparision of nonparametric & semiparametric methods**
  The subset size of Bayesian procedures increases with P^*, the required posterior

		Subset size		% capture the weakest		% Set stays the same	
m		100	360	100	360	100	360
	ASTM	6.46	4.21	100	100	11	64
P^*=95	NP1	3.83	2.53	99	100	69	87
	NP2	3.10	2.04	99	100	84	96
	SP	3.90	2.14	99	100	75	100
P^*=99	NP1	5.37	3.72	100	100	34	57
	NP2	4.51	2.96	100	100	49	67
	SP	5.42	2.86	99	100	44	100
P^*=99.5	NP1	5.78	4.32	100	100	22	35
	NP2	5.03	3.54	100	100	30	49
	SP	5.94	3.16	100	100	32	99

Table 2: Setting 2

		Subset size		% capture the weakests		% Set stays the same	
m		100	360	100	360	100	360
	ASTM	6.95	6.64	100	100	1	5
P^*=95	NP1	4.84	3.83	82	89	26	47
	NP2	4.38	2.97	83	92	34	73
	SP	4.91	3.02	90	95	28	73
P^*=99	NP1	6.15	5.51	93	99	10	17
	NP2	5.84	4.48	95	98	11	29
	SP	6.30	4.49	98	99	12	43
P^*=99.5	NP1	6.41	5.93	95	99	6	9
	NP2	6.17	5.06	97	100	6	18
	SP	6.66	5.06	99	100	5	35

Table 3: Setting 3

probability of subset of controlling species contains the one with smallest percentile. Ideally the size of the subset of controlled species (CS) should be small. When the sample sizes are small, the nonparametric method with $v = m$ (denoted as NP2) yields the smallest CS in all settings. However, as sample size increases to 360, the semiparametric procedure returns as small or smaller CS than nonparametric procedures. The Bayesian approaches considered in this paper, yield for each subpopulation index $k \in \{1, \dots, 7\}$ a posterior probability that k is the one with the smallest fifth percentile. That probability should be highest when $k = \tau$, the index of the species corresponding to that smallest value. With sample sizes of 360, a current standard size in lumber testing experiments, the semiparametric method proves slightly superior when τ is unique among the 7 species (i.e., Settings 1 and 2) (See Figure 4 in Appendix 2).

Removal of the strongest species affects the choice of CS more when P^* is large, in agreement with intuition as a large P^* entails a larger CS, which in turn is more likely to capture τ. When the sample size is large (360) and there is a unique weakest species (Settings 1 and 2), the semiparametric method returns very stable CS for all values of P^* and all settings, under strongest species removal scenario whereas the nonparametric procedure returns relatively unstable CS for large P^*. This could be because when the sample size is large, the semiparametric method tends to return the smallest CS among other methods, which is less likely to capture the strongest species. Although the $\eta_{0.05}$ estimates generated by the semiparametric model are biased when the species are not from the mixture of Weibull distributions (as discussed in Appendix 2), this property does not affect the stability of the CS and the semi-parametric model returns the most stable subset in Settings 1 and 2.

- **Comparison of the ASTM and Bayesian methods**
  When sample size is small, the ASTM procedure tend to return a larger CS than any of the Bayesian methods for any of the P^*s we considered. As a consequence, the CS from ASTM tends to be more unstable under the subset withdrawal senario than Bayesian procedures. When sample size is large, we observe that the semiparametric procedure always outperforms the ASTM procedure in terms of stability of CS. In particular, in the presence of two species with small fifth percentiles and five with higher ones (Setting 3), the ASTM tends to return a conservative CS, meaning that it selects all the species in the subset even when the sample size is large ($m = 360$). This ASTM's tendency to select conservatively large CSs, may lead to the greater instability we observe in our empirical findings under species removal scenarios in our simulation studies, which are broadly in line with those in Taylor et al. (2008)

7 Concluding remarks

Our empirical assessments based on the simulation studies, suggest both Bayesian methods for subset selection presented in this paper have promise as methods for selecting subsets of controlling species (CSs), in that the CSs tend to be more stable than the current standard procedure of ASTM under withdrawal of a species. These studies also show that the nonparametric empirical Bayes procedure works reasonably well as a competitor to the fully Bayesian semiparametric method. The former has the advantage of simplicity and low computational cost. On the other hand, care must be in tuning the precision parameter v (the weight placed on the base distribution in the Dirichlet process prior) in applying the method. That is because, not surprisingly and as the simulation studies show, that approach's performance degrades when the true distribution does not coincide with the base distribution assumption and a relatively large value of v is assigned. This may limit its value in the formulation of species grouping protocols, where exact algorithms need to be given. The simulation study also found its competitor, the semiparametric approach returns the most stable subsets for reasonably large sample size even though the estimated fifth percentile $\eta_{0.05}$ is biased when the true distribution is non-Weibull.

ASTM D1990 accepts the sample size of a single specie larger than 100, however our simulation study showed that 100 could be too small to develop a stable CS based on the current ASTM procedure. Our proposed semiparametric procedure can return a stable CS with such a small sample size, if one allows relatively small acceptable probability P^* (e.g. 95 %). For our non-parametric procedure to return a stable CS with a small sample size, strong prior belief (i.e. large v) as well as large P^* are necessary.

Overall the semiparametric approach enjoys a number of advantages over the nonparametric empirical Bayes approach in addition to those mentioned above. First it is fully Bayesian, which assures coherence in any inferential findings derived from its use. Second, it provides a simple technical mechanism (the prior) for expert opinion to be introduced into the problem of species grouping. Work will be needed on how best to elicit that opinion. Second, unlike the nonparametric empirical Bayes procedure, tuning a precision parameter is unnecessary as our model's hierarchical structure means it incorporates the learning of the precision parameter. Third, the approach of necessity produces a sampler for the unknown population's posterior strength distribution. This is a major byproduct of the work which is being used in work now underway to characterize lot properties, i.e. metrics calculated for sets of pieces of lumber of fixed size, say 10 pieces. For example, the minimum strength of such a lot. Note that the method returns a a posterior piecewise credibility interval around the estimates of the density of T, which can be useful for analyzing the behaviour of T. Overall these results would offer a new approach to grading lumber based on lot properties rather than the strength properties of individual pieces.

While the approaches being developed, like those in this paper offer new ways of grouping species, it should be emphasized that they build on the idea of controlling species in a protocol in ASTM D1990 that has generally proven quite successful over time. So the new methods we have introduced are designed as refinements of that approach rather than as replacements. Moreover they would need to be compared to Van Eeden and Zidek (2012), an exercise that was initiated in Taylor et al. (2008) for just the ASTM and nonparametric Bayes procedure. Preliminary findings suggest those refinements may offer greater stability in the published design values under the addition and withdrawal of species than the current method while those in this paper.

Acknowledgement. Constance van Eeden reviewed and commented on this manuscript, as well as suggested a number of references. We are indebted to Conroy Lum from FPInnovations for introducing the second author to the topic addressed in this report and for many helpful discussions during the course of the work. Thanks also to David Dunson from Duke University for acquainted the second author with the value of hierarchical mixture model approach used in this report as well as to Alexander Bouchard-Côte for helpful discussions during the course of the work reported in the paper. We appreciate the assistance of Dr Davor Cubranic in the development of the CRAN R package that complements this paper.

Appendix

1 Details of the MCMC procedure

This appendix presents the algorithm used to implement the method in Section 4 and sample from the the posterior distribution (4.11). Repeat the following MCMC with seven steps B times to obtain posterior sample of size B.

- Update S_j from a closed form multinomial conditional posterior with probability $P(S_j = h|\{V_h\}_{h=1}^{M-1}, \lambda_h^*, \beta_h^*) \propto V_h \prod_{l<h}(1-V_l) Weibull(t_j; \beta_h^*, \lambda_h^*)$ for $j = 1, \cdots, m$
- Update $(V_h|\{S_k\}_{k=1}^m, v) \sim Beta(V_h; 1 + \sum_{j=1}^m I(S_j = h), v + \sum_{j=1}^m I(S_j > h))$ for $h = 1, \cdots, M-1$.
- Update $(\lambda_h^*|\gamma, \beta_h^*, \{t_j\}_{j;S_j=h})$ via Metropolis Hasting (MH) algorithm with a normal proposal distribution for $h = 1, \cdots, M$.
- Update $(\beta_h^*|\phi, \lambda_h^*, \{t_j\}_{j;S_j=h})$ via MH algorithm with a normal proposal distribution for $h = 1, \cdots, M$.
- Update $(\phi|\{\beta_h^*\}_{h=1}^M) \sim Pareto(\phi; a_\phi + M, max\{l_\phi, \beta_1^*, \cdots, \beta_M^*\})$.
- Update $(\gamma|\{\lambda_h^*\}_{h=1}^M) \sim Pareto(\phi; a_\gamma + M, max\{l_\gamma, \lambda_1^*, \cdots, \lambda_M^*\})$.
- Update $(v|\{V_h\}_{h=1}^{M-1})$ via MH algorithm with a normal proposal distribution.

2 More details of the simulation studies

Figure 4 shows the boxplots of the estimates of the posterior probabilities that the species k to be the weakest from the Bayesian procedures ($k = 1, \cdots, 7$). The boxplots of the weakest species and the second weakest species are highly variable in all the simulation settings. However, as the sample size increases to the current industrial standard of 360, the boxplots of the true weakest species shift upward and the boxplots of the specie with the second weakest species shift downward; the methods more confidently select the target specie as the one with the smallest 5^{th} quantiles. When sample size is large (360), the weakest species' boxplot from semiparametric method lie slightly higher than the others in all settings.

For the semiparametric method, Figure 5 shows boxplots of the estimates of fifth quantiles over 100 simulations. For Weibull mixture species (Species 5 and 6 in Setting 1), the quantiles are slightly under estimated by semiparametric method. It shows that when the true distributions are log-normal (Species 2, 4, and 7 in Setting 1), the estimates are negatively biased; the interquartile range does not cover the true 5^{th} quantiles. For single Weibull species, the semiparametric method returns unbiased estiamtes. However, the selected subsets captures the weakest species more often/as often as the nonparametric model in Setting 1.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2(6): pp. 1152–1174. URL `http://www.jstor.org/stable/2958336`

ASTM Standard D1990 (2007). "Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens." Technical Report DOI: 10.1520/D1990-07, ASTM International, DOI: 10.1520/D1990-07.

ASTM Standard D2915 (2011). "Standard Practice for Sampling and Data-Analysis for Structural Wood and Wood- Based Products." Technical Report DOI: 10.1520/D2915-10, ASTM International, West Conshohocken, PA.

Berger, J. and Deely, J. (1988). "A Bayesian Approach to Ranking and Selection of Related Means with Alternatives to Analysis-of-Variance Methodology." *Journal of the American Statistical Association*, 364–373.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions via Polya Urn Schemes." *Ann. Stat.*, 1: 353–355.

Broderick, T., Jordan, M., and Pitman, J. (2011). "Beta Processes, Stick-Breaking, and Power Laws." *Arxiv preprint arXiv:1106.0539*.

Escobar, M. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of American Statistical Association*, 90: 577–588.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problem." *Annals of Statistics*, 1: 209–230.

— (1974). "Prior Distribution on Spaces of Probability Measures." *Annals of Statistics*, 2: 615–629.

Fong, K. and Berger, J. (1993). “Ranking, Estimation and Hypothesis Testing in Unbalanced Models – a Bayesian approach.” *Statistics and Decisions*, 11: 1–24.

Ishwaran, H. and James, L. F. (2001). “Gibbs Sampling Methods for Stick-Breaking Priors.” *Journal of the American Statistical Association*, 96(453): 161–173.

Ishwaran, H. and Zarepour, M. (2000). “Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models.” *Biometrika*, 87: 371–390.

Johnson, R., Evans, J., and Green, D. (1999a). “Nonparametric Bayesian Predictive Distributions for Future Order Statistics.” *Statistics & probability letters*, 41: 247–254.

— (1999b). “Some Bivariate Distributions for Modeling the Strength Properties of Lumber.” Research Paper FPL-RP-575., USDA Forest Service.

Kottas, A. (2006). “Nonparametric Bayesian Survival Analysis using Mixtures of Weibull Distribution.” *Journal of Statistical Planning and Inference*, 136(3): 578–596.

Kottas, A. and Gelfand, A. E. (2002). “A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models.” *Journal of Computational and Graphical Statistics*, 1(2): 289–305.

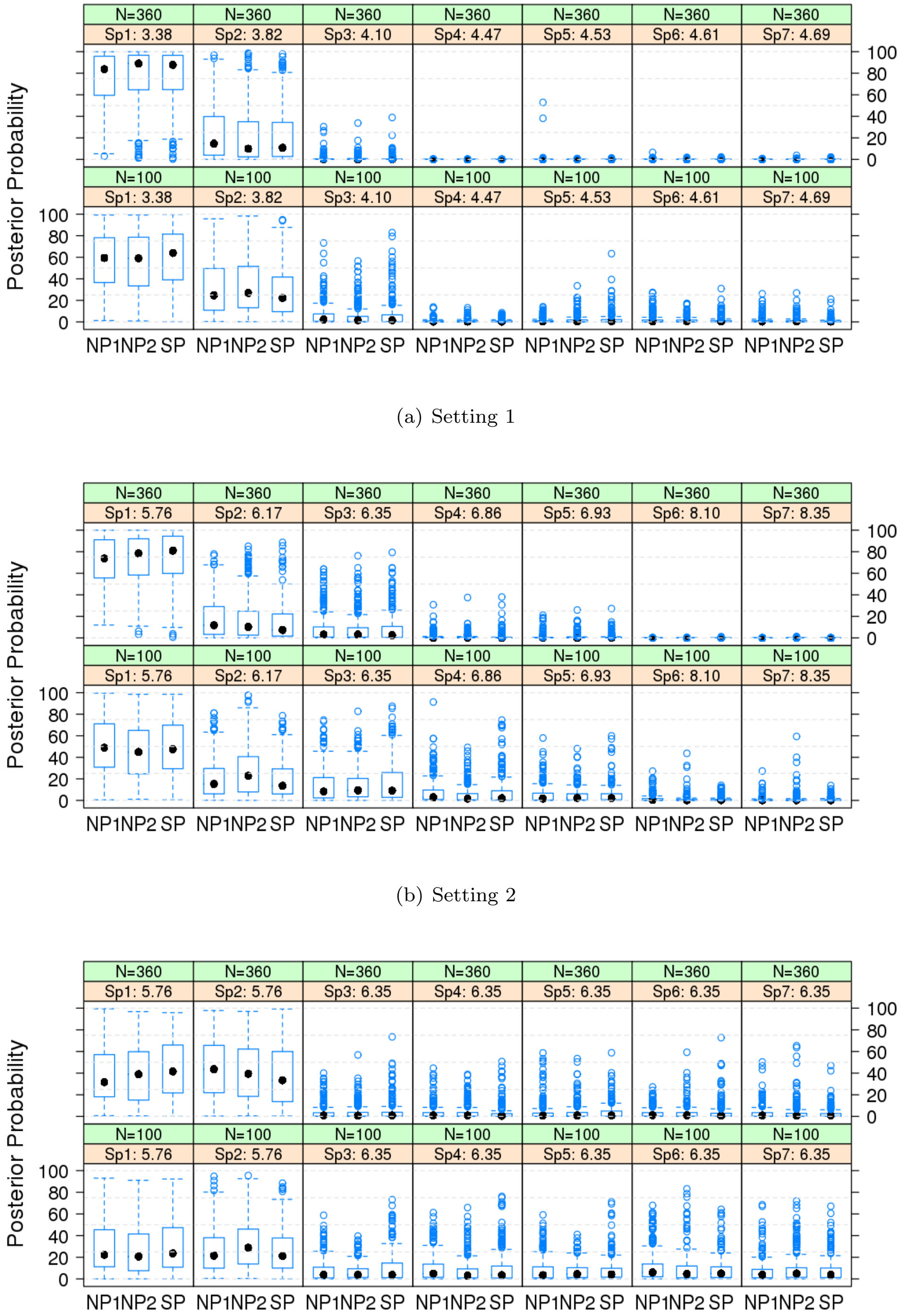
Liu, Y. (2010). “Lower Quantile Estimation of Wood Strength Data.” Master’s thesis, University of British Columbia.

Sethuraman, J. (1994). “A constructive definition of Dirichlet priors.” *Statistica Sinica*, (4): 639–650.

Taylor, C., Van Eeden, C., and Zidek, J. (2008). “Selecting the Controlling Subset of Lumber Species in Setting Grade Standards.”

Van Eeden, C. and Zidek, J. (2012). “Subset selection – Extended Rizvi–Sobel for Unequal Sample Sizes and its Implementation.” *J Nonparametric Statist*, (DOI:10.1080/10485252.2012.660482).

]

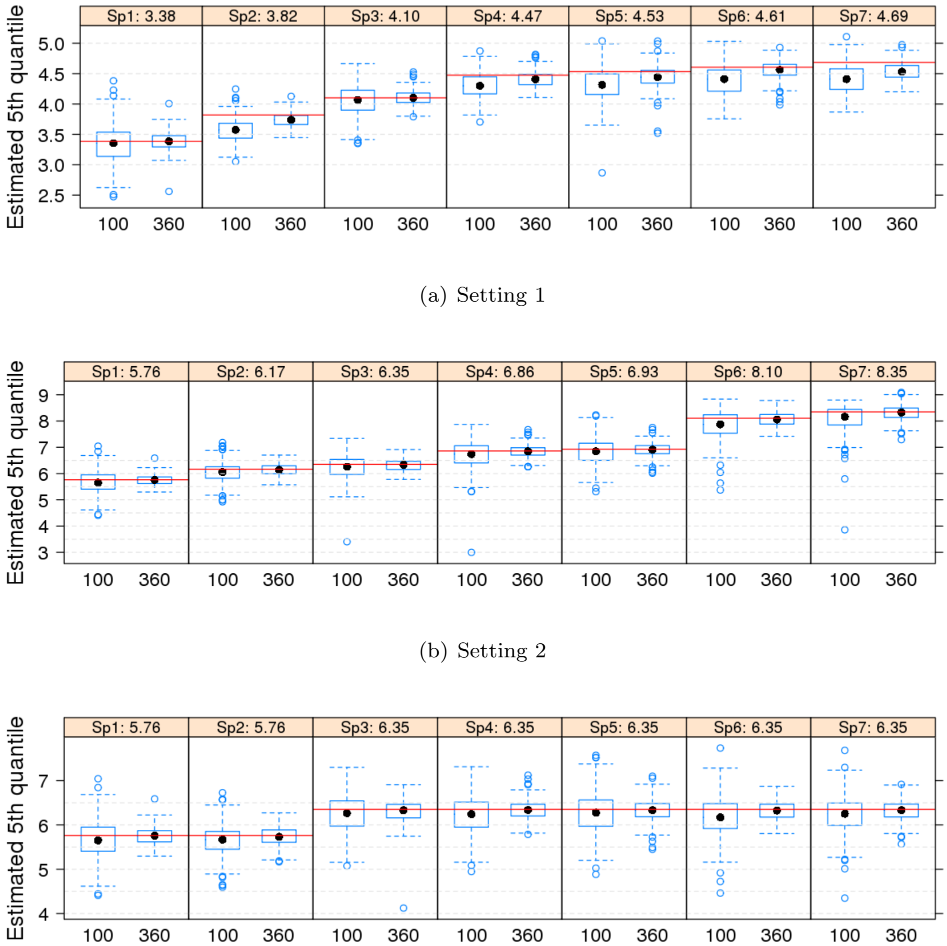


(a) Setting 1

(b) Setting 2

(c) Setting 3

Figure 4: The extimated posterior probabilities that spcecie k has the smallest 5th quantile.



(a) Setting 1

(b) Setting 2

(c) Setting 3

Figure 5: The estimated $\eta_{0.05}$ from the semi-parametric model. The horizontal lines represent the true $\eta_{0.05}$.