Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Rank Conditional Coverage and Confidence Intervals in High-Dimensional Problems

Jean Morrison & Noah Simon

To cite this article: Jean Morrison & Noah Simon (2018) Rank Conditional Coverage and Confidence Intervals in High-Dimensional Problems, Journal of Computational and Graphical Statistics, 27:3, 648-656, DOI: 10.1080/10618600.2017.1411270

To link to this article: https://doi.org/10.1080/10618600.2017.1411270

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2018, VOL. 27, NO. 3, 648–656
https://doi.org/10.1080/10618600.2017.1411270

Rank Conditional Coverage and Confidence Intervals in High-Dimensional Problems

Jean Morrison[a] and Noah Simon[b]

[a]Department of Human Genetics, University of Chicago, Chicago, IL; [b]Department of Biostatistics, University of Washington, Seattle, WA

ABSTRACT
Confidence interval procedures used in low-dimensional settings are often inappropriate for high-dimensional applications. When many parameters are estimated, marginal confidence intervals associated with the most significant estimates have very low coverage rates: They are too small and centered at biased estimates. The problem of forming confidence intervals in high-dimensional settings has previously been studied through the lens of selection adjustment. In that framework, the goal is to control the proportion of noncovering intervals formed for selected parameters. In this article, we approach the problem by considering the relationship between rank and coverage probability. Marginal confidence intervals have very low coverage rates for the most significant parameters and high rates for parameters with more boring estimates. Many selection adjusted intervals have the same behavior despite controlling the coverage rate within a selected set. This relationship between rank and coverage rate means that the parameters most likely to be pursued further in follow-up or replication studies are the least likely to be covered by the constructed intervals. In this article, we propose rank conditional coverage (RCC) as a new coverage criterion for confidence intervals in multiple testing/covering problems. The RCC is the expected coverage rate of an interval given the significance ranking for the associated estimator. We also propose two methods that use bootstrapping to construct confidence intervals that control the RCC. Because these methods make use of additional information captured by the ranks of the parameter estimates, they often produce smaller intervals than marginal or selection adjusted methods. These methods are implemented in R (R Core Team, 2017) in the package `rcc` available on CRAN at *https://cran.r-project.org/web/packages/rcc/index.html*. Supplementary material for this article is available online.

ARTICLE HISTORY
Received July 2016
Revised August 2017

KEYWORDS
Bootstrap/resampling; Multiple comparisons; Selective inference; Winner's curse

1. Introduction

In many fields including genomics, proteomics, biomedical science, and neurology it is common to conduct "high-dimensional" studies in which thousands or millions of parameters are estimated. Often, one of the main goals of these studies is to select or prioritize a small subset of features for description and further investigation. Estimates of selected parameters are often reported unadjusted and, when confidence intervals are not entirely omitted, either marginal or Bonferroni-corrected intervals are given. Many previous authors have demonstrated the undesirable features of these practices (Sun and Bull 2005; Efron 2011; Simon and Simon 2013 among others). In particular, parameter estimates selected for reporting are usually chosen because they are the largest or most significant. Unfortunately, these most extreme estimates are also highly biased. This is known informally as the "winner's curse": Large statistics tend to come from large parameters, but they also tend to be large by chance.

There is a related phenomenon for confidence intervals: Marginal confidence intervals almost always fail to cover parameters associated with the most significant estimates because they fail to account for the bias of these estimates. This results in overly short intervals that are too far from zero. Often this problem is recognized by investigators, but the most commonly used alternative, the Bonferroni correction, yields enormous, uninformative intervals. Problems with the Bonferroni correction have been described by Benjamini and Yekutieli (2005), Zhong and Prentice (2008), Weinstein, Fithian, and Benjamini (2013) and several others. This method does not recenter the intervals and inflates their size symmetrically to control the family-wise error rate. Intuitively, we know that the largest estimates are more often too large than too small so most of the upper extension provided by the Bonferroni confidence intervals is unnecessary. Furthermore, the family-wise error rate is a much more conservative criterion than is typically desired.

We will discuss several alternative coverage criteria for high-dimensional settings and introduce a new criterion, rank conditional coverage (RCC), which is the expected coverage probability of an interval given the ranking of its corresponding estimate. The RCC captures the idea that, when many parameters are estimated, the rank of an estimate provides information about its bias and the coverage probability of the associated confidence interval. This criterion applies to all high-dimensional studies regardless of whether or not selection is performed but is particularly relevant when rank is used as a selection criterion or parameters are prioritized by the ranks of their estimates. One advantage to obtaining confidence intervals that control the RCC, rather than selection adjusted confidence intervals, is

CONTACT Jean Morrison jeanm@uchicago.edu Department of Human Gentetics, University of Chicago, Chicago, IL 60637, USA.
Color versions of one or more of the figures in the article are available online at www.tandfonline.com/r/JCGS.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

that these intervals are not dependent on a particular selection procedure—that is, the interval calculated for a particular estimate will not change if the selection threshold is moved. We discuss parametric and nonparametric bootstrap-based approaches for obtaining RCC controlling confidence intervals and explore their behavior and RCC in a few common settings via simulation. The proposed methods are implemented in R (R Core Team 2017) and available in the package `rcc` on CRAN at *https://cran.r-project.org/web/packages/rcc/index.html*.

1.1. Coverage Criteria After Selection

Consider an analysis in which we would like to obtain estimates and confidence intervals for parameters $\theta_1, \dots, \theta_p$. Suppose we have point estimates for each parameter $\hat{\theta}_1, \dots, \hat{\theta}_p$ and a *marginally valid* procedure for constructing confidence intervals. By marginally valid we mean that for each parameter θ_j, the coverage probability of the α level confidence interval CI_j satisfies $P[\theta_j \in CI_j(\alpha)] = 1 - \alpha$.

In the high-dimensional setting, the marginal confidence interval will control the *average coverage*—that is, if we construct 90% marginal confidence intervals, we can expect them to cover 90% of the parameters. Typically, however, the entire set of parameter estimates is not of interest and only the most significant estimates are reported. Benjamini and Yekutieli (2005) demonstrated that, for common selection rules, the expected rate of coverage for marginal confidence intervals within a selected subset of parameters will be much lower than the nominal level, α. The marginal confidence interval achieves its average coverage by under-covering parameters associated with the most extreme or significant estimates and over-covering more boring parameters.

To quantify coverage of confidence intervals constructed for a selected set, Benjamini and Yekutieli (2005) proposed the false coverage statement rate (FCR), which measures the expected proportion of noncovering intervals within the selected set. If no parameters are selected, it is counted as no false statements made.

$$\text{False Coverage Statement Rate} = E[Q] \qquad (1)$$

$$Q = \begin{cases} \frac{\sum_{j \in S} 1_{\theta_j \notin CI_j}}{|S|} & |S| > 0 \\ 0 & |S| = 0. \end{cases}$$

Benjamini and Yekutieli (2005) proposed a procedure providing FCR controlling confidence intervals which, like the Bonferroni procedure, symmetrically inflates the size of marginally valid intervals for all parameters. When selection is based on parameter estimates exceeding a threshold, these are equivalent to coverage $(1 - \frac{|S|\alpha}{p})$ marginal intervals. This uniform inflation can be excessive in some cases. For example, if one parameter is very large, it will nearly always be selected. Thus, there is no need to inflate that interval at all (a $1 - \alpha$ marginal interval will still control FCR). Additionally, the fact that highest ranked estimates are more often too large than too small suggests that confidence intervals should have longer tails extending toward the bulk of the estimates than extending away from the bulk.

These issues are partially accounted for in more recent literature: Zhong and Prentice (2008), Weinstein, Fithian, and Benjamini (2013), and Reid, Taylor, and Tibshirani (2017) all propose FCR controlling intervals which return to the marginal interval for very large parameter estimates. Zhong and Prentice (2008) and Weinstein, Fithian, and Benjamini (2013) both condition on selecting all estimates larger than a (possibly data-dependent) cutoff. Zhong and Prentice (2008) use a likelihood-based approach to obtain asymptotically correct FCR while Weinstein, Fithian, and Benjamini (2013) calculate exact intervals under the assumption that parameter estimates are independent with a known symmetric unimodal distribution. Reid, Taylor, and Tibshirani (2017) conditioned on the identity of the selected set and construct exact intervals for finite sample sizes assuming that parameter estimates are drawn from independent Gaussian distributions. All three of these intervals are asymmetric about the original point estimate.

Despite these advances, controlling FCR remains an unsatisfying solution to confidence interval construction for high-dimensional problems. FCR controlling methods achieve the correct coverage rate within a subset in the same way that the marginal intervals achieve the correct rate in the larger set—with under-coverage of the (more interesting) highest ranked parameters and over-coverage of (less interesting) more moderately ranked parameters. We illustrate this pattern through a simple example in Section 1.5

1.2. Rank Conditional Coverage

In Section 1.1, we observed that for unadjusted confidence intervals as well as the selection adjusted alternatives, the rank of an estimate is informative about the probability that the associated confidence interval covers its target (see also Section 1.5). We find this phenomenon undesirable since it means that parameters associated with top ranked estimates are covered at a much lower rate than parameters associated with less significant estimates. Additionally, this observation indicates that there is an opportunity to use more information and construct better intervals.

We first introduce the concept of rank conditional coverage (RCC) as a way to quantify the relationship between rank and coverage probability. In the majority of cases, the most interesting ranking of parameters is based on either the size of an associated test statistic or a p-value. In general, we assume that we have some ranking function s where $s(i)$ gives the index of the ith ranked estimate. For example, if we are ranking simply based on the size of estimates, then

$$\hat{\theta}_{s(1)} \geq \cdots \geq \hat{\theta}_{s(p)}.$$

In this article, we will use the convention that a smaller rank indicates that an estimate is more significant, so the most significant estimate will have rank 1. In our examples, we focus on simple common rankings but the RCC can be defined for any scheme. In fact, the ranking scheme need not give a rank to every estimate. For example, if the parameter estimates can be grouped into highly correlated subsets such as LD blocks in a genetic study, we might choose the most significant estimate in each block and rank only within this selected set. This type of ranking scheme is discussed at greater length along with simulation results in Section 3 of the Appendix.

We define the RCC at rank i of a set of confidence intervals $CI_1 \ldots CI_p$ as

$$\text{Rank Conditional Coverage}_i = P[\theta_{s(i)} \in CI_{s(i)}] \quad (2)$$

$$= \sum_{j=1}^{p} P[\theta_j \in CI_j | s(i) = j] \cdot P[s(i) = j]. \quad (3)$$

This quantifies how often the interval formed around the ith ranked estimate contains its target parameter. This is an appealing criterion, since we have a strong interest in ensuring that intervals around our most promising candidate features contain their targets. Something to note here is that we are not conditioning on which specific features achieve a given rank. Rather, we are averaging over all features, weighted by their probability of achieving that rank. While FCR summarizes the average coverage of a confidence interval procedure applied to a set of selected parameters in a single number, RCC gives a separate estimate of coverage probability for each rank and is not directly related to a selection procedure.

1.3. Implications of Controlling RCC

Intervals that control RCC do not provide guarantees for particular parameters. For example, suppose θ_1 is of special interest. If we use an RCC controlling method with level α, we cannot say that, if the experiment were repeated many times, the proportion of experiments for which θ_1 is contained in CI_1 is expected to be $1 - \alpha$. Thus, if there is particular prior interest on one or a few parameters, RCC is not the correct criterion to control.

Using RCC controlling intervals, we are guaranteed that the expected proportion of experiments for which the top ranked parameter is covered by its interval is $1 - \alpha$. A similar statement could be made for any rank. While this property may seem less intuitive on its surface, it has important implications when parameters are prioritized based on the ranks of their estimates.

For example, suppose that a researcher publishes the results of many genome-wide association studies, each time reporting the most significant effect size estimates. If these estimates were paired with confidence intervals controlling the RCC at 90%, the researcher could expect that 90% of the published intervals for 1st ranked estimates (or 2nd, etc.), averaging over studies, contain their parameters. Most follow-up studies are conducted specifically for the most promising parameters, so this is precisely the type of guarantee needed to ensure these follow-up studies are worthwhile.

This guarantee is different from the guarantee made by the FCR. We find in simulations that FCR controlling methods typically do not control the RCC and that, using these, top ranked parameters are less likely to be covered than lower ranked parameters within the selected set. Conversely, for selection rules that choose a fixed number of the top ranked parameters, confidence intervals that control RCC for every rank also control FCR. For selection rules that select a data-dependent number of parameters, controlling the RCC does not necessarily guarantee control of the FCR. Both Weinstein, Fithian, and Benjamini (2013) and Reid, Taylor, and Tibshirani (2017) considered selection rules that choose either a fixed number of the top ranked parameters or are based on a threshold. We find, in simulations, that for threshold-based selection rules, the RCC controlling intervals proposed in this article often control the FCR despite the lack of guarantee for these cases. These results and further discussion of the connection between FCR and RCC are included in Section 1 of the Appendix.

An advantage of using the RCC over the FCR is that RCC controlling confidence intervals can be divorced from the selection procedure. For example, if FCR controlling intervals are published for the top 10 parameter estimates but we are only able to follow up on the top 5 then we will need to recompute new, wider intervals to guarantee coverage within the smaller set. By contrast, the interpretation of RCC controlling intervals is independent of the selection rule or the numbers of parameters are selected.

1.4. Relationship of RCC to Empirical Bayes Approaches

The observation that motivates the RCC is that, in a study estimating many parameters, the full set of estimates can provide information about the true underlying parameter values. This is the same idea that motivates empirical Bayes (EB) approaches to simultaneous inference problems. In a Bayesian paradigm, we are interested in estimating the posterior distributions of $\theta_1, \ldots, \theta_p$, which, assuming conditional independence of the estimates and using Bayes rule, we can express as

$$p(\theta_j|\hat{\theta}_j) \propto p(\hat{\theta}_j|\theta_j)p(\theta_j).$$

The idea of EB approaches such as those of Efron (2008) and Stephens (2017) is to assume a theoretical distribution for $\hat{\theta}_j|\theta_j$ (e.g., $\hat{\theta}_j|\theta_j \sim N(\theta_j, 1)$) and use the large number of parameter estimates to estimate the prior $p(\theta_j)$. For example, in the adaptive shrinkage (`ash`) method proposed by Stephens (2017), $p(\theta_j)$ is assumed to be unimodal and centered at zero and, in one of several proposed variations, is estimated as a mixture of normal distributions. Provided the EB modeling assumptions hold, we can expect that, averaging over many realizations, the $1 - \alpha$ EB credible intervals contain the true parameter $1 - \alpha$ percent of the time and are immune to selection bias. That is, in a Bayesian system where a single realization of an experiment also includes resampling the parameter values, EB credible intervals should control the RCC.

The bootstrapping approaches we describe in Section 2 differ from EB methods in that they require fewer modeling assumptions and are derived from a frequentist perspective. We tend to view the parameters θ_j as fixed and use the large number of estimates to learn about the distribution of the bias $\hat{\theta}_{s(i)} - \theta_{s(i)}$. This method requires no assumptions about the form of $p(\theta_j)$. In a nonparametric bootstrapping variation, we are also able to avoid assumptions about the form of $p(\hat{\theta}_j|\theta_j)$, provided we have access to the individual level data used to produce the original estimates. The flexibility of the nonparametric method comes at the expense of increased computational effort. The parametric bootstrap can be fairly efficient and, in the example in Section 1.5, is eight times faster than `ash`. The nonparametric bootstrap can be quite costly since we must be willing to repeat

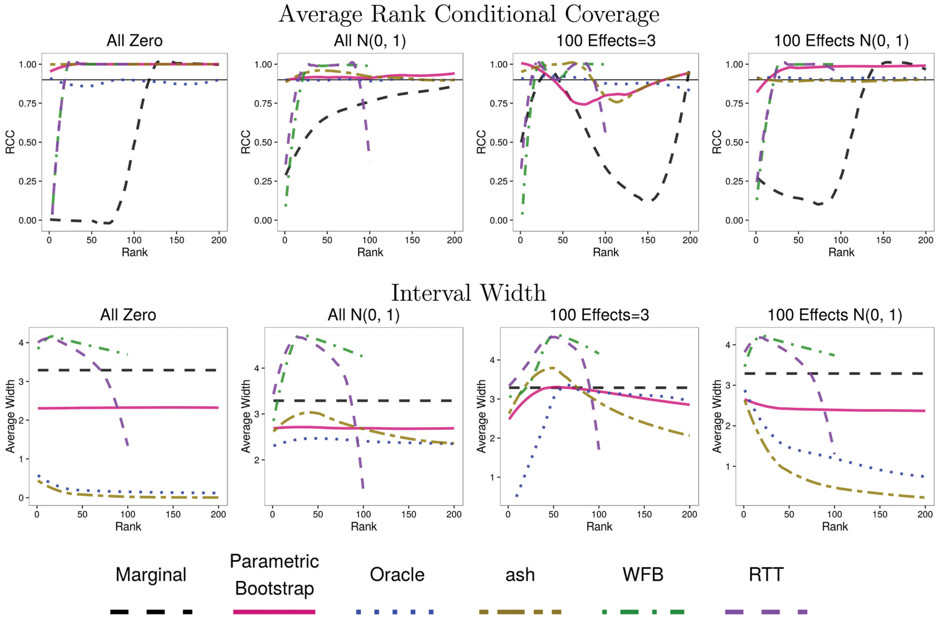


Figure 1. Rank conditional coverage estimated using 100 simulations (top) and average interval widths (bottom) for the example described in Section 1.5 The top 200 ranks are shown and both coverage and width are smoothed using loess. The four sets of true parameters are described in Section 1.5 The horizontal line in the top plots shows the nominal level 90%. `ash` refers to the EB method of Stephens (2017). WFB and RTT refer to the methods of Weinstein, Fithian, and Benjamini (2013) and Reid, Taylor, and Tibshirani (2017), respectively.

the entire analysis hundreds of times. However, in cases in which the parameter estimates are not independent or the theoretical distributions of test statistics are poor approximations, the nonparametric bootstrap is the most appropriate choice.

1.5. Example

Consider estimates $\hat{\theta}_1, \ldots, \hat{\theta}_{1000}$, where $\hat{\theta}_j$ is drawn from an $N(\theta_j, 1)$ distribution. Suppose the estimates are then ranked according to their absolute value and confidence intervals are constructed. Repeating this experiment multiple times, we can measure how often the parameter achieving rank i is covered, giving an estimate of the RCC at rank i. Figure 1 shows the RCC for the first 200 ranks and for several different configurations of parameter values:

1. All parameters are equal to zero.
2. All parameters are small and nonzero: θ_j is generated from an $N(0, 1)$ distribution but fixed for all simulations.
3. A few large nonzero parameters: $\theta_j = 3$ for $j = 1 \ldots 100$ and $\theta_j = 0$ for $j > 100$.
4. A few small nonzero parameters: θ_j drawn from an $N(0, 1)$ distribution but fixed over all simulations for $j = 1 \ldots 100$ and $\theta_j = 0$ for $j > 100$.

The standard marginal confidence intervals are $CI_i = \hat{\theta}_j \pm \Phi^{-1}(1 - \alpha/2)$. In configuration 1, this interval has an RCC of $\sim 0\%$ for the 65 most extreme observations but an RCC of $\sim 100\%$ for statistics closer to the median giving an overall average of 90% coverage.

We see a similar pattern in the intervals constructed using the methods of Weinstein, Fithian, and Benjamini (2013, WFB) and Reid, Taylor, and Tibshirani (2017, RTT). Both provide intervals only for a selected subset of parameters (we selected parameters associated with the top 100 estimates). Both methods control FCR but do so by under-covering parameters associated with the most significant statistics and over-covering parameters with more moderate statistics. In settings 2 and 3, the intervals of Reid, Taylor, and Tibshirani (2017) have poor RCC for both the most and least extreme parameters selected. The credible intervals generated by the `ash` method of Stephens (2017) control the RCC in settings 1, 2, and 4 and for most ranks in setting 3. These credible intervals also sometimes achieve a very small average interval width because `ash` attempts to shrink parameter estimates to zero. If the posterior probability that the parameter is equal to zero is larger than the desired level, the resulting credible interval will simply be $[0, 0]$.

Figure 1 also shows the results of the parametric bootstrapping method described in Section 2. This method provides an RCC larger than or equal to the nominal level for most ranks and most settings. As with `ash`, the setting in which the bootstrapping method performs the worst is setting 3. In most cases, the bootstrap intervals are much shorter than the marginal and FCR controlling intervals.

The fact that the bootstrapping method does not always achieve exactly the nominal level of RCC is a result estimating $E[\hat{\theta}_i]$. If these values were known, we could produce an "oracle" estimate which, with enough Monte Carlo samples, would achieve exactly the desired RCC for all ranks. The oracle is shown in Figure 1 and provides the motivation for bootstrapping methods proposed in this article. A detailed walk-through of the results in this section and code for replicating Figure 1 is available in *https://jean997.github.io/rccSims/compare_cis.html*.

The rest of this article is organized as follows: In Section 2, we introduce the oracle RCC controlling confidence intervals. These are then extended to parametric and nonparametric methods that use bootstrapping to estimate unknown parameters. In Section 3, we explore the performance of these methods in two simulation studies designed to mimic common high-dimensional analyses. We conclude with a discussion in Section 4.

2. Parametric and Nonparametric Bootstrapping to Build Confidence Intervals

2.1. Rank Conditional Confidence Intervals

First consider estimating a single parameter of a single distribution $\theta = T(F)$. Let $\delta = \hat{\theta} - \theta$ be the bias of the estimate and $H(x) = P[\delta \leq x]$ be the cdf of δ. If H is known, a pivotal exact $1 - \alpha$ confidence interval can be constructed as

$$\left(\hat{\theta} - H^{-1}(1 - \alpha/2),\ \hat{\theta} - H^{-1}(\alpha/2)\right). \tag{4}$$

In the high-dimensional setting, we are attempting to estimate p parameters $\theta_j = T_j(F)$. We can construct a rank conditional analog of the classical pivotal interval in (4). Define $\hat{\theta}_{s(i)}$ as in Section 1.2 where $s(i)$ gives the index of the ith ranked parameter estimate. We define the bias of the estimates at each rank

$$\delta_{[i]} = \hat{\theta}_{s(i)} - \theta_{s(i)}, \tag{5}$$

where the subscript $[i]$ indicates ranked-based indexing. Let $H_{[i]} = P[\delta_{[i]} \leq x]$ be the cdf of $\delta_{[i]}$. Where $H_{[i]}$ are known, an exact $1 - \alpha$ confidence interval for $\theta_{s(i)}$ could be constructed as

$$CI^{\text{exact}}_{s(i)} = \left(\hat{\theta}_{(i)} - H^{-1}_{[i]}(1 - \alpha/2),\ \hat{\theta}_{(i)} - H^{-1}_{[i]}(\alpha/2)\right). \tag{6}$$

We note that the rank conditional intervals are not pivotal because the distribution of $\delta_{[i]}$ depends on $\theta_1, \ldots, \theta_p$. This makes them more difficult to obtain when $H_{[i]}$ are unknown but does not impact the coverage probability of (6).

Lemma 1. The intervals in (6) have exact $1 - \alpha$ coverage:

$$P[\theta_{s(i)} \in CI^{\text{exact}}_{s(i)}] = 1 - \alpha.$$

Proof. This proof is identical to the proof for the classical interval in (4) given by Wasserman (2005) among others. Let $a = \hat{\theta}_{s(i)} - H^{-1}_{[i]}(1 - \alpha/2)$ and $b = \hat{\theta}_{s(i)} - H^{-1}_{[i]}(\alpha/2)$:

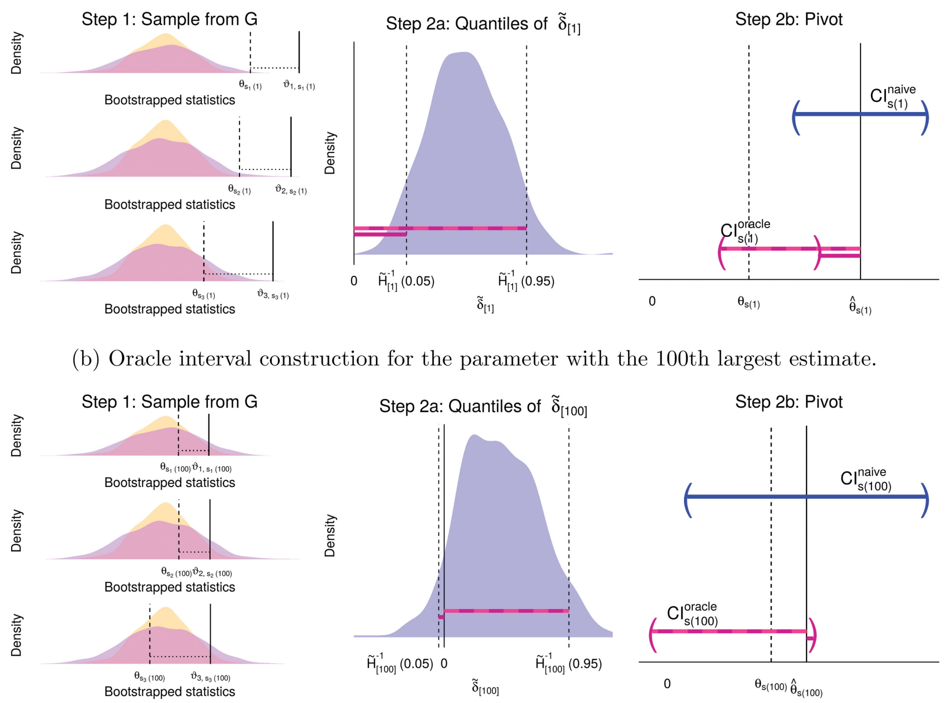
$$\begin{aligned}
P[a \leq \theta_{s(i)} \leq b] &= P[\hat{\theta}_{s(i)} - b \leq \delta_{[i]} \leq \hat{\theta}_{s(i)} - a] \\
&= H_{[i]}(\hat{\theta}_{s(i)} - a) - H_{[i]}(\hat{\theta}_{s(i)} - b) \\
&= H_{[i]}\left(H^{-1}_{[i]}(1 - \alpha/2)\right) - H_{[i]}\left(H^{-1}_{[i]}(\alpha/2)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

□

2.2. Generating Oracle Intervals with Monte Carlo Sampling

Construction of the intervals in (6) requires knowledge of the quantiles of the cdfs $H_{[i]}$, but working directly with $H_{[i]}$ may be difficult. If, instead, we can easily sample from the joint distribution G of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top$, then the quantiles of $H_{[i]}$ can be computed via Monte Carlo. We now describe this oracle Monte Carlo procedure which is detailed in algorithm 1 and illustrated in Figure 2.

First, we draw K independent p-vectors $\boldsymbol{\vartheta}_1 \ldots \boldsymbol{\vartheta}_K$ from G. Let s_k be the ranking permutation function for $\boldsymbol{\vartheta}_k$ and $\vartheta_{k,s_k(i)}$ be the ith ranked element of $\boldsymbol{\vartheta}_k$. Define the observed bias in sample k


(a) Oracle interval construction for the parameter with the largest estimate.

(b) Oracle interval construction for the parameter with the 100th largest estimate.

Figure 2. Generating oracle confidence intervals using algorithm 1 for rank $i = 1$ in (a) and $i = 100$ in (b). Left panels: A smoothed histogram of the true parameters is shown in the background. Smoothed histograms of three sets of Monte Carlo replicates from G are overlayed. Solid vertical lines mark the locations of the largest (a) and 100th largest (b) element of the Monte Carlo sample. Dashed vertical lines mark the corresponding true parameter value. The distance between these lines is the bias, $\delta_{k,[i]}$. Middle: A smoothed histogram of biases from 500 Monte Carlo samples with 0.05 and 0.95 quantiles marked by vertical dashed lines. Right: The oracle interval is constructed by pivoting the quantiles in the middle panel around the observed test statistic $\hat{\theta}_{s(i)}$, marked by a solid vertical line. The horizontal lines extending from $\hat{\theta}_{s(i)}$ are the same length as the corresponding lines in the middle panel. The dashed vertical line shows the location of the true parameter. The naive interval is shown for comparison. The vertical axis is meaningless.

at rank i as

$$\tilde{\delta}_{k,[i]} = \vartheta_{k,s_k(i)} - \theta_{s_k(i)}. \tag{7}$$

The left panels of Figure 2(a) and 2(b) show smoothed histograms of the true parameter values and the wider distribution of samples from G. The middle panels of these figures show smoothed histograms of $\tilde{\delta}_{k,[i]}$ for $i = 1$ and 100, respectively, using $K = 500$ Monte Carlo samples.

After computing the bias at each rank for each sample, we use the sample quantiles of $\{\tilde{\delta}_{1,[i]} \ldots \tilde{\delta}_{K,[i]}\}$ to estimate the quantiles of $H_{[i]}$, for $i = 1, \ldots, p$. We denote these sample quantiles as $\tilde{H}_{[i]}^{-1}(\cdot)$. The 0.05 and 0.95 quantiles of these distributions are shown as dashed vertical lines in the middle panels of Figure 2(a) and 2(b).

Substituting these estimates into the interval in (6) gives the oracle confidence interval

$$\mathrm{CI}_{s(i)}^{\text{oracle}} = \left(\hat{\theta}_{s(i)} - \tilde{H}_{[i]}^{-1}(1-\alpha/2),\ \hat{\theta}_{s(i)} - \tilde{H}_{[i]}^{-1}(\alpha/2)\right). \tag{8}$$

This procedure is illustrated in the right panels of Figure 2(a) and 2(b). For both of the ranks shown, the oracle Monte Carlo intervals are shorter than the marginal interval and contain the true parameter value.

These intervals are called oracle confidence intervals because they use knowledge of G and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Under these circumstances, $\tilde{H}_{[i]}^{-1}(x)$ will converge to $H_{[i]}^{-1}(x)$ as the number of Monte Carlo samples increases, so using (8), we can achieve the correct $1-\alpha$ confidence level (within any ϵ tolerance). This can be seen in Figure 1 where the oracle intervals have very close to the target 90% RCC at all ranks and are shorter than other methods. The following sections describe bootstrapping methods for estimating $H_{[i]}^{-1}$ when G and $\boldsymbol{\theta}$ are unknown.

Algorithm 1 Generating oracle intervals

```
$G$ and $\theta_1, \ldots, \theta_p$ are known.
1. For $k$ in $1 \ldots K$:
    a. Sample $\boldsymbol{\vartheta}_k = (\vartheta_{k,1} \ldots \vartheta_{k,p})$ from $G$.
    b. Calculate the bias at each rank $\tilde{\delta}_{k,[i]}$ as in (7).
2. For $i$ in $1 \ldots p$
    a. Calculate empirical quantiles $\tilde{H}_{[i]}^{-1}(x)$ of $\{\tilde{\delta}_{1,[i]} \ldots \tilde{\delta}_{K,[i]}\}$
    b. Generate $\mathrm{CI}_{s(i)}^{\text{oracle}}$ as in (6)
```

The intervals, $\mathrm{CI}_{s(i)}^{\text{oracle}}$, in (8) do not necessarily contain $\hat{\theta}_{s(i)}$. This is particularly true if the point-estimates, $\hat{\theta}_{s(i)}$, have not been adjusted for multiplicity/selection bias (e.g., if each $\hat{\theta}_{s(i)}$ is a maximum likelihood estimate). In Figure 2(a), both the 0.05 and 0.95 quantiles of the observed bias are positive so the confidence interval lies completely below $\hat{\theta}_{s(1)}$. A more natural point estimate to pair with this confidence interval is the debiased estimate proposed by Simon and Simon (2013), $\hat{\theta}_{s(i)} - \frac{1}{K}\sum_{k=1}^{K} \tilde{\delta}_{k,[i]}$, which will generally lie within the confidence interval.

The following bootstrap methods for confidence interval construction are extensions of point estimation methods proposed by Simon and Simon (2013) and Tan, Simon, and Witten (2014). These proposals estimate the mean of $\delta_{[i]}$ which can be used to debias point estimates of $\theta_{s(i)}$. We consider two bootstrapping strategies: a parametric bootstrap, useful when the distribution and covariance of the parameter estimates are known or can be approximated well, and a nonparametric bootstrap which is more widely applicable but is also more computationally costly. Both of these are general strategies, where the specifics of the algorithm may vary depending on the specific application.

A related bootstrapping procedure has been proposed by Claggett, Xie, and Tian (2014) for different purposes. These authors are interested in constructing confidence intervals for the quantiles, or ranked elements of $\boldsymbol{\theta}$. By contrast, we are interested in estimating confidence intervals for the parameters corresponding to the ranked elements of $\hat{\boldsymbol{\theta}}$.

2.3. Parametric Bootstrap

The parametric bootstrap parallels the Monte Carlo algorithm in Algorithm 1, replacing G and $\boldsymbol{\theta}$ with estimates based on the data. We assume that G is a member of a parametric family of distributions and estimate its parameters. In principal, we could use any family of distributions, but in this discussion, we will assume that $\hat{\theta}_j \sim N(\theta_j, \sigma_j^2)$ where σ_j^2 is either known or can be estimated and $\hat{\theta}_j$ are independent given $\boldsymbol{\theta}$. This type of parametric bootstrap is best suited for scenarios in which the estimator has an asymptotically normal distribution with known variance such as linear regression. We estimate $\hat{G}$ by replacing θ_j with an estimate such as $\hat{\theta}_j$ itself or debiased estimates of Simon and Simon (2013) or Tan, Simon, and Witten (2014). The latter choices will involve two stages of bootstrapping, one to generate a debiased mean estimate and the second to generate confidence intervals (see Section 2 of the Appendix).

The quantiles estimated through Monte Carlo simulation in (6) are replaced by bootstrapped quantiles $\hat{H}_{[i]}^{-1}(x)$ obtained by sampling p-vectors from $\hat{G}$ rather than from G. This gives the bootstrap intervals

$$\mathrm{CI}_{s(i)}^{\text{boot}} = \left(\hat{\theta}_{s(i)} - \hat{H}_{[i]}^{-1}(1-\alpha/2),\ \hat{\theta}_{s(i)} - \hat{H}_{[i]}^{-1}(\alpha/2)\right). \tag{9}$$

This procedure is described in Algorithm 2 for the case when $G = N(\boldsymbol{\theta}, \mathbf{I})$ where $\mathbf{I}$ is the $p \times p$ identity matrix. Many variations on this procedure are possible. For example, it may be easier to specify a distribution for a transformation of θ_j.

This method provides an RCC closer to the nominal level when $\hat{G}$ is closer to G. This means that, in the case of asymptotically normal test statistics, we achieve better performance using better estimates of $\boldsymbol{\theta}$, such as those proposed by Simon and Simon (2013). Algorithm 1 in Section 2 of the Appendix shows the additional steps necessary when using this debiased mean estimate. If the ranking scheme is based on the absolute value of the parameter estimates (e.g., using the magnitude of a t-statistic), it is necessary to reflect the interval across zero for negative parameter estimates. Algorithm 2 in Section 2 of the Appendix gives the parametric bootstrap procedure for absolute value-based rankings. Algorithm 2 and the two variations described in the Appendix are implemented in the `par_bs_ci` function of the R package `rcc`.

2.4. Nonparametric Bootstrap

The parametric bootstrap can be applied when G is well approximated by a member of a parametric family. It is particularly convenient for statistics which are asymptotically normal and either independent or have a covariance that can be estimated well. Many high-dimensional problems possess complex dependence structures which are not easy to estimate. Furthermore, not all estimators have known asymptotic distributions. In these cases, the parametric bootstrap, the selection adjusted FCR controlling methods, and EB methods that assume conditional independence between estimates are all unsuitable. It is not possible to estimate a general G without making any structural assumptions about the true parameters. However, it is possible to generate bootstrap samples nonparametrically if individual data are available.

```
Algorithm 2 Simple parametric bootstrap for asymptotically normal estimates
1. For $k$ in $1 \dots K$:
   a. Sample $\vartheta_{k,i}$ from an $N(\hat{\theta}_i, 1)$ distributionfor $i$ in $1 \dots p$.
   b. Calculate the bias at each rank $\hat{\delta}_{k,[i]}$ as
         $\hat{\delta}_{k,[i]} = \vartheta_{k,s_k(i)} - \hat{\theta}_{s_k(i)}$
2. For $i$ in $1 \dots p$:
   a. Calculate empirical quantiles $\hat{H}^{-1}_{[i]}(x)$ of $\{\hat{\delta}_{1,[i]} \dots \hat{\delta}_{K,[i]}\}$
   b. Generate $\mathrm{CI}^{\mathrm{boot}}_{s(i)}$ as in (9)
```

The nonparametric bootstrap is based on sampling from the data used to compute $\hat{\boldsymbol{\theta}}$ and computing new estimates using the resampled data. This is implicitly sampling from a distribution $\hat{G}$ without requiring an analytical form. We assume the data consist of n independent data vectors $\mathbf{y}_1, \dots, \mathbf{y}_n$. These may be vectors of genotypes, biometric, or image data for n individuals. They may be a mix of data types and include covariates. We assume only that there is a procedure which takes $\mathbf{y}_1, \dots, \mathbf{y}_n$ as inputs and generates estimates $\hat{\boldsymbol{\theta}}$ and statistics indicating the significance of each estimate. A bootstrap p-vector can be generated by sampling n data vectors from $\mathbf{y}_1, \dots, \mathbf{y}_n$ with replacement and applying the original estimation procedure. From this point confidence intervals may be constructed identically to the parametric case.

More formally, if $\mathbf{y}_1, \dots, \mathbf{y}_n$ are iid draws from Π, and $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(\mathbf{y}_1, \dots, \mathbf{y}_n)$ is a function of those observations, then $G \equiv G(\Pi)$ is directly a function of Π. To estimate G, we can use the estimate induced by Π_n, the empirical distribution of the $\mathbf{y}_i$: $\hat{G}_{\mathrm{emp}} \equiv G(\Pi_n)$. From here we can estimate the quantiles in (6) by $\hat{H}^{-1}_{\mathrm{emp}[i]}(x)$ obtained by sampling repeatedly from $\hat{G}_{\mathrm{emp}}$. This leads us to the nonparametric bootstrap intervals:

$$\mathrm{CI}^{\text{np-boot}}_{s(i)} = \left(\hat{\theta}_{s(i)} - \hat{H}^{-1}_{\mathrm{emp}[i]}(1 - \alpha/2),\ \hat{\theta}_{s(i)} - \hat{H}^{-1}_{\mathrm{emp}[i]}(\alpha/2)\right). \tag{10}$$

The specifics of this procedure are shown in Algorithm 3 and implemented in the `nonpar_bs_ci` function of the R package `rcc`. Nonparametric bootstrapping can potentially be very time consuming. If the original analysis was computationally expensive it may be infeasible to repeat it many times to obtain confidence intervals.

```
Algorithm 3 Nonparametric bootstrap
1. For $k$ in $1 \dots K$:
   a. Sample $\mathbf{y}_{k,1}, \dots \mathbf{y}_{k,n}$ with replacement from $\{\mathbf{y}_1, \dots \mathbf{y}_n\}$
   b. Using the sampled data, calculate estimates $\boldsymbol{\vartheta}_k = (\vartheta_{k,1} \dots \vartheta_{k,p})$.
   c. Estimate the bias at each rank $\hat{\delta}_{k,[i]}$ as
         $\hat{\delta}_{k,[i]} = \vartheta_{k,s_k(i)} - \hat{\theta}_{s_k(i)}$
2. For $i$ in $1 \dots p$:
   a. Calculate empirical quantiles $\hat{H}^{-1}_{\mathrm{emp}[i]}(x)$ of $\{\hat{\delta}_{1,[i]} \dots \delta_{K,[i]}\}$
   b. Generate $\mathrm{CI}^{\text{np-boot}}_{s(i)}$ as in (10)
```

3. Simulations

3.1. Linear Regression with Correlated Features

In this set of simulations, we explore how correlation among parameter estimates effects the rank conditional coverage rates of different methods of confidence interval construction. Code replicating these results can be found at *https://jean997.github.io/rccSims/linreg_sims.html*.

We consider a common analysis procedure used in genetic and genomic studies. In these studies, researchers measure far more features (such as gene expression levels) than there are samples and focus on estimating the marginal association between each feature individually and an outcome. We consider a setting in which the features occur in correlated blocks leading to correlated parameter estimates.

In each simulation, we simulate 1000 normally distributed features for 100 samples. Let $x_{i,j}$ denote the value of the jth feature for the ith individual and $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,1000})^T$. The features are simulated as

$$\mathbf{x}_i \sim N_{1000}(0, \Sigma),$$

where the covariance matrix, Σ, is block diagonal with 100 10 × 10 blocks. The diagonal elements of each block are equal to 1 and the off diagonal elements are equal to ρ. The outcome for individual i is simulated as

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i \qquad \epsilon_i \sim N(0, 1),$$

where elements of $\boldsymbol{\beta}$, the vector of conditional effect sizes, are equal to 0 at all but 100 elements. In each block the effect size for the fifth feature is drawn from an $N(0, 1)$ distribution while the effects for the other features are 0. These effects are fixed over all simulations.

In this analysis, we estimate the marginal rather than conditional effect sizes, $\boldsymbol{\beta}^{(\mathrm{marg})} = \Sigma \boldsymbol{\beta}$. We estimate $\beta^{(\mathrm{marg})}_j$ through a univariate linear regression of $\mathbf{x}_{\cdot j} = (x_{1,j}, \dots, x_{n,j})^T$ on $\mathbf{y}$. This is a standard analysis strategy for many genomic studies such as genome-wide association studies and gene expression studies.

We consider four levels of correlation between the features by setting ρ equal to $0, 0.3, 0.8$, and -0.1. Rank conditional coverage and interval widths averaged over 400 simulations for each scenario are shown in Figure 3. In these results, parameter ranking is based on the absolute value of the t-statistic $\hat{\beta}^{(\mathrm{marg})}_j / \hat{\mathrm{se}}(\hat{\beta}^{(\mathrm{marg})}_j)$. In Appendix Section 3, we consider ranking

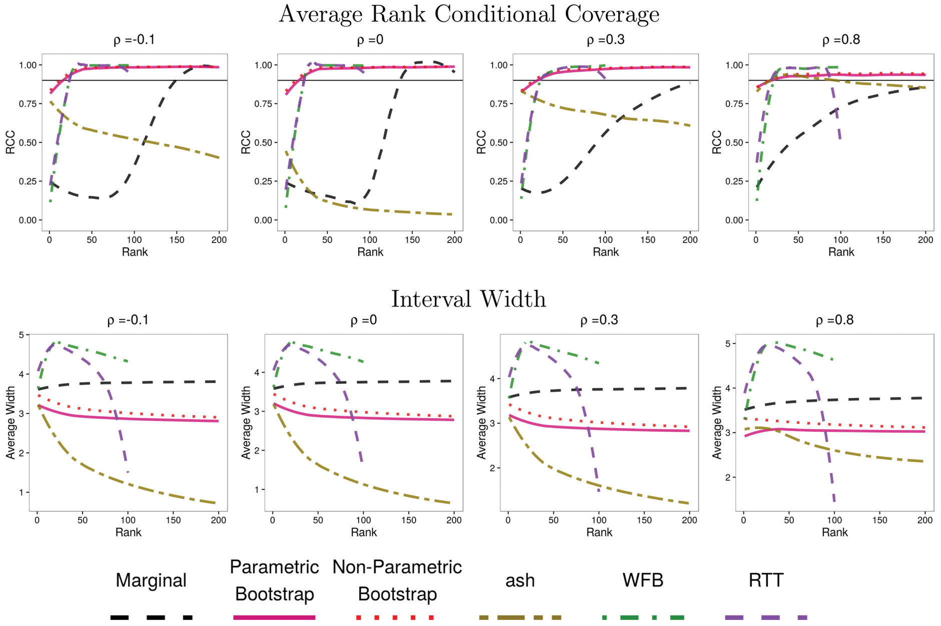


Figure 3. Simulation results for Section 3.1 Rank conditional coverage (top) and interval widths (bottom) are shown for the top 20% of parameters averaged over 400 simulations. Parameters are ranked by absolute value of the test statistic. Coverage rates and widths are smoothed using loess. In the top panel, a horizontal line shows the nominal level 90%. `ash` refers to the EB method of Stephens (2017). WFB and RTT refer to the methods of Weinstein, Fithian, and Benjamini (2013) and Reid, Taylor, and Tibshirani (2017), respectively.

the parameters by first selecting the parameter with the most significant estimate in each block and then ranking only these 100 selected parameters based on the absolute value of the t-statistic.

We find that both the parametric and nonparametric confidence intervals perform well in all four settings and are quite similar, even though the parametric bootstrap assumes independence between the estimates. None of the other methods provides an RCC close to the nominal level except for `ash` in the highest correlation scenario. The `ash` method does poorly in the other scenarios (while it did quite well in the example in Section 1.5) because the marginal effects are not sparse and `ash` attempts to shrink parameters to zero. We found similar results repeating these simulations with block sizes of 2, 20, 100, and using a mixture of block sizes (see Appendix Section 4).

3.2. Treatment Effects in Nested Subgroups

In Section 3.1, we found that the parametric bootstrap performed well even when the assumption of independence between estimates was violated. Here, we provide an example of how the parametric bootstrap can fail when estimates are very highly correlated. Code replicating these results can be found at *https://jean997.github.io/rccSims/biomarker_sims.html*.

This example is motivated by the use of biomarkers in clinical trials. Suppose we have conducted a clinical trial in which participants are randomized into two groups. For participant i, we record the treatment group, $\text{trt}_i \in \{0, 1\}$, an outcome y_i and the value of a biomarker, w_i. We expect that the treatment will have a greater effect in individuals with higher values of the biomarker but do not know the exact relationship between the biomarker and the treatment effect. In an exploratory analysis, we define a series of cut-points $c_1, \ldots, c_p$. For each cut-point, we estimate the difference in treatment effects for participants with biomarker measurements above and below the cut-point:

$$\beta_j = \left(E[y_i|\text{trt}_i = 1, w_i > c_j] - E[y_i|\text{trt}_i = 0, w_i > c_j]\right) \\ - \left(E[y_i|\text{trt}_i = 1, w_i \le c_j] - E[y_i|\text{trt}_i = 0, w_i \le c_j]\right).$$

We estimate β_j as the OLS estimate fitting the regression

$$y_i = \beta_0 + \beta_1 \text{trt}_i + \beta_{2,j} 1_{w_i > c_j} + \beta_j \text{trt}_i * 1_{w_i > c_j} + \epsilon_i,$$

where $1_{w_i > c_j}$ is an indicator that $w_j > c_j$. We then rank these estimates by the absolute value of their t-statistics to select a cut-point that gives the most significant difference in treatment effect between groups. This cut-point might be used to design future clinical trials.

In each simulation, we generate data for 200 study participants, 100 randomized to the treatment arm, and 100 randomized to the control arm. We simulate the value of the biomarker as uniformly distributed between 0 and 1. The true relationship between the biomarker, the treatment, and the outcome given by

$$E[y_i|\text{trt}_i, w_i] = \begin{cases} 0 & w_i < 0.5 \\ \left(w_i - \frac{1}{2}\right) \cdot \text{trt}_i & w_i \ge 0.5 \end{cases}.$$

The observed outcome for individual i ($i \in 1, \ldots, 200$) is $y_i = E[y_i|w_i, \text{trt}_i] + \epsilon_i$ where $\epsilon_i \sim N(0, 0.25)$.

We chose 100 cut-points evenly spaced between 0.1 and 0.9. Rank conditional coverage and interval width averaged over 400 simulations are shown in Figure 4. In this scenario, parameter estimates are very highly correlated. This results in very poor performance for the parametric bootstrap which assumes independence between estimates. Interestingly, the standard marginal intervals do well despite making the same assumption. The nonparametric bootstrap also controls the RCC though it has slight under-coverage for the least significant parameters. Unlike the marginal intervals, the nonparametric bootstrap controls the RCC by modeling the correlation structure between

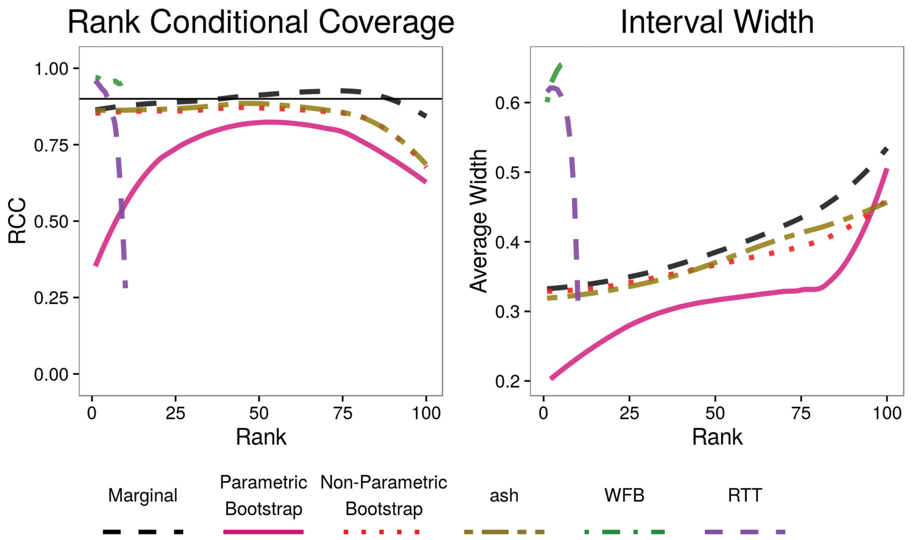


Figure 4. Simulation results for Section 3.2 Rank conditional coverage (left) and interval widths (right) are shown for the top 20% of parameters averaged over 400 simulations. Parameters are ranked by the absolute value of the test statistic. Coverage rates and widths are smoothed using loess. In the left-hand panel, a horizontal line shows the nominal level 90%. `ash` refers to the EB method of Stephens (2017). WFB and RTT refer to the methods of Weinstein, Fithian, and Benjamini (2013) and Reid, Taylor, and Tibshirani (2017), respectively.

parameter estimates and also performs well in the simulations in Section 3.1 making it a more reliable choice.

4. Discussion

Interval estimation when the number of parameters is large is a challenging problem often ignored in large-scale studies. Out of caution, these studies are frequently limited to hypothesis testing but this limitation is unnecessary in many cases. We have shown that the full set of parameter estimates contains information and can be used to correct bias and generate useful confidence intervals. We have also introduced a more granular, informative concept of coverage which can be applied to confidence intervals constructed for numerous parameters.

Rank conditional coverage is an important criterion to consider in evaluating confidence intervals for large parameter sets. As a finer grained criterion, it reveals problems that are ignored by the FCR. In many cases, using an FCR controlling procedure after selecting top parameter-estimates results in very low coverage probabilities for the very largest parameters.

In our simulations, we found that rank conditional coverage is a more difficult criterion to control than the false coverage statement rate of Benjamini and Yekutieli (2005). The two proposed bootstrapping methods almost always outperformed other methods and produced smaller intervals than all methods except the `ash` method of Stephens (2017).

Supplementary Materials

Appendix: Contains a discussion of the connection between the FCR and RCC, two variations of Algorithm 2, and additional simulation results referenced in the text.

R-package `rcc`: R-package implementing Algorithms 1, 2, 3, and 3. (GNU zipped tar file)

R-package `rccSims`: R-package replicating the simulations shown in Sections 1.5 and 3. (GNU zipped tar file)

Funding

J.M. and N.S. were supported by NIH Grant DP5OD019820.

References

Benjamini, Y., and Yekutieli, D. (2005), "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters," *Journal of the American Statistical Association*, 100, 71–81. [648,649,656]

Claggett, B., Xie, M., and Tian, L. (2014), "Meta-Analysis With Fixed , Unknown , Study-Specific Parameters Parameters," *Journal of the American Statistical Association*, 109, 1660–1671. [653]

Efron, B. (2008), "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23, 1–22. [650]

——— (2011), "Tweedies Formula and Selection Bias," *Journal of the American Statistical Association*, 106, 1602–1614. [648]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [xxxx]

Reid, S., Taylor, J., and Tibshirani, R. (2017), "Post-Selection Point and Interval Estimation of Signal Sizes in Gaussian Samples," *The Canadian Journal of Statistics*, 45, 128–148. [649,650,651]

Simon, N., and Simon, R. (2013), "On Estimating Many Means, Selection Bias, and the Bootstrap," arXiv preprint arXiv:1311.3709. [648,653]

Stephens, M. (2017), "False Discovery Rates: A New Deal," *Biostatistics*, 18, 275–295. [650,651,656]

Sun, L., and Bull, S. B. (2005), "Reduction of Selection Bias in Genomewide Studies by Resampling," *Genetic Epidemiology*, 28, 352–367. [648]

Tan, K., Simon, N., and Witten, D. (2014), "Selection Bias Correction and Effect Size Estimation under Dependence," arXiv preprint arXiv:1405.4251. [653]

Wasserman, L. (2005), *All of Nonparametric Statistics*, New York, NY: Springer. [652]

Weinstein, A., Fithian, W., and Benjamini, Y. (2013), "Selection Adjusted Confidence Intervals With More Power to Determine the Sign," *Journal of the American Statistical Association*, 108, 165–176. [648,649,650,651]

Zhong, H., and Prentice, R. L. (2008), "Bias-Reduced Estimators and Confidence Intervals for Odds Ratios in Genome-Wide Association Studies," *Biostatistics*, 9, 621–634. [648,649]