Bayesian nonparametric subset selection procedures with Weibull components.

Yumi Kondo, James V. Zidek

Statistics Department
University of British Columbia

May 27, 2013
Because of species similarities or marketing convenience, it is desirable to combine two or more species into a single marketing group. When this is done, it is necessary to determine the design value, for the combined group of species (ASTM) D1990.

Similar structural properties $\Rightarrow$ Single Marketing group
Design Value of a marketing group

The breaking strength of a randomly selected wood specimen in the marketing group must exceed design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)
Design Value of a marketing group

The breaking strength of a randomly selected wood specimen in the marketing group must exceed design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed **design value** with probability above 95%.

How should the **design value** be found? **“Species grouping”** (ASTM D1990)

Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed **design value** with probability above 95%.

How should the **design value** be found? “Species grouping” (ASTM D1990)

Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed **design value** with probability above 95%.

How should the **design value** be found? “**Species grouping**” (ASTM D1990)

Find subset of controlling species (CS)

Combine their samples to get:
- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed the design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)

Combine their samples to get:
- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed **design value** with probability above 95%.

How should the **design value** be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**Design Value of a marketing group**

The breaking strength of a randomly selected wood specimen in the marketing group must exceed design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)

Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
Design Value of a marketing group

The breaking strength of a randomly selected wood specimen in the marketing group must exceed design value with probability above 95%.

How should the design value be found? “Species grouping” (ASTM D1990)
Find subset of controlling species (CS)

Combine their samples to get:

- (1) conservative design values;
- (2) stability under species withdrawal and new entries.
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

\( H_0 \): equality of proportions of species below the combined populations’ 5\(^{th}\) percentile.

- **\( H_0 \) not rejected**: Treat all species as controlling– compute lower 5\% tolerance limit (TL) from combined sample.
- **\( H_0 \) rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
$H_0$: equality of proportions of species below the combined populations’ 5th percentile.

- $H_0$ not rejected: Treat all species as controlling—compute lower 5% tolerance limit (TL) from combined sample.
- $H_0$ rejected: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 

![Histogram of species breaking strength](image1)

![Histogram of species breaking strength](image2)
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

**$H_0$: equality of proportions of species below the combined populations’ 5$^{th}$ percentile.**

- **$H_0$ not rejected**: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.

- **$H_0$ rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 
ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)

\[ H_0: \text{equality of proportions of species below the combined populations’ 5}\text{th} \text{ percentile.} \]

- **\( H_0 \) not rejected**: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.
- **\( H_0 \) rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

\[ H_0: \text{equality of proportions of species below the combined populations’ 5}^{th} \text{ percentile.} \]

- **\( H_0\) not rejected**: Treat all species as controlling compute lower 5\% tolerance limit (TL) from combined sample.
- **\( H_0\) rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)

$H_0$: equality of proportions of species below the combined populations’ 5th percentile.

- $H_0$ not rejected: Treat all species as controlling compute lower 5% tolerance limit (TL) from combined sample.
- $H_0$ rejected: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 

not reject H0
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

\( H_0 \): equality of proportions of species below the combined populations’ 5\(^{th}\) percentile.

- **\( H_0 \) not rejected**: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.
- **\( H_0 \) rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

**$H_0$: equality of proportions of species below the combined populations’ 5\textsuperscript{th} percentile.**

- **$H_0$ not rejected**: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.

- **$H_0$ rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 
$H_0$: equality of proportions of species below the combined populations’ 5$^{th}$ percentile.

- $H_0$ not rejected: Treat all species as controlling—compute lower 5% tolerance limit (TL) from combined sample.
- $H_0$ rejected: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 

![Histogram of breaking strength](image)
**H₀**: equality of proportions of species below the combined populations’ 5<sup>th</sup> percentile.

- **H₀ not rejected**: Treat all species as controlling—compute lower 5% tolerance limit (TL) from combined sample.
- **H₀ rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects H₀.
ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)

\( H_0 \): equality of proportions of species below the combined populations’ 5\textsuperscript{th} percentile.

- \( H_0 \) not rejected: Treat all species as controlling– compute lower 5\% tolerance limit (TL) from combined sample.
- \( H_0 \) rejected: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)

$H_0$: equality of proportions of species below the combined populations’ $5^{th}$ percentile.

- $H_0$ not rejected: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.
- $H_0$ rejected: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects $H_0$. 
**ASTM D1990’s procedure (Sequential hypothesis testing) to find Controlling Species (CS)**

\( H_0 \): equality of proportions of species below the combined populations’ 5\(^{th}\) percentile.

- **\( H_0 \) not rejected**: Treat all species as controlling– compute lower 5% tolerance limit (TL) from combined sample.

- **\( H_0 \) rejected**: A stepwise pairwise comparison procedure, starting from the species with the smallest sample percentile until the test rejects \( H_0 \).
Bayesian subset selection: set new objective

Treat inferential objective as subset selection not sequential hypothesis testing.

Independent samples from $K$ species & $\tau = \text{species with smallest } 5^{th} \text{ percentile}$. Then find the smallest subset $S \subset \{1, \cdots, K\}$ such that $P(\tau \in S | \text{data}) \geq P^\ast$. e.g. $P^\ast = 0.99$

- $\pi_k \equiv P(\tau = k | \text{data})$ and $\pi(k)$ be the $k$th largest $\pi$. The optimum Bayesian choice of $S$ would be the subset corresponding to the smallest number $d$ of species with $P^\ast < \pi(K - d + 1) + \cdots + \pi(K - 1) + \pi(K)$

$$
\pi_k = \int_0^\infty \prod_{i \neq k} P(u < \eta_{i0.05} | \text{data}) dP(\eta_{k0.05} \leq u | \text{data}).
$$

where $\eta_{i0.05}$ is the $5^{th}$ percentile of breaking strength of $i^{th}$ species.

$\Rightarrow$ Solving our problem reduces to characterizing $P(\eta_{k0.05} \leq u | \text{data})$. 

Bayesian subset selection: set new objective

Treat inferential objective as subset selection not sequential hypothesis testing.

Independent samples from $K$ species & $\tau =$ species with smallest 5\textsuperscript{th} percentile. Then find the smallest subset $S \subset \{1, \cdots, K\}$ such that $P(\tau \in S|\text{data}) \geq P^*$. e.g. $P^* = 0.99$

- $\pi_k \equiv P(\tau = k|\text{data})$ and $\pi(k)$ be the $k$th largest $\pi$. The optimum Bayesian choice of $S$ would be the subset corresponding to the smallest number $d$ of species with $P^* < \pi(K-d+1) + \cdots + \pi(K-1) + \pi(K)$

$$\pi_k = \int_0^\infty \prod_{i \neq k} P(u < \eta_{i0.05} \mid \text{data})dP(\eta_{k0.05} \leq u \mid \text{data}).$$

where $\eta_{i0.05}$ is the 5\textsuperscript{th} percentile of breaking strength of $i$\textsuperscript{th} species.

$\Rightarrow$ Solving our problem reduces to characterizing $P(\eta_{k0.05} \leq u \mid \text{data})$.

Two proposed approaches:
- Nonparametric Bayesian approach
- Semi-parametric Bayesian approach
Bayesian subset selection: set new objective

Treat inferential objective as **subset selection** not sequential hypothesis testing.

Independent samples from $K$ species & $\tau$ = species with smallest $5^{th}$ percentile. Then **find the smallest subset** $S \subset \{1, \cdots, K\}$ such that $P(\tau \in S | data) \geq P^*$. e.g. $P^* = 0.99$

- $\pi_k = P(\tau = k | data)$ and $\pi(k)$ be the $k$th largest $\pi$. The optimum Bayesian choice of $S$ would be the subset corresponding to the smallest number $d$ of species with $P^* < \pi(K-d+1) + \cdots + \pi(K-1) + \pi(K)$

\[
\pi_k = \int_0^\infty \prod_{i \neq k} P(u < \eta_{i0.05} | data) dP(\eta_{k0.05} \leq u | data).
\]

where $\eta_{i0.05}$ is the $5^{th}$ percentile of breaking strength of $i^{th}$ species. 

$\Rightarrow$ Solving our problem reduces to characterizing $P(\eta_{k0.05} \leq u | data)$.

Two proposed approaches:
- Nonparametric Bayesian approach
- Semi-parametric Bayesian approach
Approach 1: A Nonparametric Bayesian Approach

Model each species separately;

Let $T$ be a breaking strength of a single species

Assume that CDF $G$ of $T$ is from Dirichlet process with two-parameter Weibull base distributions.

$$T \mid G \overset{i.i.d.}{\sim} G$$

$$G \overset{ind.}{\sim} Dir(G_0, v)$$

where $G_0(t) = Weibull(t; \beta, \lambda)$.

Then the posterior distribution of the $5^{th}$ percentile of $T$ from the $k^{th}$ species is:

$$P(\eta_{0.05} < t \mid data) = 1 - P(G(t) < 0.05 \mid data)$$

$$= 1 - Beta(0.05; \nu_m(t), \nu + m - \nu_m(t)),$$

where $\nu_m(t) = \nu G_0(t) + m \hat{F}(t)$, and $\hat{F}(t)$ is the empirical distribution function.
Approach 1: A Nonparametric Bayesian Approach

- Model each species separately;
- Let $T$ be a breaking strength of a single species

Assume that CDF $G$ of $T$ is from Dirichlet process with two-parameter Weibull base distributions.

$$T | G \overset{i.i.d.}{\sim} G$$

$$G \overset{ind.}{\sim} \text{Dir}(G_0, v)$$

where $G_0(t) = \text{Weibull}(t; \beta, \lambda)$.

Then the posterior distribution of the 5th percentile of $T$ from the $k$th species is:

$$P(\eta_{0.05} < t | \text{data}) = 1 - P(G(t) < 0.05 | \text{data})$$

$$= 1 - \text{Beta}(0.05; \nu_m(t), v + m - \nu_m(t)),$$

where $\nu_m(t) = vG_0(t) + m\hat{F}(t)$, and $\hat{F}(t)$ is the empirical distribution function.
Approach 2: Semi-parametric approach (DP mixture of Weibull distributions)

- Model each species separately

\[ T | \beta, \lambda \sim \text{Weibull}(\cdot; \beta, \lambda) \]
\[ \beta, \lambda | H \sim H \]
\[ H \sim \text{DP}(H_0, \nu), \]

where \( H_0(\beta, \lambda) = \text{Unif}(\beta; 0, \phi)\text{Unif}(\lambda; 0, \gamma). \)

Another way to express the DP mixture of Weibull distributions is:

\[
F_T(t) = \int \int \text{Weibull}(t | \beta, \lambda) H(d\beta, d\lambda) \\
= \sum_{h=1}^{\infty} \pi_h \text{Weibull}(t; \beta_h, \lambda_h), \quad (\beta_h^*, \lambda_h^*) \overset{i.i.d.}{\sim} H_0.
\]

No analytic form of the posterior distribution of \( \eta_{0.05} \) is available. Design MCMC to return the posterior samples of the marginal CDF of the breaking strength then invert the sampled CDF to obtain the posterior samples of \( \eta_{0.05} \).
Real data examples: datasets of breaking strength from 3 species

Figure: Estimated posterior density of $T$ from semi-parametric procedure

Figure: Estimated posterior CDF of $\eta_{0.05}$ from non/semi-parametric procedures
Real data examples: datasets of breaking strength from 3 species

**Figure:** Estimated posterior density of $T$ from semi-parametric procedure

**Figure:** Estimated posterior CDF of $\eta_{0.05}$ from non/semi-parametric procedures

S1, size = 282
S2: size = 98
S3: size = 174
Real data examples: datasets of breaking strength from 3 species

Figure: Estimated posterior density of $T$ from semi-parametric procedure

Figure: Estimated posterior CDF of $\eta_{0.05}$ from non/semi-parametric procedures

S1, size = 282
S2: size = 98
S3: size = 174
Simulation studies: which procedure returns “stable” subset?

- Seven species make up a single marketing group;
- Generate $m$ breaking strengths for each of the seven species 300 times ($m = 100, 360$);
- Breaking strengths of each species are generated from mixture distributions;
- Gaps between successively larger fifth percentiles of the seven species are set equal to our proprietary datasets of seven species.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Subset size</th>
<th>% capture the weakest</th>
<th>% Set stays the same</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100 360</td>
<td>100 360</td>
</tr>
<tr>
<td>ASTM</td>
<td>5.92</td>
<td>2.85</td>
<td>100 100</td>
</tr>
<tr>
<td>$P^*=95$</td>
<td>NP1</td>
<td>3.13 1.71</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>NP2</td>
<td>2.41 1.93</td>
<td>98 97</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td>2.91 1.50</td>
<td>100 100</td>
</tr>
<tr>
<td>$P^*=99$</td>
<td>NP1</td>
<td>4.79 2.35</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>NP2</td>
<td>3.91 2.46</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td>4.61 1.88</td>
<td>100 100</td>
</tr>
<tr>
<td>$P^*=99.5$</td>
<td>NP1</td>
<td>5.35 2.99</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>NP2</td>
<td>4.55 3.22</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>SP</td>
<td>5.39 2.07</td>
<td>100 100</td>
</tr>
</tbody>
</table>
Simulation studies: which procedure returns “stable” subset?

- Seven species make up a single marketing group;
- Generate $m$ breaking strengths for each of the seven species 300 times ($m = 100, 360$);
- Breaking strengths of each species are generated from weibull distributions;
- Gaps between successively larger fifth percentiles of the seven species are set equal to our proprietary datasets of seven species.

<table>
<thead>
<tr>
<th>$m$</th>
<th>Subset size</th>
<th>$P^* = 95$</th>
<th>$P^* = 99$</th>
<th>$P^* = 99.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ASTM</td>
<td>NP1</td>
<td>NP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.46 4.21</td>
<td>3.83 2.53</td>
<td>3.10 2.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>99 100</td>
<td>99 100</td>
</tr>
<tr>
<td></td>
<td>% capture the weakest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>99 100</td>
<td>99 100</td>
</tr>
<tr>
<td></td>
<td>% Set stays the same</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 64</td>
<td>69 87</td>
<td>75 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NP1</td>
<td>NP2</td>
<td>SP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.37 3.72</td>
<td>4.51 2.96</td>
<td>5.42 2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>100 100</td>
<td>99 100</td>
</tr>
<tr>
<td></td>
<td>% capture the weakest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>100 100</td>
<td>99 100</td>
</tr>
<tr>
<td></td>
<td>% Set stays the same</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>34 57</td>
<td>49 67</td>
<td>44 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SP</td>
<td>NP1</td>
<td>NP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.78 4.32</td>
<td>5.03 3.54</td>
<td>5.94 3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>% capture the weakest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
</tr>
<tr>
<td></td>
<td>% Set stays the same</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>22 35</td>
<td>30 49</td>
<td>32 99</td>
</tr>
</tbody>
</table>
When sample size is small, the ASTM procedure tend to return a larger controlling species than any of the Bayesian methods for any of the $P^*$ 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 semi-parametric procedure always outperforms the ASTM procedure in terms of stabil- ity of CS.

Our semi-parametric procedure could be useful refinement to the current ASTM procedure.

All the R functions required for our procedures are implemented in R package DPw that is available at http://cran.r-project.org/web/packages/DPw/ (Actual computations are done in C language).