Bootstrapping data arrays of arbitrary order

Art B. Owen  Dean Eckles
Stanford University  Stanford University and Facebook, Inc.

June 2011 revised November 2011

Abstract
In this paper we study a bootstrap strategy for estimating the variance of a mean taken over large multifactor crossed random effects data sets. We apply bootstrap reweighting independently to the levels of each factor, giving each observation the product of independently sampled factor weights. No exact bootstrap exists for this problem (McCullagh, 2000). We show that the proposed bootstrap is mildly conservative, meaning biased towards overestimating the variance, under sufficient conditions that allow very unbalanced and heteroscedastic inputs. Earlier results for a resampling bootstrap only apply to two factors and use multinomial weights that are poorly suited to online computation. The proposed reweighting approach can be implemented in parallel and online settings. The results for this method apply to any number of factors. The method is illustrated using a 3 factor data set of comment lengths from Facebook.

Keywords: Bayesian pigeonhole bootstrap, online bagging, online bootstrap, relational data, tensor data, unbalanced random effects

1 Introduction
Large sparse data sets with two or more crossed random effects commonly arise from electronic commerce and Internet services, and we may expect them to arise in other settings as automated data gathering becomes more prevalent. Such data often have no IID structure for us to draw on. For example, with the famous Netflix data (Bennett and Lanning, 2007) multiple ratings from the same viewer are dependent. Similarly multiple ratings on the same movie are dependent. Neither rows nor columns are IID, and a crossed random effects model with interactions is a more reasonable structure.

In Internet data there can easily be more than two crossed factors. The individual factors could be user account numbers, IP addresses, URLs, search query strings or identifiers for documents placed in web pages. The response variable might be a measure of user engagement such as time spent reading, or system performance such as the load times for pages under different versions of software.
The crossed random effects setting is challenging for inference. Methods in Searle et al. (1992) rely on Gaussian data assumptions and outside of balanced cases, the necessary linear algebra becomes prohibitively expensive on large problems.

We might therefore turn to resampling. For IID data, the bootstrap provides reliable variance estimates and confidence intervals under very weak assumptions on the mechanism generating our data. But McCullagh (2000) proved that there does not exist an exact bootstrap algorithm for crossed random effects. Specifically if

\[ X_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \]

for independent mean 0 random variables \( a_i, b_j \) and \( \varepsilon_{ij} \) with variances \( \sigma_A^2, \sigma_B^2 \) and \( \sigma_E^2 \) respectively, then no resampling method, from a very broad class, will provide an unbiased estimate of \( \text{Var}(IJ)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij} \).

One approach to bootstrapping crossed data is to independently bootstrap the indices of each factor. In bootstrapping a factor we are putting a random multinomially distributed weight on the levels of that factor. For an \( r \)-fold data set, the observation \( X_{i_1,i_2,...,i_r} \) gets a weight \( W^{*b}_{i_1i_2...i_r} = \prod_{j=1}^{r} W^{*b}_{j,i_{j}} \) where \( W^{*b}_{j,i_{j}} \) is the weight on level \( i_{j} \) of the \( j \)th factor in the \( b \)th bootstrap reweighting. For each \( j \) and each \( b \), weight vectors \( (W^{*b}_{j,1}, W^{*b}_{j,2}, \ldots, W^{*b}_{j,N_j}) \) are sampled independently. Given weights on all the data, we compute a weighted version of the statistic(s) of interest to get the \( b \)th bootstrap value.

Bootstrapping with a product of multinomial weights has been studied before, for \( r = 2 \). Brennan et al. (1987) and Wiley (2001) use it to study variance components in educational test data. McCullagh (2000) shows that independently bootstrapping the rows and columns of a data matrix gives a mildly conservative estimate of variance. That is, it has a positive bias that is usually relatively small. McCullagh (2000) considered balanced crossed random effects (no missing values) with homoscedastic variance components. Owen (2007) shows that this bootstrap remains conservative (and usually mildly so) for sparsely sampled unbalanced crossed random effects allowing for heteroscedasticity. That framework allows every row and column (e.g. customer and movie) and even every interaction to have its own variance. Resampling is then reliable and it spares the analyst from having to estimate all of those variances.

The random weighting that we favor is a product of completely independent weights: \( W^{*b}_{i_1i_2...i_r} = \prod_{j=1}^{r} W^{*b}_{j,i_{j}}, \) where for each \( b \) and each \( j \), \( W^{*b}_{j,i_{j}} \) are IID weights with mean 1 and variance 1. For these large data sets, methods that reweight data via IID random weights (Rubin 1981) Newton and Raftery 1994) are an appealing alternative to the multinomial weights used in resampling. First, it is simpler to apply independent reweighting to large scale parallelized computations, as is done in online bagging and boosting (Oza 2001; Lee and Clyde 2004). The reason is that large data sets are stored in a distributed fashion and then multinomial sampling brings substantial communication and synchronization costs. Second, resampling simplifies variance expressions by avoiding the negative dependence from the multinomial distribution. This makes it easier to develop expressions for problems with more than two factors.

Using notation and approximations defined below, the main facts are as
follows. For $r = 2$ factors, we suppose the data are sampled by a random effects model with variance components $\sigma^2_{1(1)}$, $\sigma^2_{1(2)}$, and $\sigma^2_{1(1,2)}$ corresponding to the main effects and interaction respectively. We can express the variance of the sample average of $N$ observations in the form $(\nu(1)\sigma^2_{1(1)} + \nu(2)\sigma^2_{1(2)} + \sigma^2_{1(1,2)})/N$. The subscripted $\nu$ quantities are easily computable from the data and we give explicit formulas. Naive bootstrapping produces an estimate close to $(\sigma^2_{1(1)} + \nu(1)\sigma^2_{1(2)} + \sigma^2_{1(1,2)})/N$ which is grossly inadequate because it turns out that often $\nu(j) \gg 1$. For instance, in the Netflix data set, the largest $\nu(j)$ is about 56,200.

Resampling both rows and columns leads to a variance estimate close to $((\nu(1) + 2)\sigma^2_{1(1)} + (\nu(2) + 2)\sigma^2_{1(2)} + 3\sigma^2_{1(1,2)})/N$, which is mildly conservative when $\nu(j) \gg 1$ and the $\sigma$’s are of comparable magnitude. It is up to three times as large as it should be in the event that $\sigma^2_{1(1)} \ll \sigma^2_{1(1,2)}$. Being conservative by a factor of at most 3 is far more acceptable than underestimating variance by as much as 56,200.

Our main contributions are as follows:

1) We show that a naive bootstrap suitable for IID settings severely underestimates the variance of the sample mean, when $r = 2$, while the product strategy mildly over-estimates it. These facts were known for resampling, but we show it also for reweighting.

2) We generalize the reweighting results to $r \geq 2$ factors. In particular for the homoscedastic setting, the $3\sigma^2_{1(1,2)}$ variance contribution from the case $r = 2$ becomes $(2^r - 1)\sigma^2_{1(1,2,...,r)}$. We find expressions for all $2^r - 1$ variance coefficients. Under reasonable conditions, for which we note exceptions, this bootstrap magnifies a $k$-factor variance component by roughly $2^k - 1$. Under simply described conditions, the $k = 1$ terms dominate the variance and then the variance magnification becomes negligible.

3) We introduce a heteroscedastic random effects model in which every non-empty subset of factors contributes a random effect. The product weighted bootstrap remains mildly conservative even when every factorial effect for every observation has a distinct variance, so long as all the variances are uniformly bounded away from 0 and infinity.

An outline of the paper is as follows. Section 2 introduces our notation for the random effects model and some observation counts and then defines the random effects variance that we seek to estimate. Section 3 considers naive bootstrap methods that simply resample or reweight the observations as if they were IID. They seriously underestimate the true variance unless the only nonzero variance component is that of the highest order interaction. Reweighting has a slight advantage because it allows one to step up the sampling variance to compensate for cases where the naive bootstrap variance is only a modest underestimate. Section 4 introduces a factorial reweighting bootstrap strategy. For data with $r = 2$ factors, the reweighting results closely match the resampling results from Owen (2007). This section includes an interpretable approximation to the exact bootstrap variance. Section 6 considers the heteroscedastic case, where every variance component at every combination of its factors has its own variance parameter. When the main effects are dominant, then the proposed
bootstrap closely matches the desired variance even in the heteroscedastic setting. Section 7 describes repeated observations and factors nested inside the ones being reweighted. Section 8 has a numerical example from Facebook. In that dataset, UK-based users make longer comments than do US-based users, when posting from mobile devices. The reverse holds for comments made at Facebook’s standard web interface. The differences are small, but statistically significant, even after taking account of a three factor structure (commenter, sharer, and URL). The proofs appear in an Appendix.

Although the product reweighting algorithm is simple, its analysis in the random effects context is very technical. Section 9 discusses some larger statistical issues. Among these are the reasons that we do not model the possible informativeness of the missing data mechanism, the reasons for focusing on the bootstrap variance of a sample mean, and the motivation for considering the heteroscedastic random effects model, which contains many more parameters than observations.

2 Notation and random effects model

The random variables of interest take the form \( X_{i_1, i_2, \ldots, i_r} \in \mathbb{R}^d \) for integers \( i_j \geq 1 \) and \( j = 1, \ldots, r \). To simplify notation, we write \( X_i \) for \( i = (i_1, \ldots, i_r) \). We work with \( X \) of dimension \( d = 1 \). The generalization to \( d \geq 1 \) is straightforward. We have in mind applications where each value of \( i_j \) corresponds to one level of a categorical variable with many potential values. In Internet applications, index values \( i_j \) might represent users, URLs, IP addresses, ads, query strings and so on. There may be no a priori upper bound on the number of distinct levels for \( i_j \).

The data are composed of \( N \) of these random variables, where \( 1 \leq N < \infty \). The binary variable \( Z_i \) takes the value 1 when observation \( X_i \) is present and \( Z_i = 0 \) when \( X_i \) is absent. We work conditionally on \( Z_i \) so that they are nonrandom. In practice the pattern of missingness in \( Z_i \) may be important. As with prior work, we avoid modeling \( Z_i \) in order to focus on estimating variance, apart from some brief remarks in Section 9.

The letters \( u \) and \( v \) denote subsets of \([r] \equiv \{1, \ldots, r\}\) throughout. The summation \( \sum_u \) is taken over all \( 2^r \) subsets of \([r]\), and other summations, such as \( \sum_{v \supseteq u} \) denote sums over the first named set (here \( v \)) subject to the indicated condition with the other set(s) (here \( u \)) held fixed. The index \( i_u \) extracts the components \( i_j \) for \( j \in u \). Then \( i_u = i'_{u} \) means that \( i_j = i'_j \) for all \( j \in u \).

Our \emph{r-fold crossed random effects model} is

\[
X_i = \mu + \sum_{u \neq \emptyset} \varepsilon_{i, u} 
\]  

where \( \mu \in \mathbb{R} \) and \( \varepsilon_{i, u} \) are mean 0 random variables that depend on \( i \) only through \( i_u \). We have \( \varepsilon_{i, u} = \varepsilon_{i', u} \) if \( i_u = i'_u \) and \( \varepsilon_{i, u} \) independent of \( \varepsilon_{i', u} \) otherwise. The covariance of \( \varepsilon_{i, u} \) and \( \varepsilon_{i', u'} \) is

\[
\text{Cov}(\varepsilon_{i, u}, \varepsilon_{i', u'}) = \mathbb{E}(\varepsilon_{i, u} \varepsilon_{i', u'}) = \sigma_u^2 1_{u = u'} 1_{i_u = i'_u} 
\]
for $\sigma_u^2 < \infty$.

To illustrate the model notation, suppose that $r = 2$ and one observation is at $i = (38, 44)$ while another is at $i' = (38, 19)$. Then

$$X_i = X_{(38, 44)} = \mu + \varepsilon_{(38, 44), \{1\}} + \varepsilon_{(38, 44), \{2\}} + \varepsilon_{(38, 44), \{1, 2\}} \quad \text{and}$$

$$X_{i'} = X_{(38, 19)} = \mu + \varepsilon_{(38, 19), \{1\}} + \varepsilon_{(38, 19), \{2\}} + \varepsilon_{(38, 19), \{1, 2\}}.$$  \hspace{1cm} (3)

Because $i$ and $i'$ share a value for $i_1$ they have the same random effect for the set $u = \{1\}$. That is $\varepsilon_{i, \{1\}} = \varepsilon_{i', \{1\}}$. This is the only effect that they share and so $\text{Cov}(X_i, X_{i'}) = \sigma_u^2$. More generally, suppose that two indices $i$ and $i'$ satisfy $i_j = i'_j$ for and only for $j \in u$. Then $X_i$ and $X_{i'}$ share random effects $\varepsilon_{i, v} = \varepsilon_{i', v}$ for all nonempty $v \subseteq u$ and so $\text{Cov}(X_i, X_{i'}) = \sum_{v: \emptyset \neq v \subseteq u} \sigma_v^2$.

The expression $\varepsilon_{(38, 44), \{1\}}$ is mildly redundant since the second index $i_2 = 44$ is ignored. We could have written it as $\varepsilon_{(38), \{1\}}$. Such a choice amounts to writing the general case as $\varepsilon_{i, u}$ which is more cumbersome when it appears in lengthy expressions.

The sample mean of $X$ is the ratio

$$\bar{X} = \frac{\sum_i X_i Z_i}{\sum_i Z_i}, \hspace{1cm} (4)$$

where the sums are over all index values $i$. The denominator in (4) is the total number $N$ of observations. Our goal is to estimate the variance of $\bar{X}$ by resampling methods.

2.1 Partial duplicate observations

We will need to keep track of the extent to which different observations have the same index values, in order to properly reflect correlations among the $X_i$.

For each $i$ and $u \subseteq [r]$, the number

$$N_{i, u} = \sum_{i'} Z_{i'} 1_{i_{u} = i'_{u}}$$

counts how many observations match $X_i$ for all indices $j \in u$. If $Z_i = 1$ then $N_{i, u} \geq 1$ because $X_i$ matches itself. By convention $N_{i, \emptyset} = N$ and $N_{i, \{r\}} = 1$. The quantity

$$\nu_u = \frac{1}{N} \sum_i Z_i N_{i, u} \geq 1$$

is the average number of matches in the subset $u$ for observations in the data set, and $\nu_{\{r\}} = 1$.

The most important of the $\nu_u$ are for singletons $u = \{j\}$. The value $\nu_{\{j\}}$ has a quadratic dependence on the pattern of duplication in the data. To see this, write $n_{tj} = \sum_i Z_{i} 1_{i = t}$ for the number of times that variable $j$ is equal to $t$ in the data. Then $\nu_{\{j\}} = N^{-1} \sum_{t=1}^{\infty} n_{t,j}^2$, because each $N_{i, \{j\}} = n_{i,j}$ appears $n_{i,j}$ times in the summation defining $\nu_{\{j\}}$. 

5
If $u \subseteq v$ then $\nu_u \geq \nu_v$. In some applications $\nu_u \gg \nu_v$ for proper subsets $u \subset v$. For those applications, multiple matches are very unusual. In other settings two factors, say $i_1$ and $i_2$ might be highly though not perfectly dependent (e.g., customer ID and phone number) and then $\nu_{(1,2)}$ might be only slightly smaller than $\nu_{(1)}$ or $\nu_{(2)}$. We return to this issue in Section 5.

The specific pair of data values $i$ and $i'$ match in components

$$M_{ii'} = \{ j \in [r] \mid i_j = i'_j \}.$$ 

For the motivating data, most of the $M_{ii'}$ are empty and most of the rest have cardinality $|M_{ii'}| = 1$. We have $|M_{ii'}| = r$ if and only if $i = i'$. Although $M_{ii'}$ is defined for all pairs $i$ and $i'$ we only use it when $Z_i Z_{i'} = 1$, that is when both $X_i$ and $X_{i'}$ have been observed, and the term ‘most’ above refers to these pairs.

For each $i$ and $k = 0, 1, \ldots, r$, the number

$$N_{i,k} = \sum_{i'} Z_{i'} 1_{|M_{ii'}| = k}$$

counts how many observations match $X_i$ in exactly $k$ places.

### 2.2 Random effects variance of $\bar{X}$

Here we record the true variance of $\bar{X}$, using the random effects model. This is the quantity we hope to estimate by bootstrapping.

**Theorem 1.** In the random effects model

$$\text{Var}(\bar{X}) = \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma_u^2. \quad (5)$$

The contributions of the variance components $\sigma_u^2$ are proportional to the duplication indices $\nu_u$. For large sparse data sets we often find that $1 \ll \nu_u \ll N$ when $0 < |u| < r$.

Our bootstrap approximations to this variance are centered around a quantity $(1/N) \sum_{u \neq \emptyset} \gamma_u \sigma_u^2$ for gain coefficients $\gamma_u$ that depend on the data configuration and the particular bootstrap method. Ideally we want $\gamma_u = \nu_u$. More realistically, some bootstrap methods are able to get $\gamma_u \geq \nu_u$ with $\gamma_u$ just barely larger than $\nu_u$ for the singletons $u = \{j\}$ which we expect to dominate $\text{Var}(\bar{X})$.

### 3 Naive bootstrap methods

There are two main ways to bootstrap: resampling (Efron, 1979) and reweighting (Rubin, 1981), with the distinction being that the former uses a multinomial distribution on the data while the latter applies independent random weights to the observations.

Naive bootstrap methods simply resample or reweight the $N$ observations without regard to their factorial structure. That is they use the same bootstrap one might use for IID samples. Here we show that naive bootstrap resampling and reweighting have very similar and very unsatisfactory performance.
3.1 Naive resampling

In the naive bootstrap, all $N$ observations are resampled with replacement. The naive bootstrap variance of $\bar{X}$ converges to

$$\text{Var}_{NB}(\bar{X}) = \frac{1}{N^2} \sum_i Z_i (X_i - \bar{X})^2$$  \hspace{1cm} (6)

as the number of resampled data sets tends to infinity.

**Theorem 2.** Under the random effects model (1), the expected value of the naive bootstrap variance of $\bar{X}$ is

$$E_{RE}(\text{Var}_{NB}(\bar{X})) = \frac{1}{N} \sum_{u \neq \emptyset} \sigma_u^2 \left(1 - \frac{\nu_u}{N}\right).$$  \hspace{1cm} (7)

When $r > 1$, the naive bootstrap can severely underestimate the coefficients of $\sigma^2_{\{j\}}$. We can see the effect in the Netflix data, which has $r = 2$. The gain coefficients $\nu_u$ can be computed directly. For a random variable $X$ following the random effects model (1) with variance components $\sigma^2_{\text{movies}}$, $\sigma^2_{\text{raters}}$ and $\sigma^2_{\text{movies} \times \text{raters}}$ we have

$$\text{Var}(\bar{X}) = \frac{1}{N} (56,200 \sigma^2_{\text{movies}} + 646 \sigma^2_{\text{raters}} + \sigma^2_{\text{movies} \times \text{raters}}), \quad \text{while}$$

$$\text{Var}_{NB}(\bar{X}) \leq \frac{1}{N} (\sigma^2_{\text{movies}} + \sigma^2_{\text{raters}} + \sigma^2_{\text{movies} \times \text{raters}}),$$

where $N \doteq 100,000,000$. If $X$ has a large variance components for movies, the underestimation can be severe. Even quantities dominated by a rater effect will have a naive bootstrap variance far too small.

Theorem 2 generalizes Lemma 2 of Owen (2007) which treats naive bootstrap sampling for $r = 2$. We note that Owen (2007, p. 391) has an error: it gives the coefficient of $\sigma^2_{\{1,2\}}$ as $1/N$ where it should be $1/(N-1)$.

3.2 Naive reweighting

Posterior sampling under the Bayesian bootstrap (Rubin 1981) uses independent Exp(1) weights on the sample values. This corresponds to a posterior distribution on $X$ that is a Dirichlet distribution with parameter vector $(1, 1, \ldots, 1)$ with a 1 for each observation. The corresponding prior is a degenerate Dirichlet with a parameter of 0 on all possible values for the random variable. The posterior is degenerate putting 0 probability on any value of $X$ that was never seen in the sample, thus eliminating the user’s need to know which possible values were not in fact observed. This motivation is simplest when the observations are assumed to be distinct as for example with continuously distributed values, but the method is also used on data with ties.

In the naive Bayesian bootstrap, all $N$ observations are given random weights which are then normalized. Observation $i$ gets weight $W_i \sim G$ independently
sampled. We assume that $G$ has mean 1 and variance $\tau^2 < \infty$. Typically $\tau^2 = 1$.

The original Bayesian bootstrap (Rubin 1981) had $W_i \sim \text{Exp}(1)$ but other distributions are useful too. Taking $W_i \sim \text{Poi}(1)$ gives a result very similar to the usual bootstrap, and it has integer weights. Independent Bin$(N, 1/N)$ weights would provide a more exact match, but for large $N$ there is no practical difference between Bin$(N, 1/N)$ and Poi$(1)$. See Oza (2001) and Lee and Clyde (2004) for uses of independent reweighting in bagging and boosting.

Taking $W_i \sim \text{U}\{0, 2\} = (\delta_0 + \delta_2)/2$ also has integer values. The algorithm goes ‘double or nothing’ independently on all $N$ observations. The nonzero integer values are all equal, so these weights correspond to using a random unweighted subset of the data. Double-or-nothing weighting is then a version of half-sampling methods (McCarthy, 1969) without the constraint on the sum of weights, just as Poisson weighting removes a sum constraint from the original bootstrap.

The choice of weights makes a small difference to the bootstrap performance. See Section 3.3.

Each bootstrap resampled mean takes the form

$$\bar{X}^* = \frac{T^*}{N^*}$$

where $T^* = \sum_i W_i Z_i X_i$ and $N^* = \sum_i W_i Z_i$. The bootstrap mean $T^*/N^*$ is a ratio estimator of $\bar{X}$. The asymptotic formula for the variance is

$$\tilde{\text{Var}}_{\text{NBB}}(\bar{X}^*) = \frac{1}{N^2} \text{E}_{\text{NBB}}\left((T^* - \bar{X} N^*)^2\right).$$

The tilde on $\text{Var}_{\text{NBB}}$ is a reminder that this formula is a delta method approximation: it is the variance of a Taylor approximation to $\bar{X}^*$. Because $N$ is usually very large in the target applications, we consider $\tilde{\text{Var}}_{\text{NBB}}$ to be a reliable proxy for $\text{Var}_{\text{NBB}}$.

**Theorem 3.** In the random effects model [1]

$$\text{E}_{\text{RE}}\left(\tilde{\text{Var}}_{\text{NBB}}(\bar{X}^*)\right) = \frac{\tau^2}{N} \sum_{u \neq \emptyset} \sigma^2_u \left(1 - \frac{\nu_u}{N}\right). \quad (8)$$

The naive Bayesian bootstrap using $\tau^2 = 1$ has the same average variance as the naive bootstrap. In large data sets we may find that $\nu_u \geq \tau^2(1 - \nu_u/N)$ and then the Bayesian bootstrap greatly underestimates the true variance. When $\max_{u \neq \emptyset} \nu_u$ is not too large, then Theorem 3 offers a way to counter this problem. We can take simply multiply the naive bootstrap variance by $\tau^2 = \max_{u \neq \emptyset} \nu_u$ to get conservative variance estimates. The largest $\nu_u$ comes from $u = \{j\}$ for some $j \in [r]$ and it is an easy quantity to compute.

### 3.3 Bootstrap stability

Any distribution on weights with $\text{E}(W) = 1$ and $\text{Var}(W) = \tau^2$ will have the same value for $\text{E}_{\text{RE}}(\text{Var}_{\text{NBB}}(\bar{X}^*))$. Several different weight distributions are
popular in the literature. [Oza (2001)] takes $W_{ij}$ to be independent Poisson random variables with mean 1. This creates a very close approximation to the original bootstrap’s multinomial weights. [Lee and Clyde (2004)] prefer exponential weights with mean 1 because they yield an exact online version of the Bayesian bootstrap.

In this section we look at the effect of the weight distribution. Other things being equal, we prefer a bootstrap to yield a stable variance estimate. That is, we like a smaller variance under bootstrap sampling for the estimated variance of the mean. For this purpose it is better to have weights with a small kurtosis $\kappa = \mathbb{E}((W - 1)^4)/\tau^4 - 3$. The smallest possible kurtosis for weights with mean 1 and variance 1 arises for weights uniformly distributed on the values 0 and 2.

The kurtosis of the data, $\kappa_x = (1/N) \sum_i Z_i (X_i - \bar{X})^4/\sigma^4 - 3$ where $\sigma^2 = (1/N) \sum_i (X_i - \bar{X})^2$, also plays a role. We work out the consequences for the naive bootstrap for simplicity.

If we hold the observations $X_i$ fixed and implement the bootstrap, doing some number $B$ of replicates, we will estimate the quantity

$$\tilde{\text{Var}}_{NBB}(\bar{X}^*) = \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} \mathbb{E}_{NBB}(W_i W_{i'}) Y_i Y_{i'}$$

where $Y_i = X_i - \bar{X}$. To estimate this variance we may use

$$\tilde{\text{Var}}_{NBB}(\bar{X}^*) = \frac{1}{BN^2} \sum_{b=1}^B \sum_i \sum_{i'} Z_i Z_{i'} W_{i,b} W_{i',b} Y_i Y_{i'}$$

$$= \frac{1}{B} \sum_{b=1}^B \left( \frac{1}{N} \sum_i Z_i W_{i,b} (X_i - \bar{X}) \right)^2$$

where $W_{i,b}$ are independent identically distributed random weights and $b = 1, \ldots, B$ indexes the bootstrap replications. The hat in $|9|$ represents estimation from $B$ bootstrap samples. It is possible to use $|9|$ with $B = 1$. That such a “unistrap” is possible, stems from the use of a delta method approximation.

Equation $|9|$ is not the usual estimator. The more usual variance estimate is

$$s^2_{NBB}(\bar{X}^*) = \frac{1}{B - 1} \sum_{b=1}^B (\bar{X}^*_b - \bar{X}^*)^2$$

where

$$\bar{X}^*_b = \frac{1}{N} \sum_i Z_i W_{i,b} X_i, \quad \text{and} \quad \bar{X}^* = \frac{1}{B} \sum_{b=1}^B \bar{X}^*_b.$$  \hspace{1cm} (11)

**Theorem 4.** Let $W$ and $W_{i,b}$ be IID random variables with mean 1 variance $\tau^2$ and kurtosis $\kappa_w < \infty$. Then holding $Y_i = X_i - \bar{X}$ fixed,

$$\text{Var}_{NBB}(\tilde{\text{Var}}_{NBB}(\bar{X}^*)) = \frac{\sigma^4 \tau^4}{BN^2} \left( 2 + \frac{\kappa(\kappa_x + 3)}{N} \right)$$
where \( \sigma^2 = (1/N) \sum_i Z_i Y_i^2 \), and \( \kappa_x = (1/N) \sum_i Z_i Y_i^4 / \sigma^4 - 3 \). A delta method approximation gives

\[
\text{Var}_{\text{NBB}}(s_{\text{NBB}}^2) \approx \frac{\sigma^4 \tau^4}{BN^2} \left( \frac{2B}{B - 1} + \frac{\kappa(\kappa_x + 3)}{N} \right).
\]

When \( \kappa(\kappa_x + 3) \ll N \), then \( \text{Var}_{\text{NBB}}(\bar{X}^*) \) with B reweightings has approximately the variance of \( s_{\text{NBB}}^2(\bar{X}^*) \) with \( B + 1 \) reweightings.

We find here that there are only small differences between weighting schemes, but double-or-nothing weights having the smallest possible kurtosis \( \kappa = -2 \) have the best stability. The Pois(1) distribution has \( \kappa = 1 \) and the Exp(1) distribution has \( \kappa = 6 \).

4 Factorial reweighting

Our proposal here is to apply a product of independent random weights to the data. Observation \( i \) is given weight \( W_i \geq 0 \). The weights take the form

\[
W_i = \prod_{j=1}^r W_{j,i_j}
\]  

(12)

where \( W_{j,i_j} \) are independent random variables for \( j \in [r] \) and \( i_j \geq 1 \). We assume that \( \mathbb{E}(W_{j,i_j}) = 1 \) and \( \text{Var}(W_{j,i_j}) = \tau_j^2 < \infty \). The usual choice has all \( \tau_j^2 \) equal to a common \( \tau^2 \) which in turn is usually equal to 1.

For the example in equation (3), the observation at index \( i = (38, 44) \) gets weight \( W_i = W_{1,38} W_{2,44} \). It shares one weight factor with the observation at \( i' = (38, 19) \) which has \( W_{i'} = W_{1,38} W_{2,19} \).

The reweighted mean \( \bar{X}^* \) is once again a ratio estimate with delta method approximation

\[
\text{Var}_{\text{PW}}(\bar{X}^*) = \frac{1}{N^2} \mathbb{E}_{\text{PW}} \left( (T^* - \bar{X} N^*)^2 \right)
\]  

(13)

where \( T^* = \sum_i Z_i W_i X_i \) and \( N^* = \sum_i Z_i W_i \) for \( W_i \) given by (12). The subscript PW refers to random weights taking the product form.

The bootstrap variance depends on precise details of the overlaps among different observations. We will derive some approximations to this variance.
below. For the exact variance we need to introduce some additional quantities:

\[ \rho_k = \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} 1_{M_{i,i'} = k}, \]

\[ \nu_{k,u} = \frac{1}{N} \sum_i \sum_{i'} Z_i Z_{i'} 1_{M_{i,i'} = k} 1_{i = i_u}, \]

and,

\[ \tilde{\nu}_{k,u} = \frac{1}{N^2} \sum_i \sum_{i'} \sum_{i''} Z_i Z_{i'} Z_{i''} 1_{M_{i,i'} = k} 1_{i = i''} \]

\[ = \frac{1}{N^2} \sum_i Z_i N_{i,u} N_{i,k} \]

for \( k = 0, 1, \ldots, r \) and \( u \subseteq [r] \). In words, \( \rho_k \) gives the fraction of data pairs that match in exactly \( k \) positions, while \( \nu_{k,u}/N \) gives the fraction of data pairs that match in exactly \( k \) positions including all \( j \in u \). The third quantity, \( \tilde{\nu}_{k,u} \), is \( N \) times the fraction of data triples \((i, i', i'')\) in which \( i \) matches \( i' \) in precisely \( k \) places while also matching \( i'' \) for all \( j \in u \).

These new quantities satisfy the identities

\[ \sum_{k=0}^{r} \rho_k = 1, \quad \text{and} \quad \sum_{k=0}^{r} \nu_{k,u} = \sum_{k=0}^{r} \tilde{\nu}_{k,u} = \nu_u. \]

Also it is clear that \( \nu_{k,u} = 0 \) when \( |u| > k \).

**Theorem 5.** In the random effects model \( \{1\} \)

\[ \mathbb{E}_{RE}(\text{Var}_P(X^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \gamma_u \sigma_u^2, \] (14)

where

\[ \gamma_u = \sum_{k=0}^{r} (1 + \tau^2)^k (\nu_{k,u} - 2\tilde{\nu}_{k,u} + \rho_k \nu_u). \] (15)

The quantities \( \gamma_u \) are ‘gain coefficients’ which multiply \( \sigma_u^2/N \). Ideally they should equal \( \nu_u \) and then the bootstrap variance would match the desired one. Where they differ from \( \nu_u \), the bootstrap variance is biased. Typically the bias is positive, making this bootstrap conservative. Sometimes the bias is very small.

The special case \( r = 1 \) is interesting because it corresponds to IID sampling. Then the only variance component is \( \sigma_{(1)}^2 \) which we abbreviate to \( \sigma^2 \) and equation \( \{14\} \) simplifies to

\[ \frac{\tau^2 \sigma^2}{N} \left(1 - \frac{1}{N}\right) = \frac{\tau^2 \sigma^2}{N-1}. \]

In this instance there is a (trivial) negative bias if \( \tau^2 = 1 \).

Independently reweighting rows and columns is similar to independently resampling them. That strategy of bootstrapping rows and columns has been
given several names in the literature. Brennan et al. (1987) called it “boot-p,i” because for educational testing data, it resamples both people and items. McCullagh (2000, page 294) calls the method “Boot-II”. There is also another “Boot-II” for the one way layout in that paper. Noting a similarity to Cornfield and Tukey’s pigeonhole model for analysis of variance, Owen (2007) calls this approach the “pigeonhole bootstrap”. Reweighting with a product of Rubin’s (1981) exponential weights is thus a “Bayesian pigeonhole bootstrap”.

5 Interpretable approximations

Theorem 5 gives exact finite sample formulas for the gain coefficients \( \gamma_u \), but they are unwieldy. Here we make some approximations to \( \gamma_u \) in order to get more interpretable results.

First we introduce the quantity

\[
\epsilon = \max_i \max_{u \neq \emptyset} \frac{N_i,u}{N}, \quad \max_i \max_{1 \leq j \leq r} \frac{N_i,j}{N},
\]

which measures the largest proportional duplication of indices. Though \( 1 \geq \epsilon \geq 1/N \), we anticipate that \( \epsilon \) will usually be small. For the Netflix data, \( \epsilon = 232,944/100,480,507 \approx 0.00232 \) stemming from one movie having 232,944 ratings.

Although we suppose that \( \epsilon \) is small below, it is worth pointing out that exceptions do arise, even for some very large data sets. For example if the observed data form a complete \( N_1 \times N_2 \times \cdots \times N_r \) sample, then \( \epsilon = \max_{1 \leq j \leq r} 1/N_j \). If one factor takes only a modest number of levels, then \( \epsilon \) is large. A second context where \( \epsilon \) is large arises when one of the factors is greatly dominated by one of its levels, as for example, we might find in Internet data where one factor is the country of the web user.

A second parameter to aid interpretability is

\[
\eta = \max_{\emptyset \subseteq u \subseteq v} \frac{\nu_u}{\nu_v}.
\]

By construction \( \eta \leq 1 \), and we ordinarily expect \( \eta \) to be small. Of the indices which match for \( j \in u \) only a relatively small number should also match for \( j \in v - u \) too, because each additional match in large data sets represents a coincidence. For the Netflix data

\[
\eta = \max\{\nu_{\{1,2\}}/\nu_{\{1\}}, \nu_{\{1,2\}}/\nu_{\{2\}}\} = 1/646 \approx 0.00155.
\]

While \( \eta \) is often small, there are exceptions. If two factors are very dependent then \( \eta \) need not be small. For example people’s names and phone numbers may be such variables: many or even most phone numbers are used by a small number of people (often one) and many people use only a small number of phone numbers. Then the fraction of data pairs matching on both of these variables will not be much smaller than the fraction matching on one of them.
In simplifying expressions we use $O(\eta)$ and $O(\epsilon)$. These describe limits as $\eta$ (respectively $\epsilon$) converge to 0. The implied constants may depend on $r$. In some expressions we have retained explicit constants.

**Theorem 6.** In the random effects model \( [1] \), the gain coefficient \( \{15 \} \) for \( u \neq \emptyset \) in the product reweighted bootstrap is

\[
\gamma_u = \nu_u [(1 + \tau^2)^{|u|} - 1 + \theta_u \varepsilon] + \sum_{v \subsetneq u} (1 + \tau^2)^{|v|} (\tau^2)^{|v| - |u|} \nu_v, \tag{16}
\]

where \( |\theta_u| \leq (1 + \tau^2)((1 + \tau^2)^r - 1)/\tau^2 \). For \( \tau^2 = 1 \),

\[
\gamma_u = \nu_u [2^{\mid u\mid} - 1 + \theta_u \varepsilon] + \sum_{v \subsetneq u} 2^{|v|} \nu_v,
\]

where \( |\theta_u| \leq 2^{r+1} - 2 \).

For \( r = 2 \) using \( \nu_{\{1,2\}} = 1 \) and the usual choice \( \tau^2 = 1 \), we find that

\[
\gamma_{\{j\}} = \nu_{\{j\}} (1 + \theta_{\{j\}} \varepsilon) + 2, \quad j = 1, 2, \quad \text{and}
\]

\[
\gamma_{\{1,2\}} = \nu_{\{1,2\}} (3 + \theta_{\{1,2\}} \varepsilon)
\]

where each \( |\theta| \leq 6 \). The Bayesian pigeonhole bootstrap variance closely matches the ordinary pigeonhole bootstrap variance. In the extreme setting where \( \sigma^2_{\{1\}} = \sigma^2_{\{2\}} = 0 < \sigma^2_{\{1,2\}} \) the resulting bootstrap variance is about three times as high as it should be. In a limit as \( \min_j \nu_{\{j\}} \rightarrow \infty \) and \( \epsilon \rightarrow 0 \),

\[
\frac{\mathbb{E}_{\text{RE}}(\tilde{\text{Var}}_{\text{PW}}(\tilde{X}^*))}{\text{Var}(X)} \rightarrow 1 \tag{17}
\]

holds for fixed \( \sigma^2_{\{j\}} > 0, j = 1, 2 \). For \( r = 3 \), with \( \nu_{\{1,2,3\}} = 1 \) and \( \tau^2 = 1 \)

\[
\gamma_{\{1\}} = \nu_{\{1\}} + 4 \nu_{\{1,2\}} + 4 \nu_{\{1,3\}} + 8
\]

\[
\gamma_{\{1,2\}} = 3 \nu_{\{1,2\}} + 8, \quad \text{and}
\]

\[
\gamma_{\{1,2,3\}} \approx 7,
\]

where \( \approx \) reflects an additive error of size \( \nu_u \theta_u \varepsilon \) for \( |\theta_u| \leq 14 \). In the extreme case where the only nonzero variance coefficient is \( \sigma^2_{\{3\}} \), then the product reweighted bootstrap variance is about 7 times as large as it should be. On the other hand, when the main effect variances \( \sigma^2_{\{j\}} \) are positive and \( \nu_v / \nu_u \rightarrow 0 \) for \( v \subseteq u \), then \( \{17\} \) holds. More generally, we have Theorem \( \{7\} \).

**Theorem 7.** For the random effects model \( \{1\} \) and the product reweighted bootstrap with \( \tau^2 = 1 \), the gain coefficient for nonempty \( u \subseteq [r] \) satisfies

\[
2^{\mid u\mid} - 1 - (2^{r+1} - 2) \epsilon < \frac{\gamma_u}{\nu_u} \leq 2^{\mid u\mid} (1 + 2\eta)^{\mid e-u\mid} - 1 + (2^{r+1} - 2) \epsilon.
\]

If there exist \( m \) and \( M \) with \( 0 < m \leq \sigma^2_u \leq M < \infty \) for all \( u \neq \emptyset \), then

\[
\frac{\mathbb{E}_{\text{RE}}(\tilde{\text{Var}}_{\text{PW}}(\tilde{X}^*))}{\text{Var}(X)} = 1 + O(\eta + \epsilon).
\]
The first claim of Theorem 7 can be summarized as
\[ \frac{\gamma_u}{\nu_i} = (2^{|u|} - 1)(1 + O(\eta)) + O(\epsilon) \approx 2^{|u|} - 1, \]
and the second as \( E_{\text{RE}}(\text{Var}_{\text{PW}}(\bar{X})^*)/\text{Var}(\bar{X}) \approx 1 \), where the implied constants on \( r \). They generally grow exponentially in \( r \) but the interesting values of \( r \) are small integers from 2 to 6 or so. The main effects dominate when \( \eta \) is small and they are properly accounted for when \( \epsilon \) is small.

6 The heteroscedastic model

In the \( r \)-fold crossed random effects model [1], the term \( \varepsilon_{i,u} \) has the same variance for all \( i \). This model may not be realistic. For instance, the Netflix data includes some customers whose ratings have very small variance and others with a very large variance. Similarly, but to a lesser extent, movies also differ in the variance of their ratings. Unequal variances have the potential to bias inferences, especially in unbalanced cases, because the entities with more observations on them might have systematically higher variance than the others do.

A more realistic model is the **heteroscedastic r-fold crossed random effects model**, with

\[ X_i = \mu + \sum_{u \neq \emptyset} \varepsilon_{i,u} \tag{18} \]

where \( \mu \in \mathbb{R} \) and \( \varepsilon_{i,u} \) are independent random variables with mean 0 and variance \( \sigma_{i,u}^2 \). There are more variance parameters than observations, we do not need to estimate them. Owen (2007) gives conditions under which the pigeonhole bootstrap with \( r = 2 \) produces a variance estimate with relative error tending to zero in the heteroscedastic setting. Here we investigate product reweighting with general \( r \) for model (18).

We need some new quantities. For \( u \neq \emptyset \), define

\[ \nu_{i,u} = \frac{1}{N} \sum_{i'} Z_{i'} \mathbf{1}_{i_u=i'_u} = \frac{N_{i,u}}{N}, \]

\[ \nu_{i,k} = \frac{1}{N} \sum_{i'} Z_{i'} \mathbf{1}_{|M_{i'u}|=k} = \frac{N_{i,k}}{N}, \quad \text{and} \]

\[ \nu_{i,k,u} = \frac{1}{N} \sum_{i'} Z_{i'} \mathbf{1}_{|M_{i'u}|=k \mathbf{1}_{i_u=i'_u}}. \]

We also will use

\[ \nu_u \sigma_u^2 = \frac{1}{N} \sum_{i} Z_i \nu_{i,u} \sigma_{i,u}^2, \quad \text{and} \]

\[ \nu_k = \frac{1}{N} \sum_{i} Z_i \nu_{i,k}. \]
Next, we parallel the development from the ordinary random effects model \(^1\). Theorem 8 gives the exact variance of \( \bar{\mathbf{X}} \) for heteroscedastic random effects, Theorem 9 gives the gain coefficients under product reweighting, Theorem 10 provides interpretable bounds for the gains in terms of \( \epsilon \). Finally Theorem 11 gives conditions under which the product reweighted bootstrap has a negligible bias.

**Theorem 8.** In the heteroscedastic random effect model \((18)\)

\[
\text{Var}(\bar{\mathbf{X}}) = \frac{1}{N} \sum_{u \neq \emptyset} \sum_{i} \nu_{i,u} \sigma_{i,u}^2.
\]

**Theorem 9.** In the heteroscedastic random effects model \((18)\)

\[
\mathbb{E}_{RE}(\text{Var}_{PW}(\bar{\mathbf{X}}^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \sum_{i} \gamma_{i,u} \sigma_{i,u}^2,
\]

where

\[
\gamma_{i,u} = \sum_{k=0}^{r}(1 + \tau^2)^k(\nu_{i,k,u} - 2\nu_{i,k}\nu_{i,u} + \nu_{i,u}).
\]

**Theorem 10.** In the heteroscedastic random effects model \((18)\), the gain coefficient \(\gamma_{i,u}\) of \((21)\) for \(Z_i = 1\) and \(u \neq \emptyset\), in the product reweighted bootstrap is

\[
\gamma_{i,u} = \nu_{i,u}[[1 + \tau^2]|u| - 1 + \theta_u \epsilon] + \sum_{v \supseteq u} (1 + \tau^2)^{|v|}(\tau^2)|v - u|\nu_{i,v}
\]

where \(|\theta_u| \leq (1 + \tau^2)((1 + \tau^2)^r - 1)/\tau^2\). For \(\tau^2 = 1\)

\[
\gamma_{i,u} = \nu_{i,u}[2|u| - 1 + \theta_u \epsilon] + \sum_{v \supseteq u} 2^{|v|}\nu_{i,v}
\]

where \(|\theta_u| \leq 2^{r+1} - 2\).

Theorem 10 establishes that our bootstrap is conservative in the heteroscedastic case. With \(\tau^2 = 1\) we have

\[
\frac{\gamma_{i,u}}{\nu_{i,u}} \geq 2|u| - 1 - (2^{r+1} - 2)\epsilon.
\]

For the homoscedastic random effects model, the main effects dominate when \(\eta = \max_{\emptyset \subseteq v \subseteq u} \frac{\nu_v}{\nu_u} \nu_v/\nu_u\) is small and the variance components are all within the interval \([m,M]\) for \(0 < m \leq M > \infty\). In the heteroscedastic case we might reasonably require every \(\sigma_{i,u}^2 \in [m,M]\). The analysis we used for Theorem 7 also requires the quantities

\[
\eta_i = \begin{cases} 
\max_{\emptyset \subseteq v \subseteq u} \frac{\nu_{i,v}}{\nu_{i,u}} Z_i = 1 \\
0 \quad \quad \quad \quad \quad \quad Z_i = 0
\end{cases}
\]
to be small.

For $r = 2$ the only subsets $u$ and $v$ which appear in $\eta_i$ are $u = \{j\}$ and $v = \{1, 2\}$. Furthermore $\nu_{i,\{1,2\}} = 1/N$ and so

$$\max_i \eta_i = \max_{j \in \{1,2\}} \max_j \frac{N_{i,j}}{N} = \epsilon.$$ 

Then using the same argument we used to prove the second part of Theorem 7 we get

$$\frac{\mathbb{E}_{\text{RE}}(\overline{\text{Var}}_{\text{PW}}(\overline{X}^*))}{\text{Var}(\overline{X})} = 1 + O(\epsilon), \quad \text{for } r = 2.$$ 

The case for $r > 2$ is more complicated. There may be observations $i$ with large values for $\nu_{i,v}/\nu_{i,u}$ where $\emptyset \subsetneq u \subseteq v$. We still get a good approximation from the product reweighted bootstrap because even though the individual $\eta_i$ need not always be small, sums of $\nu_{i,v}$ over $i$ are small compared to corresponding sums of $\nu_{i,u}$ for $\emptyset \subsetneq u \subsetneq v$.

**Theorem 11.** For the heteroscedastic random effects model (18), assume that there exist $m$ and $M$ with $0 < m \leq \sigma^2_{i,u} \leq M < \infty$. Then the product reweighted bootstrap with $\tau^2 = 1$, satisfies

$$\frac{\mathbb{E}_{\text{RE}}(\overline{\text{Var}}_{\text{PW}}(\overline{X}^*))}{\text{Var}(\overline{X})} = 1 + O(\eta + \epsilon).$$

7 Nested random effects

The $r$-fold crossed random effects model (1) excludes replicated observations by definition: there can be only one $X_i$ for any combination $i$ of factors. If two $X$’s are observed to share all index values $i_j$, we can incorporate them by introducing an $r + 1$st index $i_{r+1}$ which breaks the ties. Conditionally on the effects of the first $r$ indices, distinct replicates are independent. That is $\sigma^2_u = 0$ when $r + 1 \in u$ but $u \neq \{1, 2, \ldots, r\}$ or $u \cap \{r\} \neq \{r\}$. The replicate index $i_{r+1}$ is a factor that is nested within the first $r$ factors.

More generally, we could have $s$ additional indices corresponding to factors crossed with each other, but nested within our $r$ outer factors. Then the index $i \in \{1, 2, \ldots\}^{r+s}$ uniquely identifies a data point. Ordinary replication has $s = 1$. The nesting structure means that

$$\sigma^2_u = 0, \quad \text{if } u \cap \{r + 1, \ldots, r + s\} \neq \emptyset \quad \text{and} \quad u \cap \{r\} \neq \{r\}. \quad (22)$$

In words, the effect $\epsilon_{i,u}$ is 0 if the factors in $u$ include any of the inner factors without including all of the outer factors.

When one factor is nested within another, such as replicates within subjects, it is a common practice to resample or reweight the outer factor only. For example, the resampled data set might contain resampled subjects retaining the repeated measurements from each of them.
In the nested setting, the variance of $\bar{X}$ under an $r + s$ factor version of the random effects model $[1]$ is still $(1/N) \sum_{u \neq \emptyset} \nu_u \sigma_u^2$ although many of the $\sigma_u^2$ terms are zero.

For $i \in [r + s]$ let $|i| = (i_1, \ldots, i_r)$ be the indices of its outer factors. We can study these nested models by introducing the variables

$$T_{[i]} = \sum_{i'} Z_{i'} 1_{[i'] = |i|} X_{i'}, \quad \text{and}$$

$$M_{[i]} = \sum_{i'} Z_{i'} 1_{[i'] = |i|},$$

so that the sample mean

$$\bar{X} = \frac{1}{N} \sum_i Z_i X_i = \frac{\sum_{[i]} T_{[i]}}{\sum_{[i]} M_{[i]}}$$

is an $r$-factor ratio estimator.

When the numbers $M_{[i]}$ of replicates for each outer factor vary, we obtain a heteroscedastic random effects model in the first $r$ variables.

### 8 Example: loquacity of Facebook comments

We present an analysis of national differences in comment length on Facebook. In particular, Facebook users can share links with their friends. Their friends, and the posting user, can comment on the link. We compare the length of these comments produced by users in the United States using the site in American English (US users) and those produced by users in the United Kingdom using the site in British English (UK users). We restrict the analysis to US and UK users commenting on links shared by US and UK users. We additionally consider two different modes by which users can comment: the standard web interface to Facebook (web) and an application for some touchscreen mobile phones (mobile).

We treat the logarithm of the number of characters in a comment as the outcome in the following random effects model:

$$X_{cm, i} = \mu_{cm} + \sum_{u \neq \emptyset} \varepsilon_{cm, i, u}$$

where $\mu_{cm}$ is the mean log characters for country $c$ in mode $m$. Here the members of $i$ are indexes for the user sharing the link (sharer), the user commenting on the link (commenter), and the canonicalized URL being shared (URL). By definition, no comments have 0 characters, and so each $X$ in our data set is well defined.

The data consist of $X_{cm, i}$ for a sample of comments by US and UK who are using Facebook in American and British English, respectively, during a short period in 2011. This sample includes 18,134,419 comments by 8,078,531 commenters on 2,085,639 URLs shared by 3,904,715 sharers. We examine whether
these US and UK users post comments of different lengths for both of the modes. The duplication coefficients for this data are

\[ \nu_{\text{sh}} = 17.71, \quad \nu_{\text{com}} = 7.71, \quad \nu_{\text{url}} = 26,854.92 \]

\[ \nu_{\text{sh,com}} = 5.92, \quad \nu_{\text{sh,url}} = 12.91, \quad \nu_{\text{com,url}} = 5.19, \quad \text{and} \]

\[ \nu_{\text{sh,com,url}} = 4.88. \]

The coefficient for URLs is conspicuously large, indicating that a naive bootstrap would be very unreliable.

The sample mean for a country and mode is

\[ \hat{\mu}_{cm} = \frac{\sum_i Z_{cm,i} X_{cm,i}}{\sum_i Z_{cm,i}}. \]

We regard \( \hat{\mu}_{cm} \) as an estimate of \( \mu_{cm} \) conditional on the observed combinations of sharers, commenters, and URLs.

The four sample means for both countries and both modes suggest that the US users write longer comments than UK users when commenting on the web (\( \hat{\mu}_{\text{US}, \text{web}} = 3.62, \hat{\mu}_{\text{UK}, \text{web}} = 3.55 \)), while UK users write longer comments than US users when commenting via the selected mobile interface (\( \hat{\mu}_{\text{US}, \text{mobile}} = 3.5, \hat{\mu}_{\text{UK}, \text{mobile}} = 3.57 \)). Many differences between US and UK users likely contribute to this observed differences. Before searching for causes of these two differences, a data analyst would likely want to quantify the evidence for the existence and size of these differences. We test whether these two pairs of means are likely to be observed given the null hypothesis of no difference in comment length between the countries within each mode.

Using software for Hive (Thusoo et al., 2009), a Hadoop-based map–reduce data warehousing and parallel computing environment, we can compute each of these four means for a number of bootstrap reweightings of the data, while visiting each observation only once. When visiting an observation, the hashed identifiers for the factor levels for that observation are each used as seeds to random number generators. This allows all nodes to use the same \( U\{0,2\} \) draw in computing the product weight for all observations that share a particular factor level. Note that users can be both sharers and commenters. Since users can comment on their own shared links, some observations could have the same factor level identifier for both the sharer and commenter levels. We use different portions of the hashed identifier so that the weights for these two roles are not dependent. For each reweighting, we compute four reweighted sample means

\[ \hat{\mu}_{cm} = \frac{\sum_i Z_{cm,i} W_{cm,i} X_{cm,i}}{\sum_i Z_{cm,i} W_{cm,i}}, \]

corresponding to \( c \in \{\text{US, UK}\} \) and \( m \in \{\text{web, mobile}\} \).

For comparison, we conduct this analysis reweighting one, two, and all three of the factors. Figure 1 presents \( R = 50 \) bootstrapped differences in the two pairs of means when reweighting commenters, commenters and sharers, and all three factors. Inspection of these ECDFs confirms that the observed differences
Figure 1: Difference between the logged number of characters in comments by US and UK users for three different bootstrap reweightings with $R = 50$. Each data point in the plotted ECDF is the difference in means from a single bootstrap reweighting. US users post longer comments than UK users on the web, but this difference is reversed for the mobile interface studied.

Figure 2: Confidence intervals for the difference between the logged number of characters in comments by US and UK users for three different bootstrap reweightings with $R = 50$. Confidence intervals span the 2.5% and 97.5% quantiles of the normal with variance computed from the bootstrap reweightings. While all three analyses reject the null hypothesis, the one- and two-factor analyses may substantially overstate confidence about the size of the true difference, especially in the case of comments posted via the web interface.
cannot be attributed to chance, even when accounting for the random main
and interaction effects of commenters, sharers, and URLs. The bootstrapped
differences in means are strikingly more dispersed for the three-factor analysis.
Figure 2 shows 95% confidence intervals for the two differences computed as
quantiles of the normal distribution with variance computed from the bootstrap
reweightings. This highlights the substantial overstatement of certainty that can
come from neglecting the presence of additional random effects. In this case,
the three analyses would all reject the null hypothesis, but would produce quite
different confidence intervals.

For the approximations developed in Section 5 to apply, we require that
\( \epsilon \) and \( \eta \) be small – that no single level of any random effect make up a large
portion of the observations and that the number of observations matching on \( v \)
is small compared to the number matching on \( u \) factors for all \( \emptyset \subseteq u \subseteq v \). We
find that \( \epsilon = 686,990/18,134,419 = 0.0379 \), as one URL had 686,990 comments
in this sample. We also found that \( \eta = 0.767 \). Because \( \eta \) is not very small it is
possible that the variance estimates are conservative.

9 Discussion

We have worked conditionally on the observed values holding \( Z_i \) fixed. It is
clear that missingness can be informative and thereby introduce a bias into a
sample mean.

The way to correct for missingness and even whether to do such a correction
is problem dependent. In the Netflix data, the company is seeking to predict
ratings that were not made and so the bias between observed and unobserved
ratings is of interest. The people who competed in the Netflix contest were
trying to predict ratings that were actually made and then artificially withheld,
so the pairs to be predicted were not subject to this bias. For the Facebook
data, some of the observed difference between the lengths of comments by US
and UK users may be due to differences in which URLs they comment on. An
accounting of missingness might involve inferring the likely length of comments
that would have been made by US and UK users if they had the same propensity
to comment on particular URLs. An analysis made conditionally on \( Z_i \) describes
the statistical stability of comment lengths for the actual pattern of commenting,
which may be of more interest.

To make an adjustment for missing data requires some kind of assumption
about the missingness mechanism. That assumption cannot be tested within a
given data set because the necessary confirmation values are not available. It
is clear that reweighting cannot correct a sampling bias because many different
sample biases may be consistent with an observed data set. In a given problem
with our preferred adjustment for missingness built in to the statistic of interest,
we could then consider how to bootstrap the resulting bias adjusted statistics.
Alternatively, if the statistic is partially identified, then we could consider how
to bootstrap the resulting sample bounds on the statistic. It is not obvious
which bootstrap method would suit these tasks, but it seems clear that in the
random effects context product reweighting will succeed more generally than naive bootstrapping.

We have used the variance of a sample mean as a way to identify a suitable bootstrap method. Plain sample means are practically important. For example, click through rates, or feature usage rates are means or ratios of means. Even for this simple problem, naive bootstrapping methods are severely downwardly biased in the random effects setting. Product weighting replaces this bias by a small upward bias that is more acceptable in applications.

A bootstrap method that under-estimates the variance of a mean cannot be expected to work well on other problems. One that is properly calibrated or conservative for the variance of a scalar sample mean will also work in some other settings.

The extension to multivariate means is very straightforward. When $X_i \in \mathbb{R}^d$ for $d > 1$ we may replace the variances $\sigma^2_u$ or $\sigma^2_{i,u}$ by variance-covariance matrices $\Sigma_u$ or $\Sigma_{i,u}$ respectively in the variance formulas. This follows by considering the variance of $\phi^T X_i$ for vectors $\phi \in \mathbb{R}^d$.

Bootstrap correctness extends from means to other statistics. See Hall (1992) and Mammen (1992). The extension to smooth functions $g(\bar{X})$ of means is via Taylor expansion, when $g$ has a Jacobian matrix with full rank at $\mathbb{E}(X)$.

The bootstrap is usually used to get confidence intervals, not variance estimates. For an asymptotically unbiased statistic that satisfies a central limit theorem, a properly calibrated variance yields asymptotically correct bootstrap percentile confidence intervals. An overestimated variance yields conservative percentile intervals.

Another way to extend from means to other statistics is via estimating equations. If the parameter $\theta$ is defined by $\sum_i Z_i m(X_i; \theta) = 0$ then we may test the hypothesis that $\theta = \theta_0$ by testing whether $m(X_i; \theta_0)$ has mean zero. In practice one would ordinarily form a histogram of resampled $\hat{\theta}^*$ values and construct a confidence interval from them.

The heteroscedastic random effects model (18) has $2^r - 1$ variance parameters for each observation. Such a model can arise in an $r$–fold generalization of factor analysis. Suppose that $F_{i,u}$ is a nonrandom factor depending on indices in the set $u \subset \{1, 2, \ldots, r\}$ and that $L_{i,v}$ is a mean zero random loading depending on indices in the set $v \subset \{1, 2, \ldots, r\}$ where $u \cap v = \emptyset$. Let

$$X_i = \mu + \cdots + F_{i,u} L_{i,v} + \cdots + \varepsilon_{i,\{1,\ldots,r\}}$$

where the ellipses hide other factors of the type just described for different subsets of the variables. The term shown contributes $F_{i,u}^2 \text{Var}(L_{i,v})$ to the variance component for subset $v$ on observation $i$. Even if the loadings have constant variance, unequal factor values will make this variance component heteroscedastic. The factors and loadings could both have a product form so that they contribute

$$\prod_{j \in u} F_{j,i,j} \times \prod_{j \in v} L_{j,i,j}$$

to $X_i$ generalizing the SVD, but a product form is not necessary.
A generalized factor model would be extremely hard to estimate. However, the total variance from all those different variance contributions is handled by product reweighting, with a small upward bias in the bootstrap variance of a mean. A similar phenomenon is well known in the context of the wild bootstrap [Mammen (1993)] for the linear model. There a different distribution is posited for each of $n$ observations in a regression and the bootstrap process provides reliable inferences for the regression coefficients without having to accurately estimate all $n$ distributions.

**Acknowledgments**

This work was supported by grant DMS-0906056 from the U.S. National Science Foundation and by Nokia. We thank Paul Jones and Jonathan Chang for their assistance. We thank Omkar Muralidharan and the reviewers for helpful comments.

**References**

Bennett, J. and Lanning, S. (2007). The Netflix prize. In *Proceedings of KDD Cup and Workshop 2007*.

Brennan, R. L., Harris, D. J., and Hanson, B. A. (1987). The bootstrap and other procedures for examining the variability of estimated variance components. Technical report, ACT.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Lee, H. K. H. and Clyde, M. A. (2004). Lossless online Bayesian bagging. *Journal of Machine Learning Research*, 5:143–151.

Mammen, E. (1992). *When Does Bootstrap Work?*, volume 77 of *Lecture Notes in Statistics*. Springer, New York.

Mammen, E. (1993). The wild bootstrap. *Annals of Statistics*, 21(1):255–285.

McCullagh, P. (1969). Pseudo-replication: half samples. *Review of the International Statistical Institute*, 37(3):239–264.

McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli*, 6(2):285–301.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (disc: P26-48). *Journal of the Royal Statistical Society, Series B, Methodological*, 56:3–26.
Appendix: proofs

This appendix contains theorem proofs and a few lemmas. The theorems are restated to make it easier to follow the steps. Equation numbers that appear in the theorem statements from the article are preserved in this appendix.

Proof of Theorem 1

Theorem 1. In the random effects model (1)

\[
\text{Var}(\bar{X}) = \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma^2_u.
\]

Proof. The numerator of \( \bar{X} \) in (1) is \( \sum_i Z_i X_i = N \mu + \sum_i \sum_{u \neq \emptyset} Z_i \varepsilon_{i,u} \). Therefore the variance of \( \bar{X} \) under the random effects model is

\[
\text{Var}(\bar{X}) = \frac{1}{N^2} \mathbb{E} \left( \sum_i \sum_{i'} Z_i Z_{i'} \sum_{u \neq \emptyset} \sum_{u' \neq \emptyset} \varepsilon_{i,u} \varepsilon_{i',u'} \right)
\]

\[
= \frac{1}{N^2} \sum_{u \neq \emptyset} \sigma^2_u \sum_{i} \sum_{i'} Z_i Z_{i'} 1_{i_u = i'_u}
\]

\[
= \frac{1}{N^2} \sum_{u \neq \emptyset} \sigma^2_u \sum_{i} Z_i N_{i,u}
\]

\[
= \frac{1}{N} \sum_{u \neq \emptyset} \nu_u \sigma^2_u. \quad \Box
\]
Proof of Theorems 2, 3, and 4.

Here we prove the theorems about naive bootstrap sampling. Theorem 2 is about naive resampling and Theorem 3 handles naive reweighting. Theorem 4 is about bootstrap stability.

**Theorem 2.** Under the random effects model (1), the expected value of the naive bootstrap variance of \( \bar{X} \) is

\[
E_{RE}(\text{Var}_{NB}(\bar{X})) = \frac{1}{N} \sum_{u \neq \emptyset} \sigma_u^2 \left( 1 - \frac{\nu_u}{N} \right).
\]

**Proof.** A U-statistic decomposition of the sample variance is

\[
\text{Var}_{NB}(\bar{X}) = \frac{1}{2N^3} \sum_i \sum_{i'} Z_i Z_{i'} (X_i - X_{i'})^2
\]

\[
= \frac{1}{2N^3} \sum_i \sum_{i'} Z_i Z_{i'} \left( \sum_{u \neq \emptyset} \epsilon_{i,u} - \epsilon_{i',u} \right)^2.
\]

Under the random effects model

\[
E_{RE}(\text{Var}_{NB}(\bar{X})) = \frac{1}{2N^3} \sum_i \sum_{i'} Z_i Z_{i'} \sum_{u \neq \emptyset} 2\sigma_u^2 (1 - 1_{i_u = i'u})
\]

\[
= \frac{1}{N} \sum_{u \neq \emptyset} \sigma_u^2 \left( 1 - \frac{\nu_u}{N} \right). \quad \square
\]

To prove Theorem 3 we begin with a lemma on the covariance of pairs of observations under the random effects model.

**Lemma 1.** Let \( X_i \) follow the random effects model (1) and let \( Y_i = X_i - \bar{X} \). Then

\[
E_{RE}(X_i X_{i'}) = \mu^2 + \sum_{u \neq \emptyset} \sigma_u^2 1_{i_u = i'_u} \tag{23}
\]

and

\[
E_{RE}(Y_i Y_{i'}) = \sum_{u \neq \emptyset} \sigma_u^2 \left( 1_{i_u = i'_u} - \frac{N_{i,u}}{N} - \frac{N_{i',u}}{N} + \frac{\nu_u}{N} \right) \tag{24}
\]

**Proof.** Equation (23) follows directly from the random effects model definition. Expanding \( Y_i Y_{i'} \) yields

\[
X_i X_{i'} - \frac{1}{N} \sum_{i''} Z_{i''} X_i X_{i''} - \frac{1}{N} \sum_{i''} Z_{i''} X_{i'} X_{i''} + \frac{1}{N^2} \sum_{i''} \sum_{i'''} Z_{i''} Z_{i'''} X_i X_{i'} X_{i''}.
\]
Because $\mu$ cancels from $Y_i$ we may assume that $\mu = 0$ while proving (24). Now

$$E_{\text{RE}}\left(\frac{1}{N} \sum_{i'} Z_{i'} X_i X_{i'}\right) = \frac{1}{N} \sum_{u \neq \emptyset} \sigma_u^2 \sum_{i'} Z_{i'} 1_{i_u = i'_u} = \frac{1}{N} \sum_{u \neq \emptyset} \sigma_u^2 N_{i,u}.$$ 

Therefore

$$E_{\text{RE}}(Y_i Y_{i'}) = \sum_{u \neq \emptyset} \sigma_u^2 \left(1_{i_u = i'_u} - \frac{N_{i,u}}{N} - \frac{N_{i',u}}{N} + \frac{1}{N^2} \sum_{i''} Z_{i''} N_{i'',u}\right)$$

which reduces to (24).

**Theorem 3.** In the random effects model (1)

$$E_{\text{RE}}\left(\text{Var}_{\text{NBB}}(\bar{X}^*)\right) = \frac{\tau^2}{N} \sum_{u \neq \emptyset} \sigma_u^2 \left(1 - \frac{\nu_u}{N}\right).$$

**(8)**

**Proof.** Let $Y_i = X_i - \bar{X}$ and $T_y = \sum_i W_i Z_i Y_i$. Then

$$E_{\text{RE}}(\text{Var}_{\text{NBB}}(\bar{X}^*)) = \frac{1}{N^2} E_{\text{RE}}(E_{\text{NBB}}((T^* - \bar{X} N^*)^2))$$

$$= \frac{1}{N^2} E_{\text{RE}}\left(\sum_i \sum_{i'} Z_i Z_{i'} E_{\text{RE}}(Y_i Y_{i'}) E_{\text{NBB}}(W_i W_{i'})\right)$$

$$= \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} E_{\text{RE}}(Y_i Y_{i'}) E_{\text{NBB}}(W_i W_{i'}).$$

Next, $E_{\text{NBB}}(W_i W_{i'}) = 1 + \tau^2 1_{i = i'}$. Therefore

$$E_{\text{RE}}(\text{Var}_{\text{NBB}}(\bar{X}^*)) = \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} E_{\text{RE}}(Y_i Y_{i'}) + \frac{\tau^2}{N^2} \sum_i Z_i E_{\text{RE}}(Y_i^2).$$

(25)

The double sum in (25) vanishes because $\sum_i Z_i Y_i = 0$. Then from Lemma 1 the coefficient of $\sigma_u^2$ in (25) is

$$\frac{\tau^2}{N^2} \sum_i Z_i \left(1 - \frac{2 N_{i,u}}{N} + \frac{\nu_u}{N}\right) = \frac{\tau^2}{N^2} \left(N - 2 \nu_u + \nu_u\right)$$

establishing (8).

**Theorem 4.** Let $W$ and $W_{i,b}$ be IID random variables with mean 1 variance $\tau^2$ and kurtosis $\kappa_w < \infty$. Then holding $Y_i = X_i - \bar{X}$ fixed,

$$\text{Var}_{\text{NBB}}(\bar{X}^*) = \frac{\sigma^4 \tau^4}{BN^2} \left(2 + \frac{\kappa(\kappa_x + 3)}{N}\right)$$

where $\sigma^2 = (1/N) \sum_i Z_i Y_i^2$, and $\kappa_x = (1/N) \sum_i Z_i Y_i^4/\sigma^4 - 3$. A delta method approximation gives

$$\text{Var}_{\text{NBB}}(\hat{s}_N^2) \approx \frac{\sigma^4 \tau^4}{BN^2} \left(\frac{2B}{B - 1} + \frac{\kappa(\kappa_x + 3)}{N}\right).$$
Proof. First, the variance of \( \hat{\text{Var}}_{\text{NBB}}(\bar{X}^*) \) scales as \( 1/B \) so we can work with \( B = 1 \) and divide the result by \( B \). For \( B = 1 \), we drop the subscript \( b \) from \( W \)‘s. We will use the identity \( \sum_i Z_i W_i Y_i = \sum_i Z_i (W_i - 1) Y_i \). If \( B = 1 \), then \( \text{Var}_{\text{NBB}}(\hat{\text{Var}}_{\text{NBB}}(\bar{X}^*)) \) equals

\[
\mathbb{E}_{\text{NBB}}\left( \left( \sum_i Z_i W_i Y_i \right)^4 \right) - \left( \frac{\tau^2 \sigma^2}{N} \right)^2
= \frac{1}{N^4} \sum_i Z_i \mathbb{E}((W_i - 1)^4) Y_i^4 + \frac{3}{N^4} \sum_i \sum_i' Z_i Z_i' \mathbb{E}((W_i - 1)^2) Y_i^2 Y_i'^2
- \frac{3}{N^4} \sum_i Z_i \mathbb{E}((W_i - 1)^2)^2 Y_i^4 - \left( \frac{\sigma^2 \tau^2}{N} \right)^2
= \frac{\tau^4 \sigma^4 (\kappa + 3)(\kappa_x + 3)}{N^3} + \frac{3 \tau^4 \sigma^4}{N^2} - \frac{3 \tau^4 \sigma^4 (\kappa_x + 3)}{N^3} - \frac{\sigma^4 \tau^4}{N^2}
= \frac{\tau^4 \sigma^4}{N^2} \left( 2 + \frac{\kappa (\kappa_x + 3)}{N} \right).
\]

For the second part

\[
\text{Var}_{\text{NBB}}(s_{\text{NBB}}^2) = \mathbb{E}_{\text{NBB}}(s_{\text{NBB}}^2)^2 \left( \frac{2}{B - 1} + \frac{\kappa^*}{B} \right)
\]

where \( \kappa^* \) is the kurtosis of \( \bar{X}^* = \sum_i Z_i W_i Y_i / \sum_i Z_i W_i \). The delta method approximation to \( \mathbb{E}_{\text{NBB}}(s_{\text{NBB}}^2) \) is \( \tau^2 \sigma^2 / N \). For the kurtosis, we make the Taylor approximation

\[
\bar{X}^* = \bar{X} + \sum_i Z_i (W_i - 1) Y_i.
\]

The expected value of \( \bar{X}^* - \bar{X} \) reuses much of the above computation and yields

\[
\mathbb{E}_{\text{NBB}}((\bar{X}^* - \bar{X})^4) = \frac{\tau^4 \sigma^4}{N^2} \left( 3 + \frac{\kappa (\kappa_x + 3)}{N} \right).
\]

Therefore \( \kappa^* = \kappa (\kappa_x + 3) / N \) and so

\[
\text{Var}_{\text{NBB}}(s_{\text{NBB}}^2) = \frac{\tau^4 \sigma^4}{BN^2} \left( \frac{2B}{B - 1} + \frac{\kappa (\kappa_x + 3)}{N} \right). \quad \square
\]

Proof of Theorems 5, 6, and 7.

Theorem 5 gives an exact expression for the gain coefficients of the Bayesian pigeonhole bootstrap in the constant variance crossed random effects model. Theorem 6 gives an interpretable approximation to those gain coefficients. Theorem 7 shows factorial reweighting gives nearly the correct variance when \( \epsilon \) and \( \eta \) are both small.

26
Theorem 5. In the random effects model (1)

\[ E_{RE}(\widetilde{\text{Var}}_{PW}(X^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \gamma_u \sigma_u^2, \]

where

\[ \gamma_u = \sum_{k=0}^{r} (1 + \tau^2)^k (\nu_{k,u} - 2\nu_{k,u} + \rho_k \nu_u). \]  \hfill (15)

Proof. We begin along the same lines as Theorem 3 and find that

\[ E_{RE}(\widetilde{\text{Var}}_{PW}(X^*)) = \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} E_{RE}(Y_i Y_{i'}) E_{PW}(W_i W_{i'}). \]

For the product weights used in this bootstrap,

\[ E_{PW}(W_i W_{i'}) = \prod_{j: i_j = i'_j} (1 + \tau^2) = (1 + \tau^2)^{|M_{ii'}|}, \]

with \( E_{PW}(W_i W_{i'}) = 1 \) if \( i \) and \( i' \) are not equal in any components.

From Lemma 1, the coefficient of \( \sigma_u^2 \) in \( E_{RE}(\widetilde{\text{Var}}_{PW}(X^*)) \) is

\[ \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} \left( 1_{i_u=i'_u} - \frac{N_{i,u}}{N} - \frac{N_{i',u}}{N} + \frac{\nu_u}{N} \right) (1 + \tau^2)^{|M_{ii'}|} \]

\[ = \frac{1}{N^2} \sum_i \sum_{i'} Z_i Z_{i'} \left( 1_{i_u=i'_u} - \frac{2N_{i,u}}{N} + \frac{\nu_u}{N} \right) (1 + \tau^2)^{|M_{ii'}|} \]

\[ = \frac{1}{N^2} \sum_{k=0}^{r} (1 + \tau^2)^k \sum_i \sum_{i'} Z_i Z_{i'} \left( 1_{M_{ii'}=k} - \frac{2N_{i,u}}{N} + \frac{\nu_u}{N} \right) \]

\[ = \frac{1}{N} \sum_{k=0}^{r} (1 + \tau^2)^k (\nu_{k,u} - 2\nu_{k,u} + \rho_k \nu_u). \]

Next we establish an interpretable approximation to the Bayesian pigeonhole bootstrap variance, using the quantity \( \varepsilon = \max_i \max_j N_{i,(j)}/N \) which is small unless the data are extremely imbalanced.

Theorem 6. In the random effects model (1), the gain coefficient (15) for \( u \neq \emptyset \) in the product reweighted bootstrap is

\[ \gamma_u = \nu_u [(1 + \tau^2)^{|u|} - 1 + \theta_u \varepsilon] + \sum_{v \supseteq u} (1 + \tau^2)^{|v|} (\tau^2)^{|v-u|} \nu_v, \]  \hfill (16)

where \( |\theta_u| \leq (1 + \tau^2) ((1 + \tau^2)^r - 1)/\tau^2. \) For \( \tau^2 = 1, \)

\[ \gamma_u = \nu_u [2^{|u|} - 1 + \theta_u \varepsilon] + \sum_{v \supseteq u} 2^{|v|} \nu_v, \]

where \( |\theta_u| \leq 2^{r+1} - 2. \)
Proof. The second claim follows immediately from the first which we now prove. We will approximate 
\( \gamma_u = \sum_{k=0}^r (1 + \tau^2)^k(\nu_{k,u} - 2\bar{\nu}_{k,u} + \rho_k \nu_u) \). First

\[
\sum_{k=0}^r (1 + \tau^2)^k \nu_{k,u} = \frac{1}{N} \sum_{k=0}^r (1 + \tau^2)^k \sum_i \sum_{i'} Z_i Z_{i'} 1_{|M_{i,i'}| = k} 1_{i = i'} \\
= \frac{1}{N} \sum_{w \supseteq u} (1 + \tau^2)^{|w|} \sum_i \sum_{i'} Z_i Z_{i'} 1_{M_{i,i'} = w} \\
= \frac{1}{N} \sum_{w \supseteq u} (1 + \tau^2)^{|w|} \sum_i \sum_{i'} Z_i Z_{i'} \sum_{v \supseteq w} (-1)^{|v-w|} 1_{i = i'} \\
= \sum_{w \supseteq u} (1 + \tau^2)^{|w|} \sum_{v \supseteq w} (-1)^{|v-w|} \nu_v.
\]

Writing \( w \in [u,v] \) for \( u \subseteq w \subseteq v \),

\[
\sum_{w \supseteq u} (1 + \tau^2)^{|w|} \sum_{v \subseteq w} (-1)^{|v-w|} \nu_v \\
= \sum_{v \supseteq u} \sum_{w \in [u,v]} (1 + \tau^2)^{|w|} (-1)^{|v-w|} \\
= \sum_{v \supseteq u} \nu_v \sum_{w \in [u,v]} \frac{|w-u|}{\ell} \left(1 + \tau^2\right)^{|v-\ell|} \\
= \sum_{v \supseteq u} \nu_v \left(1 + \tau^2\right)^{|v|} \frac{|v-u|}{(\tau^2)^{|v-u|}}.
\]

For the other parts of \( \gamma_u \), we use quantities \( \theta \) that satisfy bounds \( 0 \leq \theta \leq 1 \). There are several such quantities, distinguished by subscripts, and defined at their first appearance. First, we have the bounds

\[
\frac{N_i,0}{N} = 1 - r\theta_{i,0}\epsilon, \quad \text{and} \quad \frac{N_i,k}{N} = \theta_{i,k}\epsilon, \quad 1 \leq k \leq r. \quad (26)
\]

Next, for \( u \neq \emptyset \),

\[
\bar{\nu}_{0,u} = \frac{1}{N^2} \sum_i Z_i N_{i,u} N_{i,0} = \frac{1}{N} \sum_i Z_i N_{i,u} (1 - r\theta_{i,0}\epsilon) = \nu_u (1 - r\theta_{0,u}\epsilon)
\]

and for \( k = 1, \ldots, r \)

\[
\bar{\nu}_{k,u} = \frac{1}{N^2} \sum_i Z_i N_{i,u} N_{i,k} = \frac{1}{N} \sum_i Z_i N_{i,u} \theta_{i,k}\epsilon = \nu_u \theta_{k,u}\epsilon,
\]

Turning to \( \rho_k \),

\[
\rho_0 = \frac{1}{N^2} \sum_i Z_i N_{i,0} = \frac{1}{N} \sum_i Z_i (1 - \epsilon r\theta_{i,0}) = 1 - \epsilon r\theta_0, \quad \text{and} \n\]

\[
\rho_k = \frac{1}{N^2} \sum_i Z_i N_{i,k} = \frac{1}{N} \sum_i Z_i \theta_{i,k}\epsilon = \theta_k\epsilon, \quad k = 1, \ldots, r.
\]
Now \(-2\bar{\nu}_{0,u} + \rho_0 u = -\nu_u + \nu_u (2\theta_{0,u} - \theta_0) r \epsilon\) and
\[
\sum_{k=1}^{r} (1 + \tau^2)^k (-2\bar{\nu}_{k,u} + \rho_k u) = \nu_u \sum_{k=1}^{r} (1 + \tau^2)^k (\theta_k - 2\theta_{k,u}) \epsilon
\]
Therefore
\[
\gamma_u = \nu_u ((1 + \tau^2)^{|u|} - 1 + \theta_u \epsilon) + \sum_{v \supseteq u} \nu_v (1 + \tau^2) (\tau^2)^{|v-u|}, \quad \text{where}
\]
\[
\theta_u = \sum_{k=1}^{r} (1 + \tau^2)^k (\theta_k - 2\theta_{k,u}).
\]
The proof follows because \(-1 \leq \theta_k - 2\theta_{k,u} \leq 1\) and \(\sum_{k=1}^{r} (1 + \tau^2)^k = (1 + \tau^2)((1 + \tau^2)^r - 1)/\tau^2\).

**Theorem 7**. For the random effects model \(\text{(1)}\) and the product reweighted bootstrap with \(\tau^2 = 1\), the gain coefficient for nonempty \(u \subseteq [r]\) satisfies
\[
2^{|u|} - 1 - (2^{r+1} - 2) \epsilon < \frac{\gamma_u}{\nu_u} \leq 2^{|u|}(1 + 2\eta)^{|v-u|} - 1 + (2^{r+1} - 2) \epsilon.
\]
If there exist \(m\) and \(M\) with \(0 < m \leq \sigma_u^2 \leq M < \infty\) for all \(u \neq \emptyset\), then
\[
\frac{E_{RE}(\text{Var}_{PW}(\bar{X}^*))}{\text{Var}(\bar{X})} = 1 + O(\eta + \epsilon).
\]

**Proof.** From Theorem 6
\[
\frac{\gamma_u}{\nu_u} \leq -1 + \sum_{v \supseteq u} 2^{|u|}\eta^{|v-u|} + (2^{r+1} - 2) \epsilon
\]
\[
= 2^{|u|}(1 + 2\eta)^{|v-u|} - 1 + (2^{r+1} - 2) \epsilon,
\]
and then using \(\nu_v > 0\),
\[
\frac{\gamma_u}{\nu_u} \geq 2^{|u|} - 1 - (2^{r+1} - 2) \epsilon.
\]

For the second claim, small \(\eta\) means that the variance is dominated by contributions \(\sigma_{u}^2\) for which \(\gamma_{(j)} \approx \nu_{(j)}\). Now
\[
\sum_{|u|=1} \gamma_u \sigma_u^2 = \sum_{|u|=1} \nu_u \sigma_u^2 [1 + O(\eta + \epsilon)]
\]
where the constant in \(O(\cdot)\) can depend on \(r\), and
\[
\sum_{|u|>1} \gamma_u \sigma_u^2 = \sum_{|u|>1} \nu_u \sigma_u^2 [2^{|u|} + O(\eta + \epsilon)] = O(\eta) \sum_{|u|=1} \nu_u \sigma_u^2.
\]
Similarly $\sum_{|u|>1} \gamma_u \sigma_u^2 = O(\eta) \sum_{|u|=1} \nu_u \sigma_u^2$. Therefore

$$\mathbb{E}_{RE}\left(\frac{\text{Var}_{PW}(\tilde{X}^*)}{\text{Var}(X)}\right) = \frac{(1 + O(\eta + \epsilon)) \sum_{|u|=1} \nu_u \sigma_u^2}{(1 + O(\eta)) \sum_{|u|=1} \nu_u \sigma_u^2} = 1 + O(\eta + \epsilon).$$

\[ \square \]

**Proof of Theorems 8 through 11.**

Here we prove the theorems for the heteroscedastic case. We begin with a lemma.

**Lemma 2.** Let $X_i$ follow the heteroscedastic random effects model \[18\] and let $Y_i = X_i - \bar{X}$. Then

$$\mathbb{E}_{RE}(X_iX_i') = \mu^2 + \sum_{u \neq \emptyset} \sigma_{i,u}^2 1_{i=\nu_a}$$

and

$$\mathbb{E}_{RE}(Y_iY_i') = \sum_{u \neq \emptyset} \left(1_{i=\nu_a} \sigma_{i,u}^2 - \nu_{i,u} \sigma_{i,u}^2 - \nu_{i',u} \sigma_{i',u}^2 + \nu_a \sigma_a^2\right).$$

**Proof.** Equation \[27\] follows directly just as the analogous expression did in Lemma \[1\]. Once again, expanding $Y_iY_i'$ yields

$$X_iX_i' - \frac{1}{N} \sum_{\nu'} Z_{i''}X_iX_{i''} - \frac{1}{N} \sum_{\nu'} Z_{i'''}X_iX_{i'''} + \frac{1}{N^2} \sum_{\nu'} \sum_{\nu''} Z_{i'''}Z_{i''}X_iX_{i''}X_{i''''}.$$ and we may assume that $\mu = 0$ while proving \[28\]. Now

$$\mathbb{E}_{RE}\left(\frac{1}{N} \sum_{\nu'} Z_{i''}X_iX_{i''}\right) = \frac{1}{N} \sum_{u \neq \emptyset} \sum_{\nu'} Z_{i''} 1_{i=\nu_a} \sigma_{i,u}^2 = \sum_{u \neq \emptyset} \sum_{i} \sigma_{i,u}^2 \nu_{i,u},$$ and

$$\mathbb{E}_{RE}\left(\frac{1}{N^2} \sum_{\nu''} \sum_{\nu'''} Z_{i'''}Z_{i''}X_iX_{i''}X_{i''''}\right) = \frac{1}{N^2} \sum_{u \neq \emptyset} \sum_{\nu''} \sum_{i} Z_{i'''}Z_{i''} 1_{i''=\nu_a} \sigma_{i,u}^2$$

$$= \frac{1}{N} \sum_{u \neq \emptyset} \sum_{\nu''} Z_{i''} \sigma_{i,u}^2 \nu_{i,u}$$

$$= \sum_{u \neq \emptyset} \nu_u \sigma_u^2,$$ which together establish \[28\].

**Theorem 8.** In the heteroscedastic random effect model \[18\]

$$\text{Var}(\tilde{X}) = \frac{1}{N} \sum_{u \neq \emptyset} \sum_{\nu} \nu_{i,u} \sigma_{i,u}^2.$$
Proof. The proof is very similar to that of Theorem 1.

**Theorem 9.** In the heteroscedastic random effects model (18)
\[
\mathbb{E}_{RE}(\tilde{\text{Var}}_{PW}(X^*)) = \frac{1}{N} \sum_{u \neq \emptyset} \sum_{i} \gamma_{i,u} \sigma_{i,u}^2,
\]
where
\[
\gamma_{i,u} = \sum_{k=0}^{r}(1 + \tau^2)^k(\mu_{i,k,u} - 2\mu_{i,k} + \varphi_k \nu_{i,u}).
\]

Proof. We begin along the same lines as Theorem 3 and find that
\[
\mathbb{E}_{RE}(\tilde{\text{Var}}_{PW}(X^*)) = \frac{1}{N^2} \sum_{i} \sum_{i'} Z_i Z_{i'} \mathbb{E}_{RE}(Y_i Y_{i'}) \mathbb{E}_{PW}(W_i W_{i'}).
\]
As in Theorem 5, \( \mathbb{E}_{PW}(W_i W_{i'}) = (1 + \tau^2)^{|M_{ii'}|} \).

From Lemma 2,
\[
\mathbb{E}_{RE}(\tilde{\text{Var}}_{PW}(X^*)) = \frac{1}{N^2} \sum_{u \neq \emptyset} \sum_{i} \sum_{i'} Z_i Z_{i'} (1 + \tau^2)^{|M_{ii'}|}
\times \left( 1_{i'=i} \sigma_{i,u}^2 - \nu_{i,u} \sigma_{i,u}^2 - \nu_{i',u} \sigma_{i',u}^2 + \varphi_k \nu_{i,u}^2 \right).
\]

The contribution from the last term in the parentheses of (30) is
\[
\frac{1}{N} \sum_{u \neq \emptyset} \nu_{i,u} \sigma_{i,u}^2 \sum_{k=0}^{r} (1 + \tau^2)^k \sum_{i} Z_i \nu_{i,k} = \frac{1}{N} \sum_{u \neq \emptyset} \nu_{i,u} \sigma_{i,u}^2 \sum_{k=0}^{r} (1 + \tau^2)^k \varphi_k.
\]
Therefore the coefficient of \( \sigma_{i,u}^2 \) in \( \mathbb{E}_{RE}(\tilde{\text{Var}}_{PW}(X^*)) \) (when \( Z_i = 1 \)) is
\[
\frac{1}{N} \sum_{i} Z_i \sum_{k=0}^{r} 1_{|M_{ii'}|=k} (1 + \tau^2)^k \left( 1_{i'=i} - 2\nu_{i,u} + \frac{\nu_{i,u}}{N} \sum_{k=0}^{r} (1 + \tau^2)^k \varphi_k \right)
\]
\[
= \frac{1}{N} \sum_{k=0}^{r} (1 + \tau^2)^k (\nu_{i,k,u} - 2\nu_{i,k} + \varphi_k \nu_{i,u}).
\]

**Theorem 10.** In the heteroscedastic random effects model (18), the gain coefficient \( \gamma_{i,u} \) of (21) for \( Z_i = 1 \) and \( u \neq \emptyset \), in the product reweighted bootstrap is
\[
\gamma_{i,u} = \nu_{i,u} [(1 + \tau^2)^{|u|} - 1 + \theta_u \varepsilon] + \sum_{v \geq u} (1 + \tau^2)^{|u|} (\tau^2)^{|v-u|} \nu_{i,v}
\]
where \( |\theta_u| \leq (1 + \tau^2)/((1 + \tau^2)^{r-1})/\tau^2 \). For \( \tau^2 = 1 \)
\[
\gamma_{i,u} = \nu_{i,u} [2^{|u|} - 1 + \theta_u \varepsilon] + \sum_{v \geq u} 2^{|v|} \nu_{i,v}
\]
where \( |\theta_u| \leq 2^{r+1} - 2 \).
Proof. From Theorem 9, \( \gamma_{i,u} = \sum_{k=0}^{r} (1 + \tau^2)^k (\nu_{i,k,u} - 2\nu_{i,k}\nu_{i,u} + \nu_{i}^2\nu_{i,u}) \). The proof is similar to that of Theorem 6, so we summarize the steps. First

\[
\sum_{k=0}^{r} (1 + \tau^2)^k \nu_{i,k,u} = \sum_{v \supseteq u} v \nu_{i,v} (1 + \tau^2)^{|v| (\tau^2)^{|u-w|}].
\]

Next, \( \nu_{i,0} = 1 - r\theta_{i,0}\epsilon \) and \( \nu_{i} = 1 - r\theta_{i}\epsilon \), while for \( k \geq 1 \), \( \nu_{i,k} = \theta_{i,k}\epsilon \) and \( \nu_{k} = \theta_{k}\epsilon \). Here, all of the \( \theta \)'s are in the interval \([0, 1]\). The result follows as in Theorem 6.

\textbf{Theorem 11.} For the heteroscedastic random effects model (18), assume that there exist \( m \) and \( M \) with \( 0 < m \leq \sigma_{i,u}^2 \leq M < \infty \). Then the product reweighted bootstrap with \( \tau^2 = 1 \), satisfies

\[
\mathbb{E} \left( \text{RE} \left( \hat{\text{Var}}_{PW}(\bar{X}^*) \right) \right) / \text{Var}(\bar{X}) = 1 + O(\eta + \epsilon).
\]

Proof. First we show that main effects dominate. For \(|u| > 1\),

\[
\sum_{i} \gamma_{i,u} \sigma_{i,u}^2 \leq M \sum_{i} \nu_{i,u} (2|u| - 1 + 2r+1\epsilon) + \sum_{v \supseteq u} 2|v| \nu_{i,v}
\]

\[
= M \left( \nu_{u} (2|u| - 1 + 2r+1\epsilon) + \sum_{v \supseteq u} 2|v| \nu_{v} \right)
\]

\[
= (2|u| - 1) M \nu_{u} (1 + O(\epsilon))
\]

\[
= O(\eta) \max_{1 \leq j \leq r} \nu_{(j)},
\]

and similarly \( \sum_{i} \nu_{i,u} \sigma_{i,u}^2 = O(\eta) \max_{1 \leq j \leq r} \nu_{(j)} \). For \( u = \{j\} \),

\[
\sum_{i} \gamma_{i,\{j\}} \sigma_{i,\{j\}}^2 \geq m \sum_{i} \nu_{i,\{j\}} (1 - 2r+1\epsilon)
\]

\[
= m\nu_{(j)} (1 + O(\epsilon)).
\]

Therefore

\[
\mathbb{E} \left( \text{RE} \left( \hat{\text{Var}}_{PW}(\bar{X}^*) \right) \right) / \text{Var}(\bar{X}) = \frac{\sum_{i,j=1}^{r} \gamma_{i,\{j\}} \sigma_{i,\{j\}}^2}{\sum_{i,j=1}^{r} \nu_{i,\{j\}} \sigma_{i,\{j\}}^2} (1 + O(\eta + \epsilon)).
\]

Next we show that the main effects are properly estimated

\[
\sum_{i} \sum_{j=1}^{r} |\gamma_{i,\{j\}} - \nu_{i,\{j\}}| \sigma_{i,\{j\}}^2 \leq M \sum_{i} \sum_{j=1}^{r} |\gamma_{i,\{j\}} - \nu_{i,\{j\}}|
\]

\[
\leq M \sum_{i} \sum_{j=1}^{r} \nu_{i,\{j\}} (2r+1\epsilon + 3r\eta)
\]

\[
= \sum_{j=1}^{r} \nu_{(j)} O(\eta + \epsilon),
\]

32
while $\sum_i \sum_{j=1}^r \nu_{i,(j)} \sigma_{i,(j)}^2 \geq m \sum_{j=1}^r \nu_{(j)}$.