An Analysis of Standard Error Estimation for Clustered Data

STA640 Final Project

Andrew Amore & Rob Kravec

4/24/2022

Abstract

Clustered data is common in real-world settings and of particular interest to the field of causal inference when working with cluster randomized trials. When performing inference tasks on clustered data, it is critical to account for the inherent grouping structure in standard error estimation, as methods that ignore clustering exhibit severe downward bias. In this project, we first discuss techniques commonly used in the literature to estimate standard errors in clustered data. Then, we perform Monte Carlo simulations with varying numbers of clusters, observations per cluster, and intracluster correlation (ICC) to evaluate the performance of three different bootstrap-based standard error estimation techniques: (1) a standard bootstrap where clustering is ignored, (2) a block bootstrap method where clusters are resampled, and (3) a "double" bootstrap that first resamples by cluster and then performs a standard bootstrap within each resampled cluster. Results from these simulations align with findings in the literature that the standard bootstrap exhibits significant downward bias in the presence of non-zero ICC, and additional trends are also explored. Lastly, these three bootstrap-based standard error estimation techniques are applied to a real-world data set that contains home sale prices in Ames, IA, clustered by neighborhood. While the true population parameter is not known in this case, it is easily observable that the standard bootstrap produces the tightest confidence interval and likely underestimates the degree of uncertainty in the inference task.

Please see the linked Github repository for all code used to generate this paper.

Introduction & Motivation

To move beyond statements of association using the potential outcome framework, the field of causal inference depends on a number of critical assumptions. One of these assumptions, the Stable Unit Treatment Values Assumption (SUTVA), states there is no interference between units (i.e., a unit's outcome is independent of treatment status assigned to other units). In many situations, individual level randomization into treatment and control groups, can satisfy SUTVA, but sometimes it may be logistically difficult or practically impossible for this assumption to hold, which can bias statistical results. For example, in medical studies many patients might see identical doctors, live in the same geographic area, or interact with one another which can make causal inference difficult.

To account for latent or observed group membership, treatment randomization can occur at a cluster level through a CRT (cluster randomized trial). In the medical study example, all individuals from the same hospital would be assigned the same treatment status, which can help account for some interactions. However, even in a CRT, SUTVA can become untenable as "individuals in the same cluster are very likely to resemble one another, not only in terms of pre-treatment characteristics, but also in terms of treatment receipt behavior and post-treatment outcomes [1]." As a result, downward bias can creep into standard error estimation if intracluster correlation is neglected, contributing to tighter confidence intervals and higher type I error than the specified testing level.

Clustering-related "error" is more than just omitted variable bias (OVB) as it can pose problems when OVB is non-existent [2]. For example, Harden used simulations to show situations where standard error estimates using certain methods are biased downwards and supported his findings with real data from political science case studies where changing the estimation method altered the statistical significance of the findings [3].

In the rest of this paper, we will assess several methods from literature via Monte Carlo simulation, apply these methods to a real data set, and (attempt to) develop a new standard error estimation technique to

better quantify uncertainty.

Review of Existing Methods

Several existing methods attempt to account for clustering during standard error estimation and adjust for any cluster induced bias. Two variations are prevalent in literature: "sandwich" estimation and bootstrap. We will introduce OLS-SE, share its relationship with sandwich formulation and move to bootstrap methods last which will be the main focus of this paper.

OLS-SE In a typical OLS linear model the error term(s) affecting all observations are assumed to be independent and drawn from the same distribution (iid). Under this model all observations "experience" the same source and strength of error (at least distributionally).

A closed form solution for the coefficients exists and is derived from the observations. However, standard error estimates on the coefficients tend to be biased downward for clustered data.

For this reason more robust estimator estimation techniques are needed. One way to deal with deal with heterogeneous effects are with "sandwich" estimators.

Robust Cluster Standard Error (RCSE) In typical OLS regression error terms are assumed to have constant variance. This assumption is relaxed in heterogeneous estimation methods (e.g., Huber-White) where variance terms are no longer constrained to be identical.

Webb et. al. indicates this estimation technique controls for error heteroskedasticty and general correlation within clusters [4].

Standard errors are estimated for each cluster using the following formulation where N_c denotes the number of observations per cluster and r_j denotes each observations residual.

$$Var(\hat{\beta}) = (X^T X)^{-1} \sum_{j=1}^{N_c} \left(X_j \hat{r}_j \hat{r}_j^T X_j^T \right) (X^T X)^{-1}$$

Consistency is based on the following assumptions: the number of clusters go to infinity, the degree of within-cluster correlation remains constant across clusters, and there are an equal number of observations in each cluster. All three assumptions are tenuous under real life conditions.

Bootstrap Methods

At a high-level bootstrap, is an iterative resampling technique which assumes a data set is a representative sample from some population. The parameter of interest is repeatedly calculated from multiple sub-sample(s), and standard error estimates are approximated from the standard deviation of the resampled parameter estimates. Bootstrapping is popular because it can be leveraged non-parametrically to estimate a parameter with minimal model constraints. Several bootstrapping methods are introduced and evaluated using simulation.

Simple "Regular" Bootstrap Under regular bootstrap, the clustering structure is completely ignored and observations are resampled independent of clustering assignment. Before beginning a bootstrap sample size, N, and iteration number, B, are specified. The steps are as follows:

1. Draw sample with replacement of length N from full data set. Typically, N is equal to the original length of the full data set
2. Compute parameter estimate from bootstrapped sample
3. Repeat B times
4. Estimate standard error from collection of bootstrap parameter estimates in steps 1-3

Cluster Bootstrap Instead of ignoring the clustering structure like in the "regular" bootstrap, the collection of clusters themselves are iteratively resampled. When a cluster is selected in the bootstrap sample all observations are utilized for the parameter estimate.

Double Bootstrap Instead of including all cluster observations when a given cluster is selected, observations from the selected cluster are bootstrap sampled again and used to estimate the parameter estimate. Thus, a bootstrap procedure is performed twice, once at each level of the data hierarchy.

Simulations

In order to assess standard error estimation methods one needs a parameter to estimate. For simplicity, coverage and interval widths are calculated for population means, but these results can easily be extended to standard error estimation of other quantities of interest, such as causal estimands.

Objective

Harden's 2011 paper titled "A Bootstrap Method for Conducting Statistical Inference with Clustered Data" informed the design of our simulations [3]. In this paper, Harden examines clustered data with various attributes and assesses three standard error estimation techniques: (1) robust cluster standard error estimation (RCSE), (2) ordinary least squares regression (OLS-SE), and (3) a cluster bootstrapping method (BCSE). For clustered data, Harden finds a downward bias in both RCSE and OLS-SE methods proportional to intracluster correlation, ρ. In contrast, BCSE estimation exhibits coverage close to theoretical guarantees for all situations Harden examined.

For our procedure, we aimed to corroborate Harden's Monte Carlo simulations under identical conditions: `total sample size = {200, 800, 1200}`, `number of clusters = {10, 25, 40, 50, 100}`, `ICC = {0.1, 0.5}` and expand the scope. While Harden compared BCSE to two non-bootstrap estimation methods, we compared BCSE to two alternative bootstrap methods described above: (1) a "simple" bootstrap ignoring cluster assignments and (2) a "double" bootstrap resampling at both cluster and observation level.

After verifying the correspondence between the BCSE performance in our simulations and Harden's paper, we extended simulations to (1) include more values of ICC and (2) experiment with more combinations of `number of clusters` and `observations per cluster` to further explore BCSE performance across a larger parameter space.

Data Generation Model

Simulated data was based on a random intercept model with no predictors: $y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ for all observations i in cluster j where $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and $\alpha_j \overset{\text{iid}}{\sim} N(0, \tau^2)$. We can define intracluster correlation (i.e., the correlation between two observations in the same group) as $\text{Corr}(y_{ij}, y_{i'j}) = \frac{\text{Cov}(y_{ij}, y_{i'j})}{\sqrt{\text{Var}(y_{ij})\text{Var}(y_{i'j})}} = \frac{\tau^2}{\tau^2+\sigma^2}$. For simplicity, we assume that the intracluster correlation is the same for all clusters (i.e, τ the same across clusters). Additionally, we assume a uniform cluster-level mean of zero across all clusters.

Figure 1 shows a toy example of our data generation process with three different ICC values. Clearly, as ICC increases, clusters become more concentrated and separated, even though all observations are drawn from a distribution with the same mean.

Simulation Procedure

500 data sets were drawn for each unique parameter combination in (`number of clusters`, `observations per cluster`, and `ICC`) using the generation process described above. For each data set, three bootstrap samples of size 100 were drawn: one for the simple bootstrap, one for BCSE, and one for the double bootstrap.

Standard errors were computed from the standard deviation of the bootstrap parameter estimates for each method. 95% confidence intervals were generated using a normal approximation (i.e., interval = sample mean $\pm 1.96 \times$ SE). Coverage for each standard error estimation technique was calculated as the percentage of the 500 bootstrap intervals containing the population mean (i.e., 0). For an unbiased standard error estimator, we expect ~95% coverage.

Simulation Results

In Figures 2 and 3, we show results for (1) coverage and (2) 95% confidence interval widths under the simulation conditions proposed in Harden's paper. Notably, we observe that BCSE gets close to theoretical coverage guarantees in the low ICC scenarios (`ICC = 0.1`), but BSCE does exhibit a slight downward bias

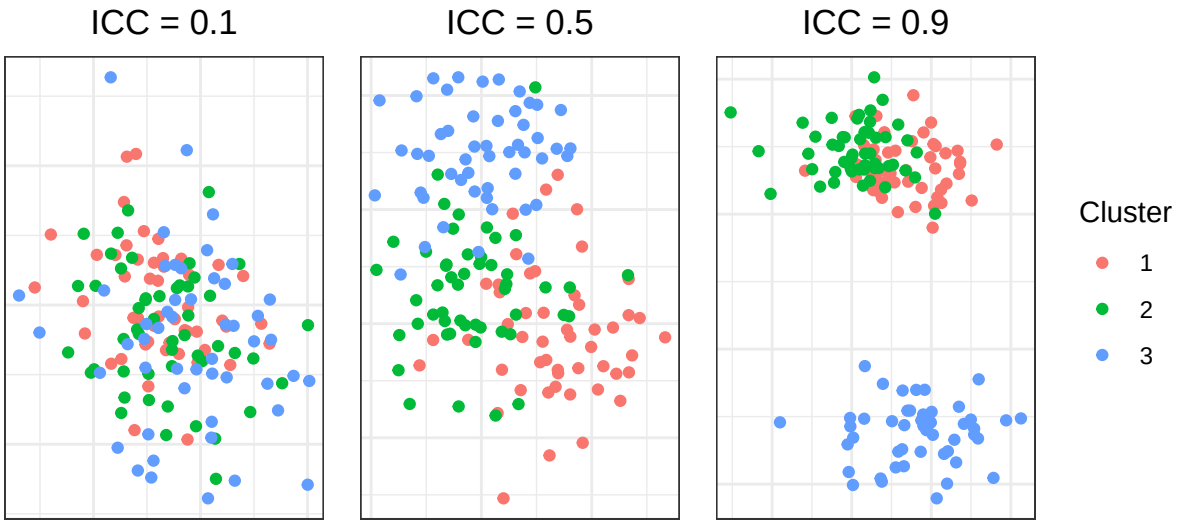

Figure 1: Data Generation Example

in the high ICC scenarios (`ICC = 0.5`). This downward bias appears to be worse for high sample size (`N = 1200`). The simple (individual) bootstrap has a much stronger downward bias, which is accentuated in the high ICC and sample size scenarios.

It is important to understand that the superior coverage for BCSE, relative to the simple bootstrap, comes at the cost of a much wider 95% confidence interval. For instance, when `N = 1200` and `ICC = 0.5`, the average 95% confidence interval width is about 1.1 for BSCE vs. 0.15 for the standard bootstrap.

Lastly, we note that the double bootstrap appears to have slightly higher coverage (and slightly wider intervals) in all tested situations, compared to the BCSE. For low ICC scenarios, the double bootstrap appears to be overly conservative, besting the theoretical guarantee. This conservatism may be warranted in the high ICC scenarios, but these results alone cannot support such a conclusion.

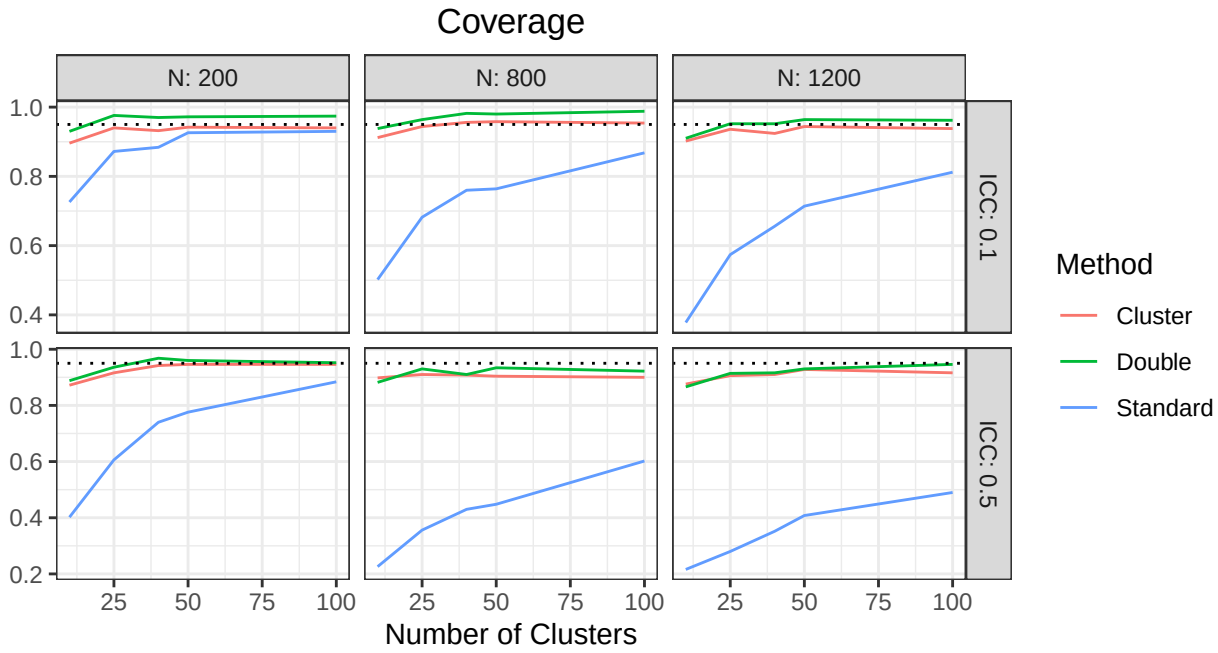

Figure 2: Harden et. al. Coverage

Given the clear poor performance of the simple bootstrap on clustered data with even moderate ICC, we focus on BCSE and double bootstrapping in the simulations to follow (for ease of visualization).

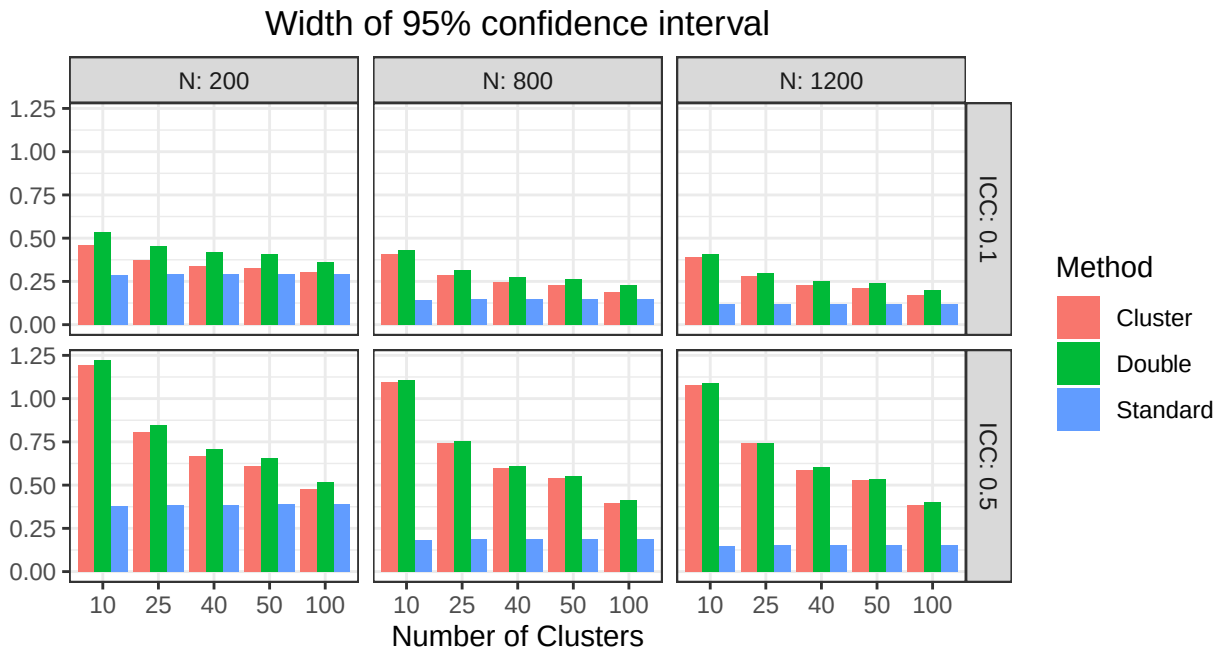

Figure 3: Confidence Interval Widths

Our more extensive simulation results reveal a number of additional interesting trends and conclusions (see Figures 4 and 5):

- For a given ICC and cluster size, coverage increases and confidence interval width decreases with an increasing number of clusters. For a smaller number of clusters, both cluster-robust bootstrapping techniques are more likely to be biased downward, which is consistent with our findings in the literature
- For a given ICC and number of clusters, coverage increases and confidence interval width decreases with an increasing cluster size. Both this trend and the one above make intuitive sense, as sample size and standard error have an inverse relationship (all else held equal)
- For a given number of clusters and cluster size, coverage decreases and confidence interval width decreases with an increasing ICC. This effect is observed much more strongly for the simple bootstrap, which does not account for any intracluster relationship
- When `ICC = 0`, the double bootstrap is overly conservative, yielding a coverage of almost 100% (with a wider than necessary confidence interval)

Case study: Ames, IA home sales

Next, we apply the three standard error estimation techniques from our simulation to a real-world data set, which contains sale prices for homes in Ames, IA from 2006 to 2010 [5].

Admittedly, our use of the data set is a bit contrived. We imagine a situation in which only the sale price and neighborhood of each house is available, creating a situation with obvious clustering. Significant unmeasured confounding exists when we focus on only these two attributes, and to mitigate this effect slightly, we filter the data set to only contain 3-bedroom homes (the most frequent number of bedrooms contained in the data). Then, ignoring the effect of time on housing prices during this five year period, our goal is to estimate the average value of a 3-bedroom home in Ames, IA from 2006 to 2010 using data from homes that sold during that interval. One could also imagine using this estimate to extrapolate to other cities with similar characteristics to Ames, IA.

One attribute that made this data set especially appealing was the large hypothesized effect of `Neighborhood` on `Sale Price`. With a calculated ICC of 0.6, we expected the standard error estimates for the simple bootstrap vs. cluster-robust bootstraps to be quite different.

Below, we display (normal approximations of) the probability distribution functions produced by the three bootstrap methods considered in this project. Relative to the cluster and double bootstraps (which produce similar PDFs), the standard bootstrap PDF is much taller and narrower. While we do not have access to

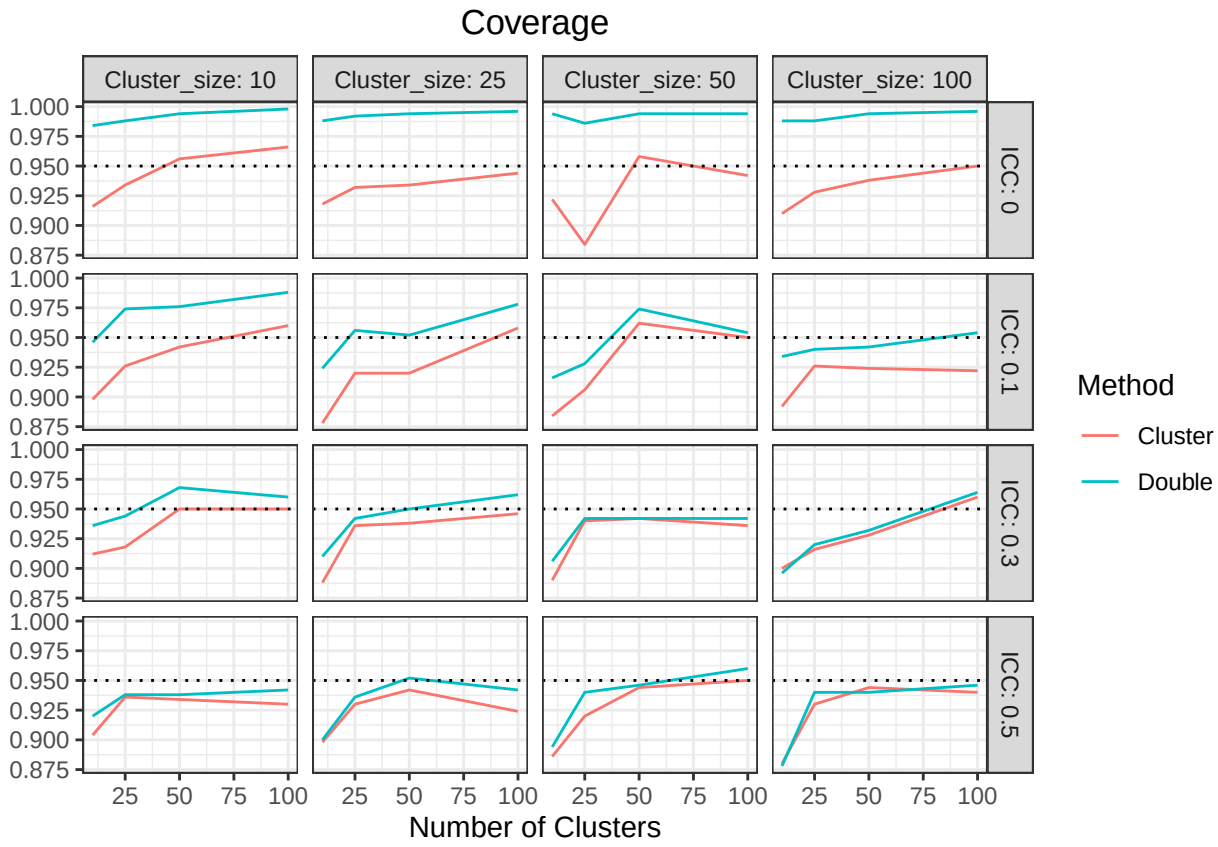

Figure 4: Coverage vs. Cluster Number

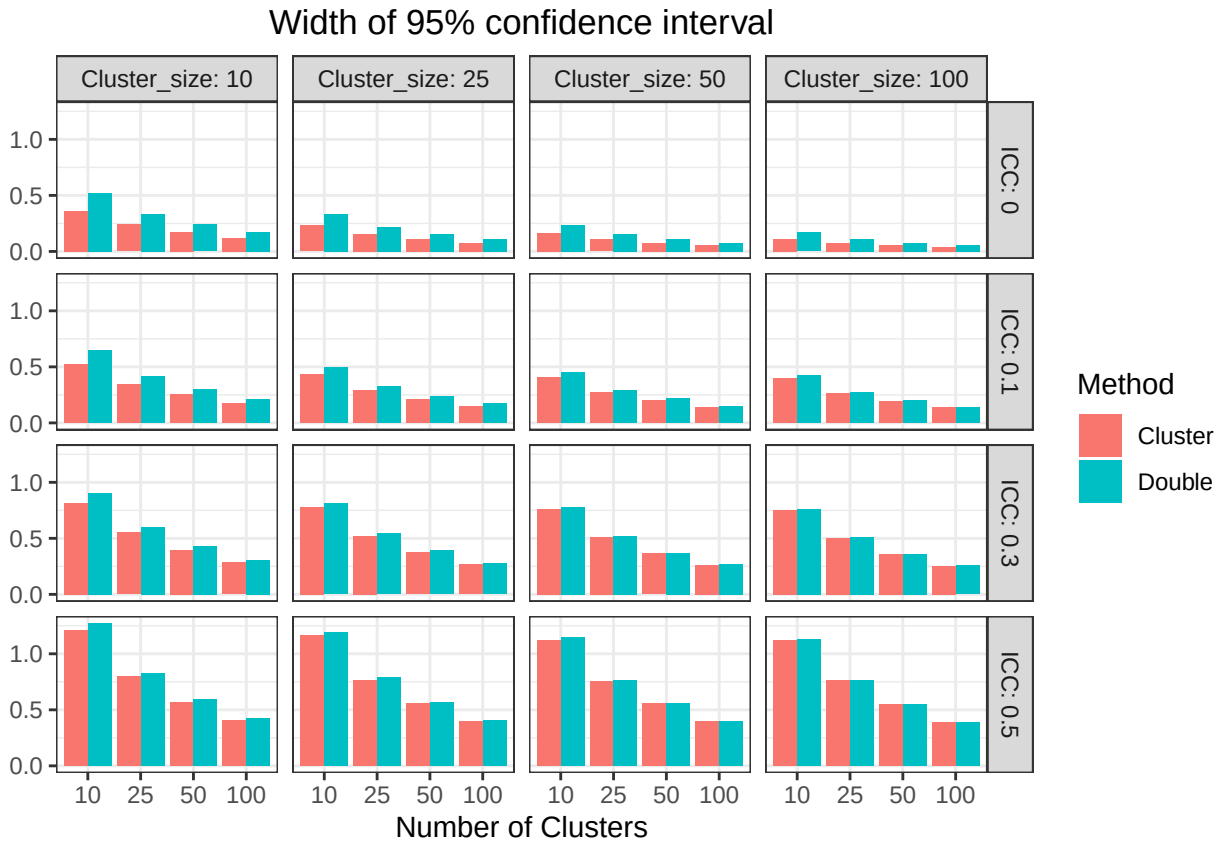

Figure 5: Confidence Interval Widths

the true population mean, we suspect that the standard bootstrap displays significant downward bias in its standard error estimation for this data set with high ICC.

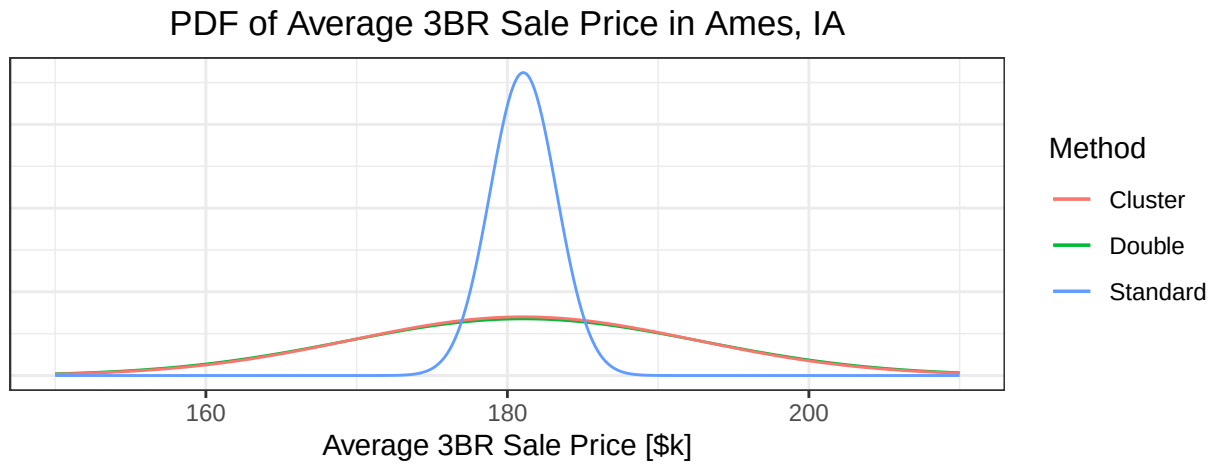

Figure 6: AMES Housing Case Study

New Method: Reweighted Double Bootstrap using a Gaussian Kernel

The simulation and case study results presented an interested opportunity: could we design a bootstrapping procedure combining the coverage guarantees of the double / cluster bootstrap without exhibiting upward bias in certain situations (e.g,. little to no ICC)?

To reduce the width of a confidence interval, one can either decrease the level of the test, increase the sample size or reduce the variance. The later seems most influenced by parameters of the bootstrap which directly affects the standard error estimates. In theory, increasing the precision of the bootstrap estimates will reduce the standard error and thus the overall confidence interval width.

We propose a weighted "2-stage" bootstrap method, a slight variation of the double bootstrap introduced above. Cluster resampling remains identical, however, the second stage is altered to assign unequal resampling weights to each cluster observation. The rationale for unequal weighting is as follows:

The closer an observation is to other observations, the more importance or signal it contains about the parameter of interest. To compute proximity to other observations, we apply a Gaussian kernel to each observation in the data set yielding a Gram matrix, which assigns more weight to points in close proximity. Each row in the Gram matrix is summed and corresponds with an overall proximity metric for each data point. This measure is normalized to yield quantities summing to 1 and are used as sampling weights for the bootstrap sampling algorithm.

Figures 9 and 10 in the appendix show performance of our new method against the three mentioned earlier. At low values of ICC we observe the most "shrinkage" of the confidence intervals. However, performance remains similar to both cluster bootstrapping methods, and we chose to move this result to the appendix. We had hoped the role of sigma in the Gaussian kernel would be akin to a tuneable hyperparameter in machine learning, providing the practitioner with a lever to affect standard error estimates. Unfortunately, we were disappointed when simulations showed negligible differences between various levels of sigma (see Figures 11 and 12 in the appendix).

Conclusions & Future Work

In this paper, we confirmed Harden's original findings indicating a downward standard error estimation bias when cluster information is disregarded and expanded his conclusion to a wider range of circumstances via Monte Carlo simulation. We found the improved coverage (and reduced type I error) of cluster based bootstrap methods to correspond with significantly wider confidence intervals, which could affect statistical significance in some settings and discourage widespread use.

In addition, we proposed a new "2-stage" kernel based bootstrapping method to try and address the wide intervals from cluster-based uncertainty. However, we were unsuccessful in developing a "better" tuneable

method.

Future extensions of our work are many. We would like to explore the kernel based smoothing idea further by trying a wider range of sigmas and or different kernel functions. We would also like to explore the role of multi-level clustering (e.g., students in schools in districts in states) and how clustering at different levels could affect overall standard error estimates.

Acknowledgements

We would like to thank Dr. Fan Li for inspiring us to pursue this topic and Rihui Ou/Yunran Chen for excellent guidance while learning causal inference.

References

1. Jo B, Asparouhov T, Muthén BO. Intention-to-treat analysis in cluster randomized trials with noncompliance. Stat Med. 2008;27(27):5565-5577. doi:10.1002/sim.3370
2. Arceneaux, Kevin, and David W. Nickerson. 2009. "Modeling Certainty with Clustered A Comparison of Methods." Political Analysis 1
3. Harden, Jeffrey J. "A Bootstrap Method for Conducting Statistical Inference with Clustered Data." State Politics & Policy Quarterly, vol. 11, no. 2, 2011, pp. 223–46, http://www.jstor.org/stable/41575822. Accessed 18 Apr. 2022.
4. Matthew D. Webb, 2014. "Reworking Wild Bootstrap Based Inference For Clustered Errors," Working Paper 1315, Economics Department, Queen's University.
5. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

Appendix

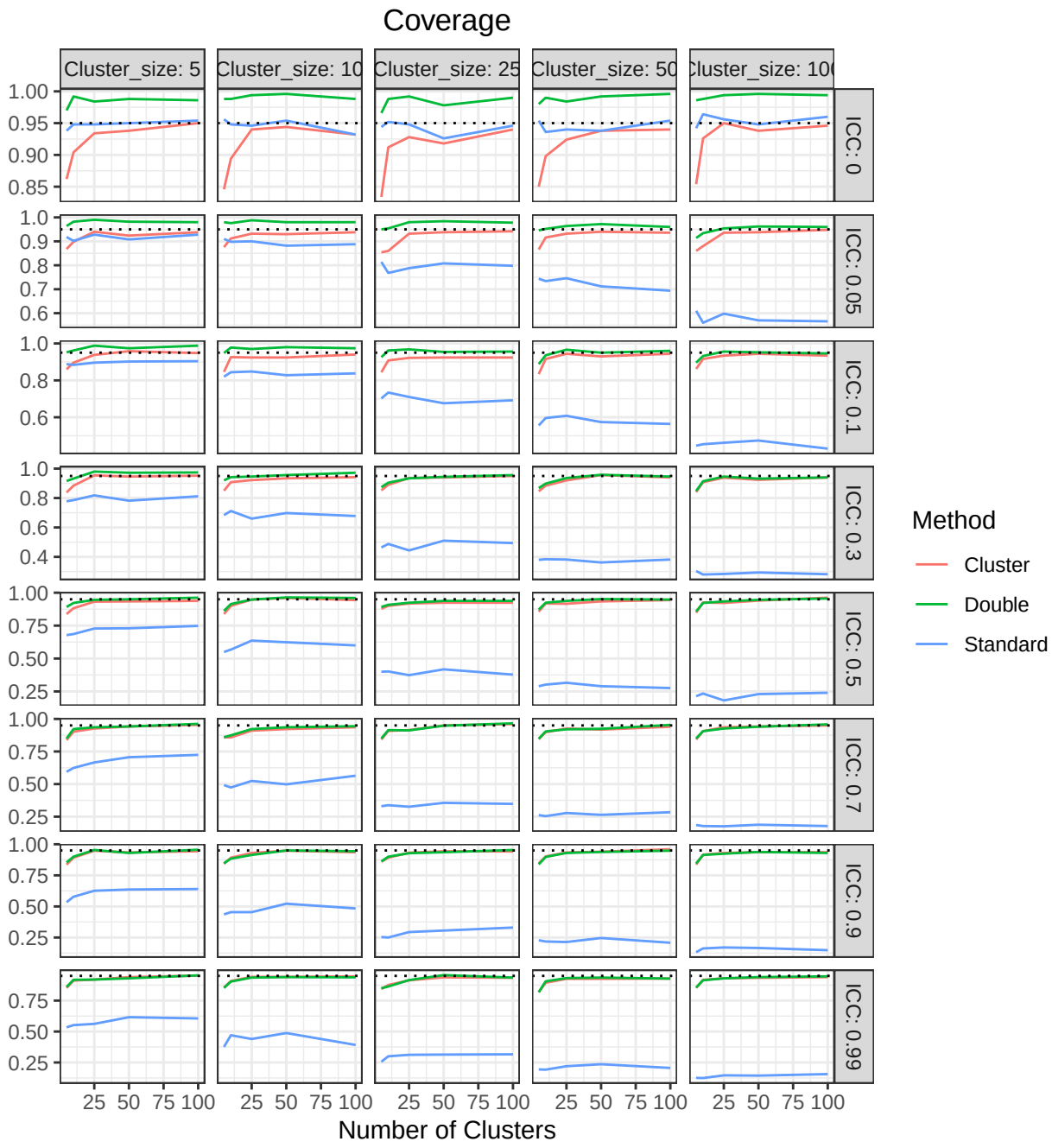

Figure 7: Simulations over Broader Parameter Set

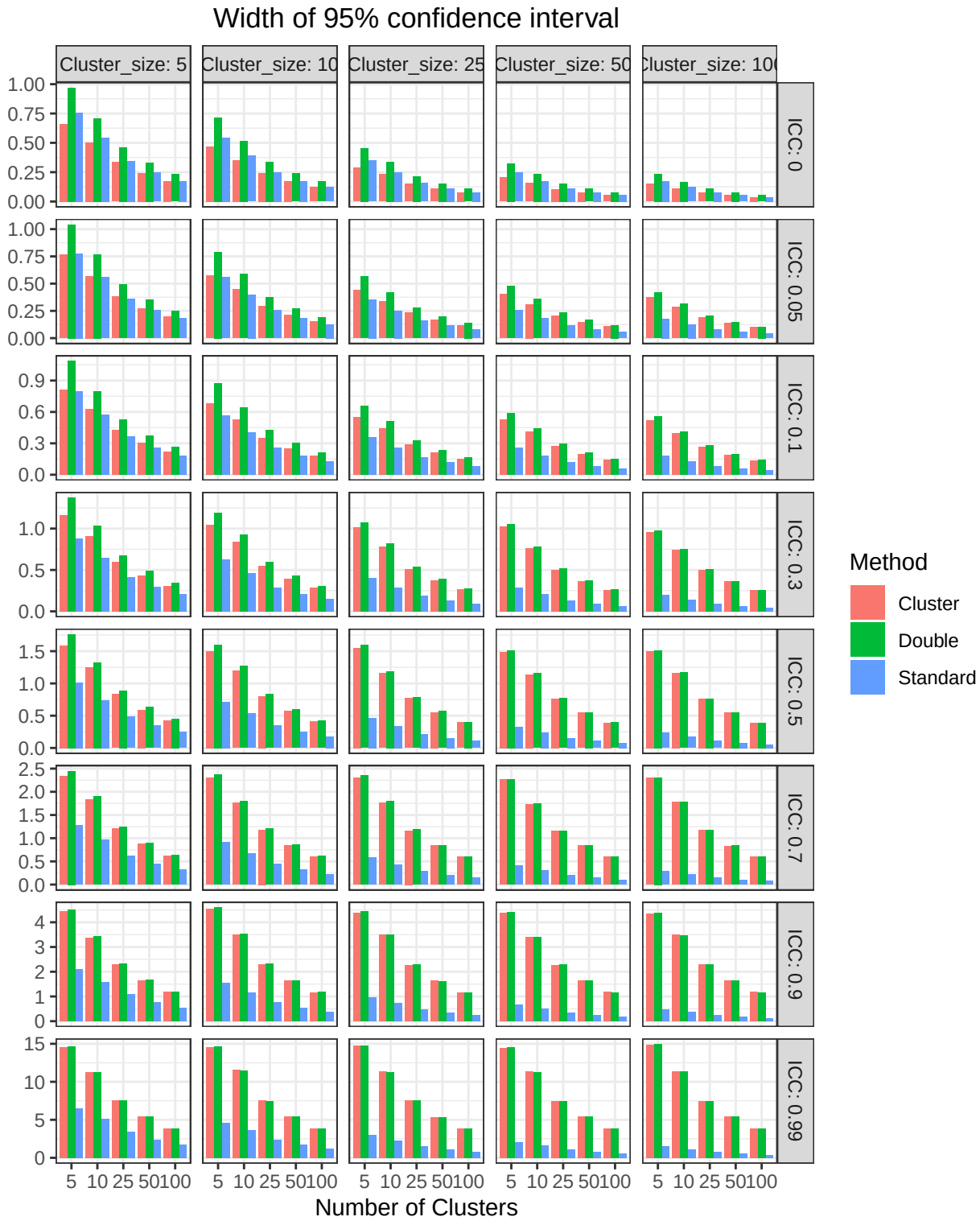

Figure 8: Corresponding CI Widths

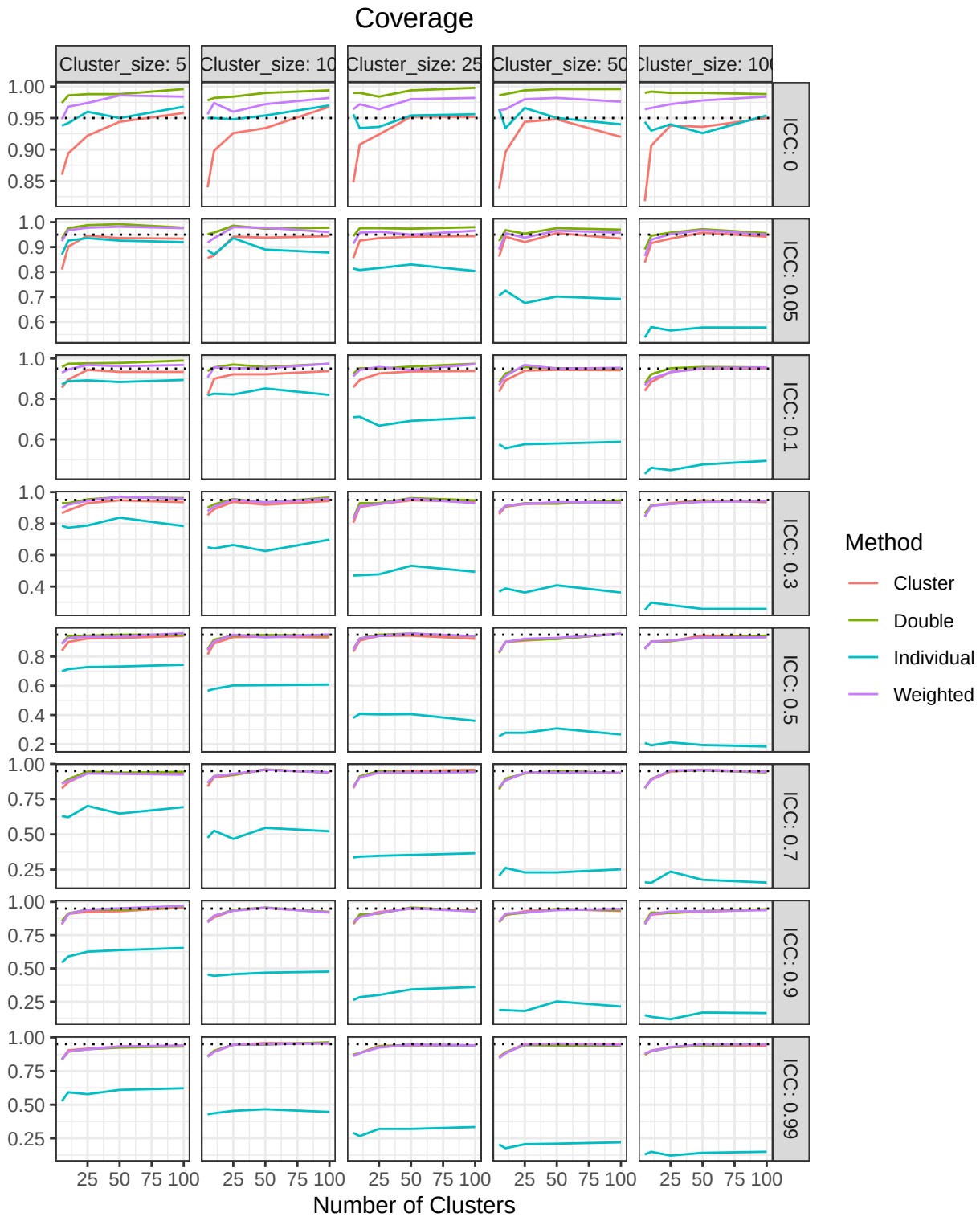

Figure 9: Broader Parameter Set with Kernel Weighted Method

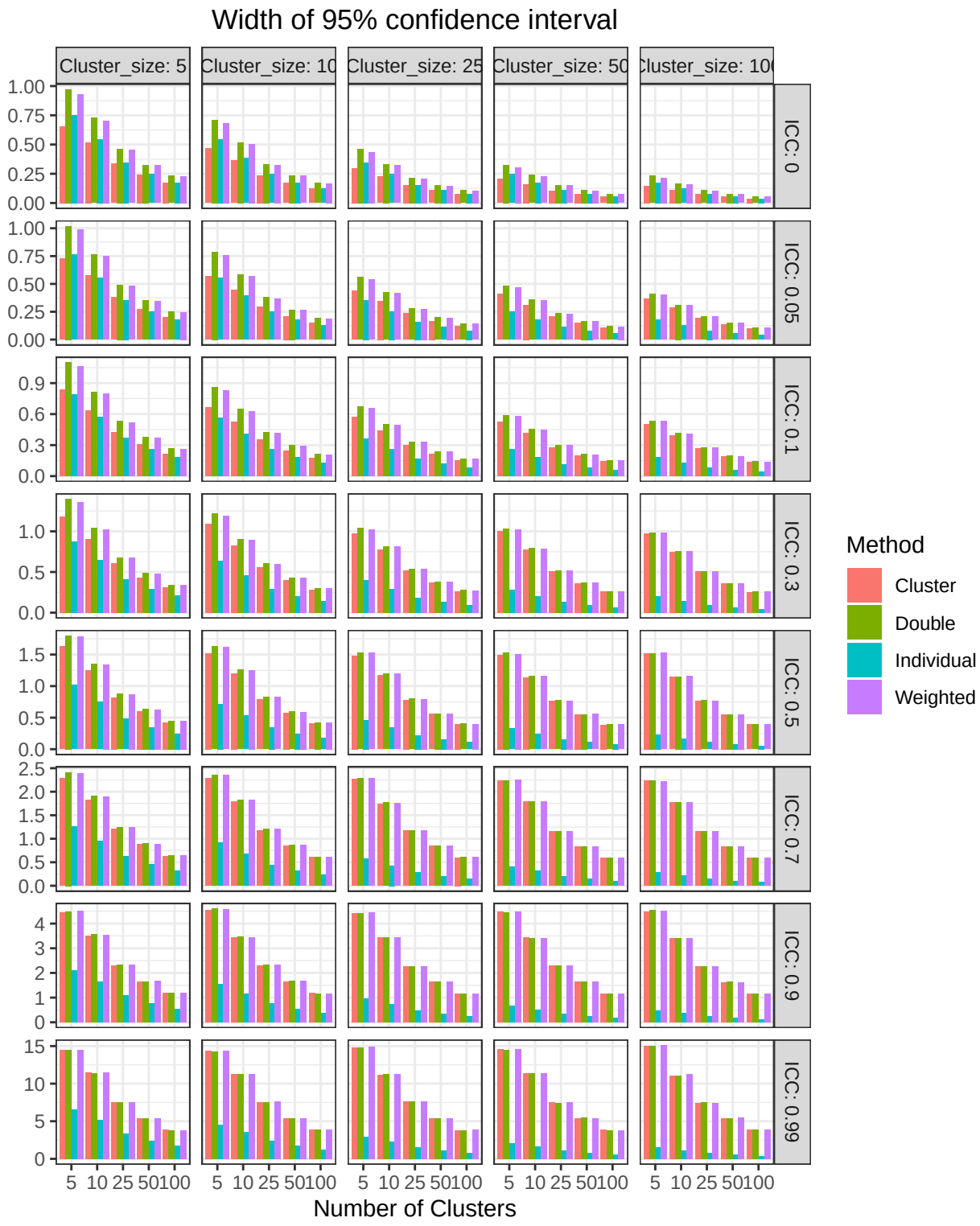

Figure 10: CI Widths with Weighted Method

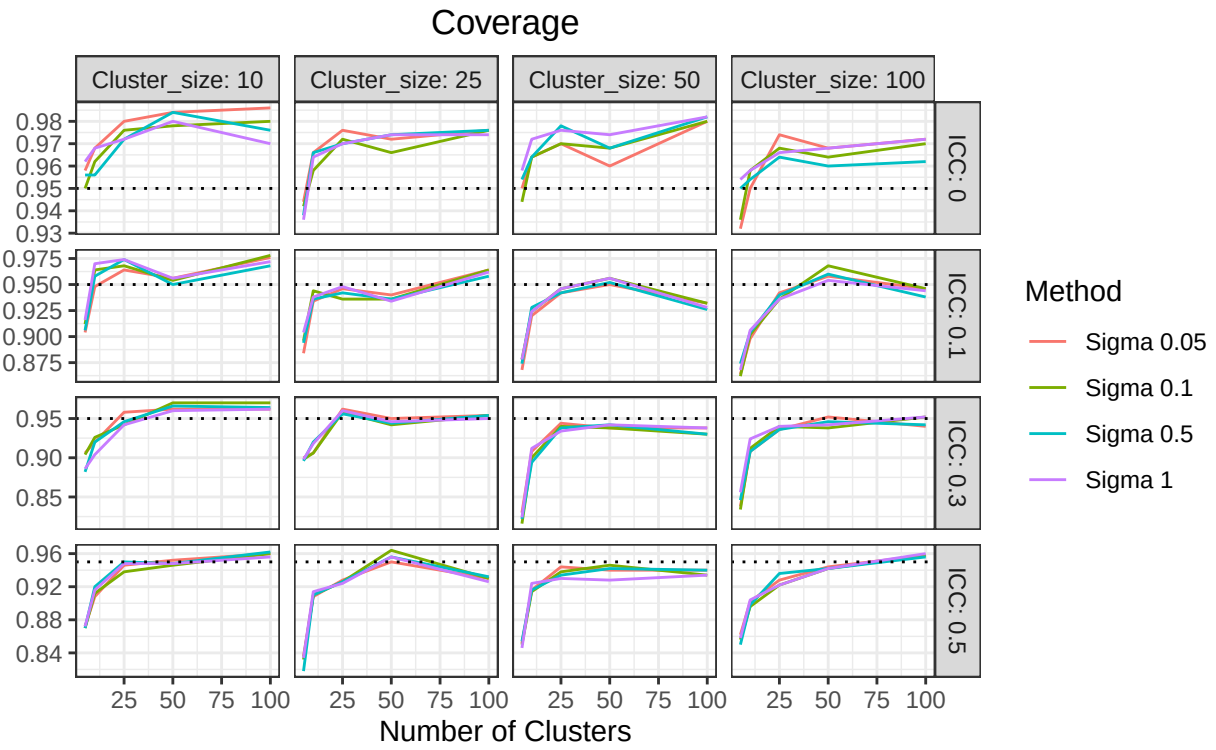

Figure 11: Sensitivity of Weighted Bootstrap to Width of Gaussian Kernel

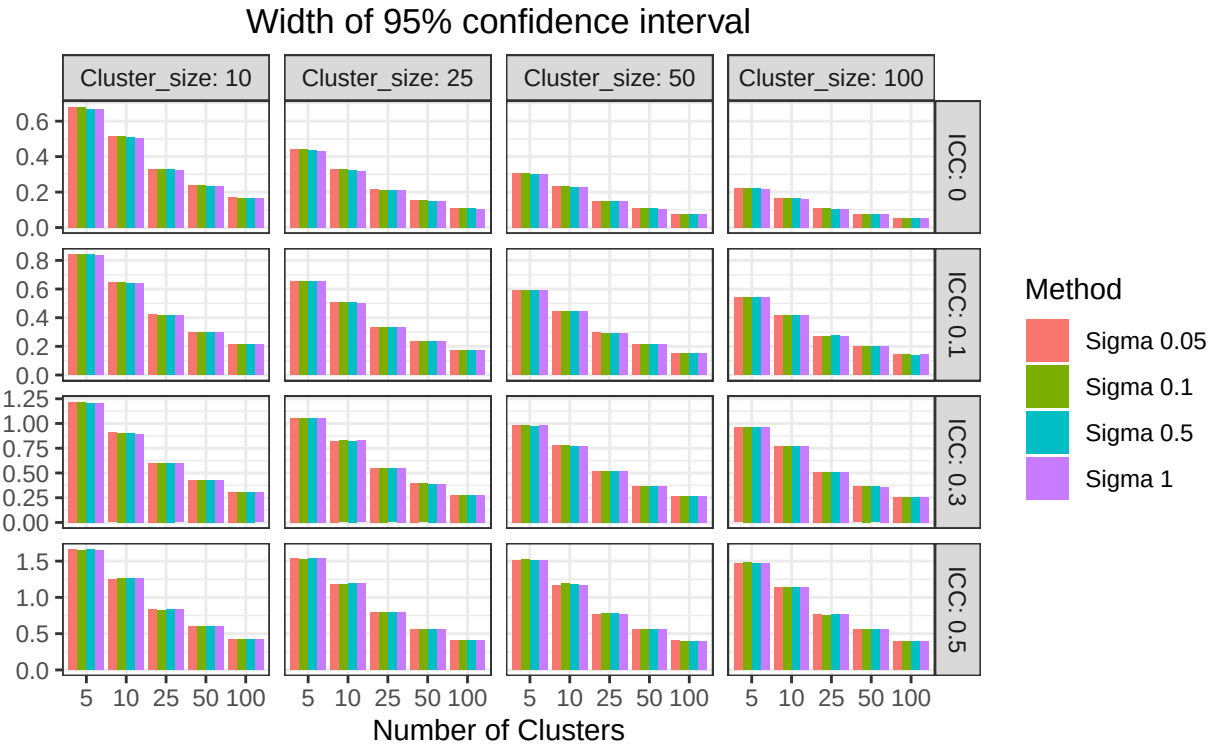

Figure 12: CI Widths