UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Moderating the Mediation Bootstrap for Causal Inference

van Garderen, K.J.; van Giersbergen, N.

DOI
10.48550/arXiv.2412.11285

Publication date
2024

Document Version
Final published version

License
CC BY

Link to publication

Citation for published version (APA):
van Garderen, K. J., & van Giersbergen, N. (2024). *Moderating the Mediation Bootstrap for Causal Inference*. (v1 ed.) ArXiv. https://doi.org/10.48550/arXiv.2412.11285

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:15 Sep 2026

arXiv:2412.11285v1 [econ.EM] 15 Dec 2024

Moderating the Mediation Bootstrap for Causal Inference

Kees Jan van Garderen
Amsterdam School of Economics
University of Amsterdam
K.J.vanGarderen@uva.nl

Noud van Giersbergen
Amsterdam School of Economics
University of Amsterdam
N.P.A.vanGiersbergen@uva.nl

Abstract

Mediation analysis is a form of causal inference that investigates indirect effects and causal mechanisms. Confidence intervals for indirect effects play a central role in conducting inference. The problem is non-standard leading to coverage rates that deviate considerably from their nominal level. The default inference method in the mediation model is the paired bootstrap, which resamples directly from the observed data. However, a residual bootstrap that explicitly exploits the assumed causal structure ($X \to M \to Y$) could also be applied. There is also a debate whether the bias-corrected (BC) bootstrap method is superior to the percentile method, with the former showing liberal behavior (actual coverage too low) in certain circumstances. Moreover, bootstrap methods tend to be very conservative (coverage higher than required) when mediation effects are small. Finally, iterated bootstrap methods like the double bootstrap have not been considered due to their high computational demands. We investigate the issues mentioned in the simple mediation model by a large-scale simulation. Results are explained using graphical methods and the newly derived finite-sample distribution. The main findings are: (i) conservative behavior of the bootstrap is caused by extreme dependence of the bootstrap distribution's shape on the estimated coefficients (ii) this dependence leads to counterproductive correction of the the double bootstrap. The added randomness of the BC method inflates the coverage in the absence of mediation, but still leads to (invalid) liberal inference when the mediation effect is small.

Keywords: finite-sample analysis, bootstrap inference, mediation, indirect effects

1 Introduction

This paper analyses various bootstrap inference methods for indirect effects in the simple mediation model. Meditation is concerned with how the effect of a causal variable X on a consequent variable Y is possibly transmitted through an intervening variable M. The analysis of mediating processes has a long history that can be traced back to the path analysis introduced by Wright (1920) and originally formulated as a statistical hypothesis by Woodworth (1928). During the 1950s, mediation analysis as we know it today was developed in the social sciences, with main contributions in psychology, see for instance Rozeboom (1956). The seminal paper of Baron and Kenny (1986) laid out statistical requirements for detecting a true mediation relationship and made a distinction between mediating and moderating variables. The current approach favored by Hayes and others, see for instance Preacher et al. (2007), has shifted the focus, but both approaches employ a set of two regression models, with coefficients that are used to measure the direct and indirect (mediation) effect. The simple mediation model is given by the following bivariate recursive system (i.e. triangular system with diagonal disturbance covariance matrix):

$$m = \theta_x x + v, \tag{1}$$

$$y = \beta_x x + \beta_m m + u, \tag{2}$$

where y, x and m denote $n \times 1$ observable vectors, while u and v are $n \times 1$ non-observed error vectors. The equations could include an intercept, but we assume without loss of generality that all variables, i.e. y, x and m, are expressed in deviation from their means.[1] The (indirect) mediation effect is the product $\gamma = \theta_x \beta_m$, which can be estimated by $\hat{\theta}_x \hat{\beta}_m$. We are interested in constructing confidence intervals for the mediation effect.

A classic method for the construction of a confidence interval of γ is based on the asymptotic standard normal approximation of the studentized quantity:

$$\frac{\hat{\gamma} - \gamma}{SE(\hat{\gamma})} \overset{a}{\sim} N(0,1),$$

where $SE(\hat{\gamma})$ denotes the standard error of $\hat{\gamma}$, typically obtained by the delta method. An alternative method is to exploit the quantiles of the product of two standard normal distributions directly; see Craig (1936) and Aroian (1947). Craig (1936) showed that this distribution of the product is symmetric with a kurtosis of 6 when the distributions are independent and both have a zero mean, i.e. $\alpha_x = \beta_m = 0$. Using numerical

[1] Variables in deviation from their means can be obtained as residuals after regression on a constant. In fact one may also use other variables, e.g. observable confounders, in such a preliminary regression. This only affects the degrees of freedom in the t-distributions below.

integration, Meeker et al. (1981) tabulated quantiles of the product of two normally distributed variables.

Asymptotic results may be of limited worth in small samples, especially when the distribution of $\hat{\gamma}$ is highly non-normal. The bootstrap addresses these small-sample issues like non-normality and is currently the preferred method to construct confidence intervals for γ in practice; see for instance MacKinnon et al. (2007), MacKinnon et al. (2004), Shrout and Bolger (2002), Montoya and Hayes (2017) and for an earlier contribution Bollen and Stine (1990). The prevalent bootstrap approach is a paired bootstrap, where resamples are drawn from the tuples (y_i, x_i, m_i) with replacement. Various influential papers recommend the bias-corrected (BC) method, e.g. MacKinnon et al. (2004, p. 120) (more than 7,260 citations per May 2, 2022) state "As a result, the single best method overall was the bias-corrected bootstrap ...", while also noting that "The bias-corrected bootstrap did have Type I error rates that were above the robustness interval for some parameter combinations ...". Such liberal inference disqualifies the BC method as a valid inference method not only from a theoretical point of view, but also in practice with deviations that can be substantial. When testing is based on these confidence intervals, as is common practice, it is not size correct and overrejection can be substantial as the simulation results show. The validity of bootstrap methods might also be fundamentally problematic given the absence of pivotal statistics and strong dependence on nuisance parameters.

There are many other bootstrap methods, however, that might be valid, including residual, single versus double, or parametric versus non-parametric bootstrap. In addition, various different versions of constructing confidence intervals exist, including the five main approaches discussed in Section 3. We follow up on MacKinnon et al. (2004) who conclude that there are several ways to improve the confidence limits worth investigating. We address the question which methods have correct (minimal) coverage rates and which method in the plethora of bootstrap approaches is preferred. The double bootstrap in particular seems an obvious candidate to correct for coverage errors associated with single bootstrap procedures. It is computationally more demanding, however, and coverage of such confidence intervals has not been investigated in the mediation setting. So, we analyze the double bootstrap in detail and find that it over-corrects for specific parameter constellations and parts of the sample space. We provide an explanation for this finding.

In order to analyze the bootstrap methods we use known asymptotic results, but also derive a number of new exact (finite-sample) distributional results. Under strict normality of the errors, we derive the joint distribution of $\hat{\theta}_x$ and $\hat{\beta}_m$ and show that, conditionally on x, the two estimators are independent and unbiased. This implies that $\hat{\theta}_x\hat{\beta}_m$ is mean unbiased for γ and suggests that the commonly used bias correction is

redundant and only introduces unnecessary and harmful randomness when constructing confidence intervals, although the BC method is based on the median instead of the mean bias. The distribution of $\hat{\gamma}$ is shown to be a Mellin convolution of a student-t and normal distribution with a skewness that only disappears if θ_x and β_m are zero. This distribution has fatter tails than the distribution of the product of two normal random variables due to the student-t distribution.

The outline of the paper is as follows. In Section 2, we will derive finite-sample properties of $\hat{\gamma}$ assuming errors are normally distributed. Although the attractiveness of the non-parametric bootstrap lies in the fact that no distributional assumptions are required, the stylized Gaussian setup will act as a benchmark: if the bootstrap does not work in this setup, then it will neither work in the non-Gaussian setup. Section 3 describes the residual and paired bootstrap and the various confidence intervals. Section 4 contains a more elaborate exposition of the double bootstrap for confidence intervals since it has not yet been investigated or applied in the mediation setting. The Monte Carlo results are shown in Section 5, where the finite-sample results derived in Section 2 are useful for explaining some of the observed results. Concluding remarks are given in Section 6.

2 Relevant Exact Finite-Sample Results

The bootstrap can be interpreted as a simulation method in which population parameters are (implicitly or explicitly) replaced by sample analogs. The residual bootstrap resamples regression residuals and could be carried out easily in the simple mediation model given in (2)-(1) recursively generating m and y, while keeping x fixed; see equations (11) and (12). On the other hand, the paired bootstrap directly resamples from the observed data; see equation (10). The former approach respects the endogeneity of y and m, and explicitly treats x as exogenous, whereas the latter approach only does so implicitly. Although we find that the paired bootstrap works similar as the residual bootstrap, the properties are more directly investigated using a parametric bootstrap approach for the residual bootstrap. In order to characterize the parametric bootstrap, we derive the joint finite-sample distribution of $\hat{\theta}_x$ and $\hat{\beta}_m$, and the distribution of their product under the normality assumption: $u_i \sim N(0, \sigma_u^2)$ independent of $v_i \sim N(0, \sigma_v^2)$ for $i = 1, ..., n$.

The obvious estimator of the indirect effect is the product of $\hat{\alpha}_x$ and $\hat{\beta}_m$ from the two regressions involving different covariates: $\hat{\theta}_x$ conditional on x, and $\hat{\beta}_m$ conditional on both x and m. Given the independence of the two error terms, the system in (2)-(1) is called recursive. Therefore, all model parameters can be estimated consistently and efficiently using ordinary least squares (OLS); see inter alia Rothenberg and Leenders

(1964). Conditional on x, equation (1) satisfies all classical Gaussian linear regression assumptions and the maximum likelihood estimator for the conditional mean parameters equals the OLS estimator. The same holds for equation (2) conditional on both x and m. We therefore have the standard exact distributional results that, conditional on x:

$$\hat{\theta}_x = (x'x)^{-1}x'm|x \sim N\left(\alpha_x, \sigma_v^2(x'x)^{-1}\right). \tag{3}$$

and, conditional on both x and m, with $X = [x : m]$ an $n \times 2$ matrix:

$$\begin{pmatrix} \hat{\beta}_x \\ \hat{\beta}_m \end{pmatrix} = (X'X)^{-1}X'y|(x,m) \sim N\left(\begin{pmatrix} \beta_x \\ \beta_m \end{pmatrix}, \sigma_u^2(X'X)^{-1}\right), \tag{4}$$

We show next in Proposition 1 that the estimators $\hat{\theta}_x$ and $\hat{\beta}_m$ are independent, which leads to analytical and numerical simplifications, and that the estimator $\hat{\beta}_m$ itself has a scaled $t_{(n-2)}$-distribution, rather than, as usual, its t-ratio. Since the student-t distribution has fatter tails than the normal distribution, especially when the degrees of freedom are small, inference based on the product of two normal distributions can be misleading.

Proposition 1 *In the Gaussian simple mediation model (2)-(1) with $u \sim N(0, \sigma_u^2 I_n)$ independent of $v \sim N(0, \sigma_v^2 I_n)$, the estimators $\hat{\theta}_x$ and $\hat{\beta}_m$ are independent given x with their joint distribution the product of the normal distribution given in equation (3)and a t-distribution with location β_m and scale parameter $\sqrt{n-2}\sigma_v/\sigma_u$, and $(n-2)$ degrees of freedom, or, expressed in terms of a standard t-distribution:*

$$f_{\hat{\beta}_m}(b) = f_{t(n-2)}\left(\sqrt{n-2}\frac{\sigma_v}{\sigma_u}(b - \beta_m)\right)\sqrt{n-2}\frac{\sigma_v}{\sigma_u}. \tag{5}$$

The probability density function (pdf) of $\hat{\gamma} = \hat{\alpha}_m\hat{\beta}_m$ can in principle be derived using:

$$f_{\hat{\gamma}}(g; \theta_x, \beta_m, \sigma_v^2, \sigma_u^2, x'x) = \int_{-\infty}^{\infty} f_{\hat{\theta}_x}(a) f_{\hat{\beta}_m}(g/a)\frac{1}{|a|}\,\mathrm{d}a, \tag{6}$$

see e.g. Mood et al. (1974, p. 187), but this does not lead to a closed-form expression. Equation (6) nevertheless useful provides a convenient way to numerically determine the density and its associated probabilities. We use equation (6) in Section 5 for the parametric bootstrap distribution with $(\theta_x, \beta_m, \sigma_v^2, \sigma_u^2, x'x)$ evaluated at $(\hat{\theta}_x, \hat{\beta}_m, s_v^2, s_u^2, x'x)$.

The next proposition gives the first three moments of the estimator $\hat{\gamma} = \hat{\theta}_x\hat{\beta}_m$. In particular, equation (7) shows that $\hat{\gamma}$ is *mean* unbiased. This does not imply that it is also *median* unbiased, however, due to the skewness given in equation (9). The distribution of $\hat{\gamma}$ is skewed when $\theta_x \neq 0$ and $\beta_m \neq 0$, although the distributions of the individual estimators $\hat{\theta}_x$ and $\hat{\beta}_m$ are both symmetric. If $\theta_x\beta_m > 0$, the distribution of $\hat{\gamma}$ positively skewed, while it is negatively skewed if $\alpha_x\beta_m < 0$. Note that the BC method uses the median to bias correct the confidence interval.

The expression for the variance is new to the literature. It can be decomposed in terms of different orders. Assuming $x'x = O(n)$, or in probability if x_i is i.i.d., the first two terms in (8) are of size $O(n^{-1})$ when $\theta_x \neq 0$ and $\beta_m \neq 0$. The last term in (8) is of smaller magnitude $O(n^{-2})$, and is always larger than zero, even when $\theta_x = \beta_m = 0$.

Proposition 2 *The expectation, variance and skewness of $\hat{\gamma} = \hat{\theta}_x \hat{\beta}_m$ in the Gaussian simple mediation model, conditional on x, are equal to:*

$$\mathbb{E}[\hat{\gamma}|x] = \gamma, \tag{7}$$

$$Var(\hat{\gamma}|x) = \theta_x^2 \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{(n-4)} + \beta_m^2 \frac{\sigma_v^2}{x'x} + \frac{\sigma_u^2}{x'x} \frac{1}{(n-4)}, \tag{8}$$

$$Skewness(\hat{\gamma}|x) = \frac{\mathbb{E}[(\hat{\gamma} - \mathbb{E}[\hat{\gamma}|x])^3|x]}{Var(\hat{\gamma}|x)^{3/2}} = \frac{6\theta_x \beta_m \sigma_u^2}{(n-4)x'x} \frac{1}{Var(\hat{\gamma}|x)^{3/2}}. \tag{9}$$

The next proposition considers the distributions of the appropriate centered t-statistic for θ_x under $H_0 : \theta_x = \theta_x^0$ and β_m under $H_0 : \beta_x = \beta_x^0$, where the superscript 0 indicates the true value. Since the system is recursive, the t-statistics for θ_x and β_m have a t_{n-1} and t_{n-2}-distribution conditional on the regressors in the model. Although the variance of $\hat{\beta}_m$ in the second model depends on the residuals $\hat{u}$ of the first model, the t-distributions are still independent from each other as shown in the Proposition 3.

Proposition 3 *In the Gaussian simple mediation model (2)-(1) with $u \sim N(0, \sigma_u^2 I_n)$ independent of $v \sim N(0, \sigma_v^2 I_n)$, the t-statistics for testing $H_0 : \theta_x = \theta_x^0$ and $H_0 : \beta_x = \beta_x^0$ are independent and t-distributed:*

$$t_{\theta_x} = \frac{\hat{\theta}_x - \theta_x^0}{SE(\hat{\theta}_x)} \sim t_{n-1} \qquad \text{and} \qquad t_{\beta_m} = \frac{\hat{\beta}_m - \beta_m^0}{SE(\hat{\beta}_m)} \sim t_{n-2}.$$

3 Bootstrap Inference

We consider two main bootstrap approaches that are used in the regression model: (i) the paired bootstrap proposed by Efron (1979) and (ii) the residual bootstrap first analyzed in Bickel and Freedman (1981). The paired bootstrap is the one generally applied in papers on mediation for bootstrap inference. The paired-bootstrap results are then interpreted conditional on x, but given the causal structure assumed in the mediation setup as $x \to m \to y$, it might be more intuitive to also consider the residual bootstrap; see Hall (1992, Section 4.3.2) for a more detailed discussion about the different assumptions underlying the paired and residual bootstrap in a regression context.

If $w_i = (y_i, x_i, m_i)$ denotes the vector containing the i-th observation, then the general idea for constructing bootstrap confidence intervals can be summarized as follows:

1. Given the data $w_1, ..., w_n$, generate a bootstrap sample of size n denoted as $w_1^*, ..., w_n^*$.

2. Calculate an appropriate quantity using the bootstrap sample. For instance, the estimate $\hat{\gamma}^* = \hat{\theta}_x^* \hat{\beta}_m^*$ or the studentized root $\tau^* = (\hat{\gamma}^* - \hat{\gamma})/SE(\hat{\gamma}^*)$.

3. Repeat steps 1 and 2, B times to obtain B bootstrap replications $\hat{\gamma}_1^*, ..., \hat{\gamma}_B^*$ or $\tau_1^*, ..., \tau_B^*$.

4. Use the B bootstrap replications to construct a confidence interval.

There are several ways to construct the bootstrap sample $w_1^*, ..., w_n^*$ in step 1. The paired bootstrap simply resamples from the original $w_1, ..., w_n$ with probabilities:

$$\mathbb{P}(w_i^* = w_j) = \frac{1}{n} \qquad \text{for } i, j = 1, ..., n. \tag{10}$$

A residual bootstrap generates $w_1^*, ..., w_n^*$ by resampling bootstrap errors from the residuals, or from a fitted (parametric) distribution, and subsequently constructing w_i^* according to the estimated model. So in the mediation model, bootstrap errors u^* and v^* are drawn and the bootstrap observables $w_i^* = (y_i^*, m_i^*, x_i)$ constructed using the estimated parameter values as:

$$y^* = \hat{\beta}_x x + \hat{\beta}_m m^* + u^*, \tag{11}$$

$$m^* = \hat{\theta}_x x + v^*. \tag{12}$$

In the non-parametric bootstrap, u_i^* and v_i^* are sampled with replacement from the rescaled OLS residuals $\sqrt{n/(n-3)}\hat{u}_i$ and $\sqrt{n/(n-2)}\hat{v}_i$ respectively for $i = 1, ..., n$, with the rescaling as originally suggested by Efron (1982). In the parametric bootstrap, one might draw $u_i^* \sim N(0, \hat{\sigma}_u^2)$ independent of $v_i^* \sim N(0, \hat{\sigma}_v^2)$, using the estimated variances, instead of resampling residuals. The residual bootstrap allows for clear-cut conditioning by keeping x fixed in (11) and (12).

The following main methods for constructing confidence intervals have been presented in the bootstrap literature: (i) *basic* (ii) *percentile* (iii) *bias-corrected (BC) percentile* (iv) *bias-corrected and accelerated (BC$_a$)* and (v) *percentile-t* methods.

(i) The *basic method* for constructing a two-sided equal-tailed $(1 - \alpha)$ confidence interval is based on the idea that the distribution of $\hat{\gamma} - \gamma$ can be approximated by $\hat{\gamma}^* - \hat{\gamma}$ leading to the following interval:

$$(\hat{\gamma} - q_{1-\alpha/2}^*, \hat{\gamma} - q_{\alpha/2}^*), \tag{13}$$

where q_α^* denotes the α-quantile of $\hat{\gamma}^* - \hat{\gamma}$, i.e. $\mathbb{P}^*[\hat{\gamma}^* - \hat{\gamma} \leq q_\alpha^*] = \alpha$; see Davison and Hinkley (1997, Section 5.2). Due to the fact that $q_\alpha^* \equiv (\hat{\gamma}^* - \hat{\gamma})_\alpha = \hat{\gamma}_\alpha^* - \hat{\gamma}$ with $\hat{\gamma}_\alpha^*$

the α-quantile of the bootstrap distribution, we can write the basic confidence interval alternatively as:

$$(2\hat{\gamma} - \hat{\gamma}^*_{1-\alpha/2}, 2\hat{\gamma} - \hat{\gamma}^*_{\alpha/2}).$$

Comparing the quantiles with those of the percentile and percentile-t explains why this interval is also known as the hybrid interval; see e.g. Shao and Tu (1996, Section 4.1). Note that the lower confidence limit is based on the upper tail of the bootstrap distribution, while the upper confidence limit uses the lower tail. Hence, asymmetry in the basic confidence interval is opposite to the asymmetry of the percentile interval. This is an attractive feature because if $\hat{\gamma}^* - \hat{\gamma}$ is positively skewed, this suggests that larger values of $\hat{\gamma}$ could more easily be generated by smaller values of γ than the other way round.

(ii) The two-sided *percentile* confidence interval is given by:

$$(\hat{\gamma}^*_{\alpha/2}, \hat{\gamma}^*_{1-\alpha/2}), \tag{14}$$

where $\hat{\gamma}^*_\alpha$ denotes the α-quantile of the bootstrap distribution, i.e. $\mathbb{P}^*[\hat{\gamma}^* \leq \hat{\gamma}^*_\alpha] = \alpha$; see e.g. Efron (1981).

(iii) Efron (1981) also introduces the *BC percentile* method as an improvement to correct for estimation bias. It uses the proportion of bootstrap replications less than the original estimate $\hat{\gamma}$:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\hat{\gamma}^*_b < \hat{\gamma}\}\right),$$

where $\Phi^{-1}(\cdot)$ denotes the inverse function of the standard normal distribution function, $\mathbb{1}\{\cdot\}$ denotes the indicator function, and $\hat{\gamma}^*_b$ is the b-th bootstrap realization. So, $\hat{z}_0$ measures the bootstrap approximation of the *median* bias of $\hat{\gamma}^*$ in normal units. If exactly half of the $\hat{\gamma}^*_b$ is less than $\hat{\gamma}$, then $\hat{z}_0 = 0$. Although Proposition 2 has established that the estimator $\hat{\gamma}$ is mean unbiased, suggesting that $\hat{z}_0$ is close to zero, the correction as defined by Efron (1981) uses the median, which differs from the mean because of the skewness.

(iv) The BC_a interval proposed by Efron (1987) not only corrects for bias, but also for skewness by the so-called acceleration constant a. There are various ways to estimate the acceleration constant a, but a commonly used estimate based on jackknife values is given by:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\gamma}_{(\cdot)} - \hat{\gamma}_{(-i)})^3}{6\{\sum_{i=1}^{n}(\bar{\gamma}_{(\cdot)} - \hat{\gamma}_{(-i)})^2\}^{3/2}}, \tag{15}$$

where $\hat{\gamma}_{(-i)}$ denotes the i-th jackknife value based on the sample information excluding the i-th observation $w_i = (y_i, x_i, m_i)$ and $\bar{\gamma}_{(\cdot)} = 1/n \sum_{i=1}^{n} \hat{\gamma}_{(i)}$ the average of the n jackknife

values $\hat{\gamma}_{(1)}, ..., \hat{\gamma}_{(n)}$. This is also the standard implementation used by, inter alia, the R-packages *Lavaan* and *Boot*. A two-sided $(1-\alpha)$ BC_a confidence interval can now be defined as the interval:

$$(\hat{\gamma}^*_{\alpha_1}, \hat{\gamma}^*_{\alpha_2}) \tag{16}$$

where the quantiles are based on the probabilities:

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}\right) \qquad \text{and} \qquad \alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}\right).$$

If the estimated skewness is 0, then $\hat{a} = 0$ and the BC_a interval reduces to the BC interval.

(v) Finally, a *percentile-t* confidence interval is defined by:

$$\left(\hat{\gamma} - \tau^*_{1-\alpha/2} SE(\hat{\gamma}), \hat{\gamma} - \tau^*_{\alpha/2} SE(\hat{\gamma})\right)$$

where τ^*_α denote the α-quantile of the studentized root $(\hat{\gamma}^* - \hat{\gamma})/SE(\hat{\gamma}^*)$; see Efron (1982).

Only the BC_a and the percentile-t methods are second-order accurate and are said to achieve asymptotic refinement, see for instance Hall (1992, Chapter 3). However, the accuracy of this latter method in practice depends on the accuracy of the standard error $SE(\hat{\gamma})$. Note that a well-behaved standard error for $\hat{\gamma}$ is problematic; see for instance simulation evidence in MacKinnon et al. (2002) for a variety of choices. The usual Sobel (1982) formula:

$$SE(\hat{\gamma}) = \sqrt{\hat{\theta}^2_x SE(\hat{\beta}_m)^2 + \hat{\beta}^2_m SE(\hat{\theta}_x)}$$

is used when reporting results for the percentile-t method. Results are also reported based on the Jacknife standard error:

$$SE_J(\hat{\gamma}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\bar{\gamma}_{(\cdot)} - \hat{\gamma}_{(-i)})^2},$$

which is related to the denominator of the estimated acceleration constant a shown in (15). Note that the asymmetry of the percentile-t method is in the same direction as the basic/hybrid method.

4 Double Bootstrap Methods

Simulation results reported in the mediation literature, for instance MacKinnon et al. (2004) and its follow-up study Fritz et al. (2012), show that bootstrap confidence intervals are liberal, i.e. the probability coverage is larger than the nominal $1-\alpha$ coverage, when the indirect effect γ is small. This does not lead to invalid inference, but to confidence

intervals that are too wide and therefore to very low probabilities of rejecting the null of no mediation. In particular, when testing the null hypothesis of no mediation, $H_0 : \gamma = 0$, by checking whether the value zero is included by the confidence interval, a liberal interval leads to a low rejection probability of the null. In fact rejection probabilities are very much lower than the significance level, which is a serious problem given that establishing a mediation effect is usually the primary purpose of this type of analysis.

When confidence intervals are liberal, a second-level bootstrap can possibly be used to estimate the overcoverage and correct for it. Such a procedure is called a double bootstrap and, despite its great potential, has hardly been investigated in the mediation setting and principal reason to investigate it here. The main idea is to adjust the quantiles used in the confidence intervals. We use the one-sided percentile method to illustrate the approach. Let $\mathcal{I}_1(\alpha; \mathcal{X}, \mathcal{X}^*) = (-\infty, \hat{\gamma}^*_{1-\alpha})$ denote the original percentile interval based on sample information $\mathcal{X}$, and resample information $\mathcal{X}^*$, as a function of the nominal coverage $1-\alpha$. The true coverage probability, denoted $\pi(\alpha) = \mathbb{P}[\gamma \in \mathcal{I}_1(\alpha; \mathcal{X}, \mathcal{X}^*)]$ could differ significantly from $1-\alpha$. Let δ_α denote the 'correct nominal' coverage such that, when used in the procedure, has $\pi(\delta_a) = 1-\alpha$. In general, δ_α is unknown since $\pi(\alpha)$ is unknown, but $\pi(\alpha)$ can be estimated by:

$$\hat{\pi}(\alpha) = \mathbb{P}[\hat{\gamma} \in \mathcal{I}_1(\alpha; \mathcal{X}^*, \mathcal{X}^{**})|\mathcal{X}].$$

In practice, this is estimated by simulation, based on B observed first-level bootstrap samples $\mathcal{X}^*_1, ..., \mathcal{X}^*_B$:

$$\hat{\pi}_B(\alpha) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\hat{\gamma} \in \mathcal{I}_1(\alpha; \mathcal{X}^*_b, \mathcal{X}^{**}_b)\}.$$

However, the distribution of $\mathcal{X}^{**}_b$ given $\mathcal{X}^*_b$ is generally unknown, but can be estimated in turn by C, second-level bootstrap samples $\mathcal{X}^{**}_{b1}, ..., \mathcal{X}^{**}_{bC}$ from $\mathcal{X}^*_b$. If $\hat{\delta}_\alpha$ solves $\hat{\pi}_B(\hat{\delta}_\alpha) = 1-\alpha$, then the double bootstrap confidence interval for γ is $\mathcal{I}_2(\hat{\delta}_\alpha; \mathcal{X}, \mathcal{X}^*)$.

Following Davison and Hinkley (1997, p. 250), define:

$$\tilde{u}^*_b = \frac{1}{C}\sum_{c=1}^{C} \mathbb{1}\{\hat{\gamma}^{**}_{bc} \leq \hat{\gamma}\}.$$

The simulation-based approximation to $\hat{\delta}_\alpha$ is given by:

$$\tilde{\delta}_\alpha = \tilde{u}^*_{[(1-a)(B+1)]},$$

where $\tilde{u}^*_{[1]} \leq ... \leq \tilde{u}^*_{[B]}$. Finally, the simulation-based approximation of the double bootstrap percentile interval is given by:

$$\mathcal{I}^{perc}_2(\tilde{\delta}_\alpha; \mathcal{X}, \mathcal{X}^*) = (-\infty, \hat{\gamma}^*_{1-\tilde{\delta}_\alpha}).$$

Similarly, for the basic confidence interval in (13), we only have to modify $\tilde{u}_b^*$ to:

$$\tilde{u}_b^* = \frac{1}{C}\sum_{c=1}^{C} \mathbb{1}\{\hat{\gamma}_{bc}^{**} \leq 2\hat{\gamma}_b^* - \hat{\gamma}\},$$

so that the double bootstrap basic interval is given by:

$$\mathcal{I}_2^{basic}(\tilde{\delta}_\alpha; \mathcal{X}, \mathcal{X}^*) = (-\infty, \hat{\gamma} - (\hat{\gamma}^* - \hat{\gamma})_{\tilde{\delta}_\alpha}).$$

Two-sided intervals can be obtained by the set difference of two one-sided intervals, i.e. $\mathcal{J}_1(\alpha_1, \alpha_2) = \mathcal{I}_1(1-\alpha_1/2) \backslash \mathcal{I}_1(\alpha_2/2)$, leading to the following two-sided double bootstrap intervals:

$$\mathcal{J}_2^{perc}(\tilde{\delta}_{\alpha/2}, \tilde{\delta}_{1-\alpha/2}) = (\hat{\gamma}^*_{1-\tilde{\delta}_{1-\alpha/2}}, \hat{\gamma}^*_{1-\tilde{\delta}_{\alpha/2}}),$$

where $\tilde{\delta}_{\alpha/2}$ and $\tilde{\delta}_{1-\alpha/2}$ are two appropriate order statistics. It is also possible to consider a $(1-\alpha)$ two-sided percentile confidence interval via a two-sided version of $\mathcal{I}_1(\alpha)$ directly, i.e. $\mathcal{J}_1(\alpha) = (\hat{\gamma}^*_{\alpha/2}, \hat{\gamma}^*_{1-\alpha/2})$; see for instance Lee and Young (1996), although they use α to denote the coverage instead of $(1-\alpha)$.

Figure 4 shows the simulated coverage based on the double bootstrap of the two-sided percentile for one particular sample: $(\alpha_x, \beta_m) = (0, 0)$, $(\hat{\theta}_x, \hat{\beta}_m) = (0.051, -0.070)$ and $n = 50$. The double bootstrap in this case suggests using a 77% confidence level $(\tilde{\delta}_\alpha = 0.77)$ to obtain a 95% confidence interval. Given that $(\hat{\theta}_x, \hat{\beta}_m)$ is close to $(0, 0)$, it is reasonable to shorten the confidence interval.

The total number of bootstrap replications equals $B \cdot C$, where B and C are the number of first- and second-level bootstrap simulations. To ease the computational burden, we exploit the fact that $\tilde{\delta}_{\alpha/2}$ and $\tilde{\delta}_{1-\alpha/2}$ are based on quantiles in both tails of $\tilde{u}^*$; see also Nankervis (2005). Note that $\tilde{u}_b^*$ can be interpreted as a p-value. If the bootstrap distribution is centered around $\hat{\gamma}_b^*$, we expect $\tilde{u}_b^*$ to be large/small when $\hat{\gamma}_b^*$ is far to the left/right of $\hat{\gamma}$. Hence, after sorting $\hat{\gamma}_b^*$, we only carry out the double bootstrap for the $\frac{1}{2}M$ smallest and $\frac{1}{2}M$ largest values of $\hat{\gamma}_b^*$. Since only M, instead of B values of $\tilde{u}_b^*$ are determined, $\tilde{\delta}_{\alpha/2}$ and $\tilde{\delta}_{1-\alpha/2}$ are based on the $(B/M \cdot \alpha/2)$ and $(1 - B/M \cdot \alpha/2)$ quantiles of $\tilde{u}^*$. In this way, the number of bootstrap replications is reduced from $B \cdot C$ to $M \cdot C$.

5 Simulation Setup and Results

We have simulated the various bootstrap methods extensively using the julia programming language, see Bezanson et al. (2017). The chosen sample sizes $n \in \{25, 50, 100, 500\}$ are broadly relevant in various subject areas. The x-vector is drawn from a standard normal distribution, rescaled to have a sample variance of 1 and kept fixed in all simulations

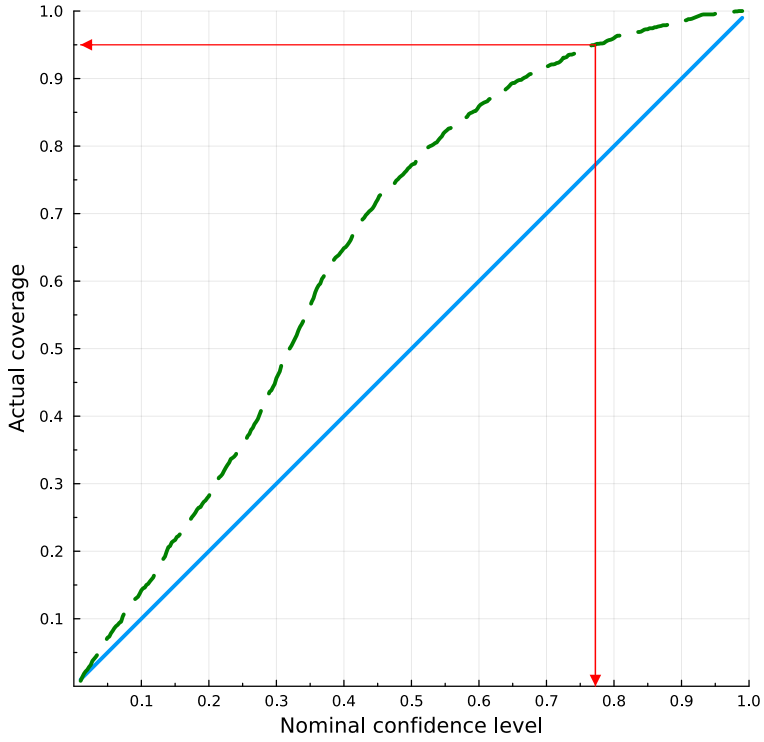

Figure 1: Estimated coverage based on the double bootstrap of the two-sided percentile confidence intervals for one particular sample: $(\alpha_x, \beta_m) = (0, 0)$, $(\hat{\theta}_x, \hat{\beta}_m) = (0.051, -0.070)$ and $n = 50$. The green dashed line shows the actual coverage obtained by the double bootstrap, while the blue solid line represents perfect coverage. A 95% confidence interval is provided by a 77% confidence level, as shown by the red lines.

since inference is conditional on x. The errors v_i and u_i are independently drawn from the standard normal distribution. For the parameters θ_x and β_m, we follow MacKinnon et al. (2004) to indicate the strength of the effects: 0.0 (none), 0.14 (small), 0.39 (medium) and 0.59 (large) and $\beta_m \geq \theta_x$. The number of Monte Carlo simulation, REP, is set to 10,000. Since the BC_a and double bootstrap intervals adjust the levels of the quantiles, the number of bootstrap replications is taken higher than the usual $1,000$; see e.g. Booth and Presnell (1998). So for each sample, the bootstrap distribution is based on $B = 1,999$ first-level bootstrap samples, while the $M = 1,000$ second-level bootstrap p-values are based on $C = 1,000$ second-level bootstrap samples; see Figure 5 for an illustration. Hence, the bootstrap p-values $\tilde{u}_b^* \in \{0.0\%, 0.1\%, ..., 99.9\%, 100.0\%\}$ are multiples of 0.1% and $\tilde{\delta}_{\alpha/2}(B+1)$ and $\tilde{\delta}_{1-\alpha/2}(B+1)$ are integers. In this way, no interpolation is needed when constructing double-bootstrap confidence intervals; see Hall et al. (2000).

We report the percentage that confidence intervals are to the left and to the right of γ and therefore do not contain the true value. We refer to them as non-coverage rejection

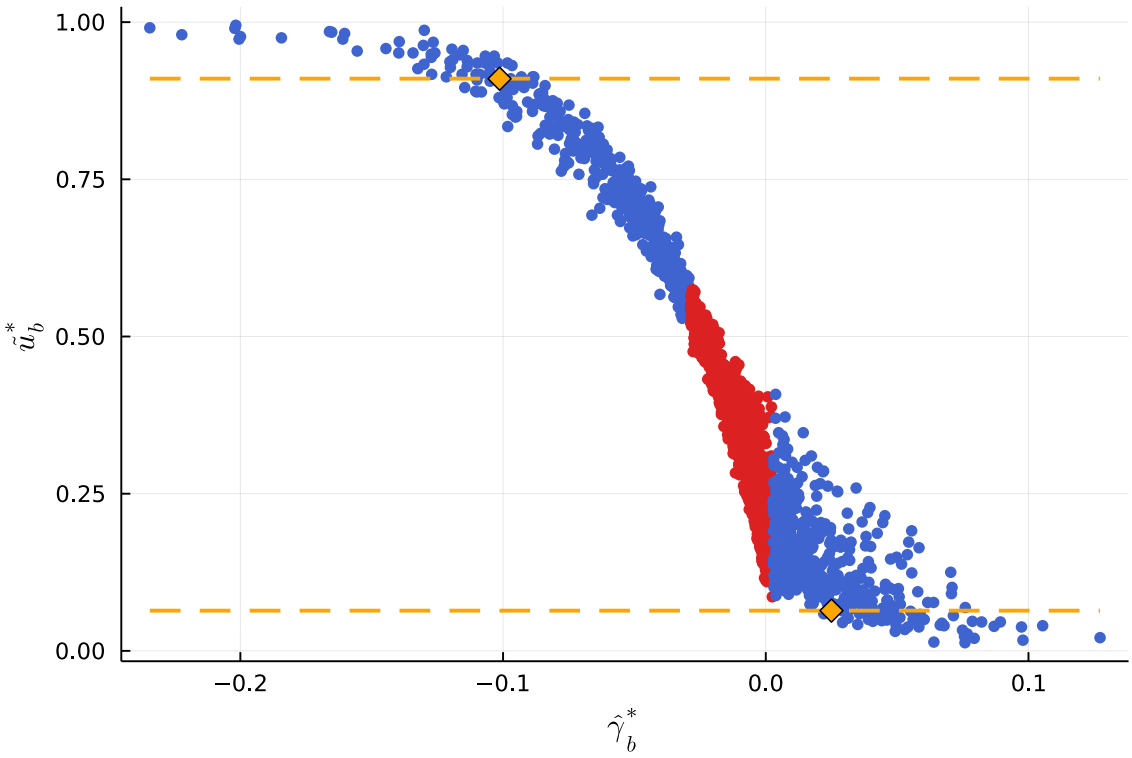

Figure 2: Scatter plot showing the p-values $\tilde{u}_b^*$ versus sorted $\hat{\gamma}_b^*$ for one particular realization of $(\hat{\theta}_x, \hat{\beta}_m)$ based on $B = 3,000$ and $M = 1,500$. The double bootstrap percentile interval can be written as $(\hat{\gamma}^*_{\tilde{\alpha}_1}, \hat{\gamma}^*_{\tilde{\alpha}_2})$, where $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ denote the 2.5% and 97.5% percentiles of $\tilde{u}_b^*$ respectively. These percentiles are indicated in the scatter plot by the two orange diamonds. Since $M/B = 1/2$, only half of the p-values would normally not be calculated in our double bootstrap procedure, corresponding to the red points that are only shown here for illustrative purposes

frequencies (ncRFs). For 95% equal-tailed confidence intervals, these percentage points should equal 2.5%, but only approximately, due to simulation error. Given that each ncRF is based on 10,000 trials, it is not significantly different from 2.5% (at the 95% confidence level) if its value is contained in the interval:

$$\left(0.025 \pm 1.96\sqrt{0.025 \cdot 0.975/10000}\right) \cdot 100\% \quad \text{i.e. } (2.194\%, 2.806\%).$$

The ncRFs for $n = 100$ are shown in Table 1, while the results for the other sample sizes can be found in tables 3 - 4 in the Appendix. An asterisk (*) after a ncRF indicates that it is significantly different from 2.5%.

We begin with the results of the percentile method. In line with earlier findings in the literature, the ncRFs of this method when $\theta_x = 0$ and β_m small are extremely low. Figure 5 illustrates this fact by showing the 10,000 estimated values of $(\hat{\alpha}_x, \hat{\beta}_m)$ in the simulation for $n = 100$ and $\theta_x = \beta_m = 0$ as dots. They are colored red if the residual

bootstrap interval based on this realization does not contain the true value $\gamma = 0$. There are 18 red realizations out of 10,000. This corresponds to a ncRF of 0.18% and far lower than the nominal 5.0%, which has serious consequences for the power of the test based on this confidence interval.

In order to explain this poor behavior, we plot, in the same figure, green lines as the boundary of an area having 95% probability of $(\hat{\alpha}_x, \hat{\beta}_m)$ lying inside. This is based on the exact distribution of $\hat{\gamma} = \hat{\theta}_x \hat{\beta}_m$, shown in Figure 5 as dashed green lines, and it is determined by numerical integration using equation (6). Quantiles for $\hat{\gamma}$ are ± 0.02236 such that:

$$\mathbb{P}\left[|\hat{\theta}_x \hat{\beta}_m| > 0.02236 \mid (\alpha_x, \beta_m, \sigma_v^2, \sigma_u^2) = (0, 0, 1, 1)\right] = 0.05.$$

The green boundary lines in Figure 5 are the restriction $|\hat{\theta}_x \hat{\beta}_m| = 0.02236$ and essentially 5% of the realizations lie outside it. For the vast majority of this 5% of $(\hat{\theta}_x, \hat{\beta}_m)$ realizations with $|\hat{\theta}_x \hat{\beta}_m| > 0.02236$, the bootstrap generates percentile confidence intervals for γ that includes 0. The red colored dots are the values for which the percentile intervals exclude the true value $\gamma = 0$. It is clear that this is nowhere near 5%: only 0.18% exclude $\gamma = 0$ and coverage is 99.82% instead of the nominal 95%. The reason is that the bootstrap distributions for $\hat{\gamma} \neq 0$ are very asymmetric and changes substantially with the estimated parameters, both in location and in their shape. When a sample is drawn and $(\hat{\theta}_x, \hat{\beta}_m)$ calculated, the bootstrap distribution approximates the pdf of $\hat{\alpha}_x \hat{\beta}_m$ with parameter values $(\hat{\theta}_x, \hat{\beta}_m, s_v^2, s_u^2)$, which differs substantially from their true values $(0, 0, 1, 1)$. This dependence on the parameters can be extreme, as seen in the asymptotic distribution of Sobel's test statistic: if $\gamma = 0$ and $(\theta_x, \beta_m) = (0, 0)$ then the asymptotic distribution of the Wald statistic for testing $\gamma = 0$ is $\frac{1}{4}\chi_1^2$, but χ_1^2 if $\gamma = 0$ and $(\theta_x, \beta_m) \neq (0, 0)$; see Glonek (1993).

Figure 5 shows three distributions: the distribution of $\hat{\gamma}$ for the true parameter values $(0, 0, 1, 1)$ as dashed green lines, the parametric bootstrap distribution of $\hat{\gamma}$ based on equation (6) with parameter values $(\hat{\theta}_x, \hat{\beta}_m, s_v^2, s_u^2)$ in purple, and very close to it, the non-parametric (residual) bootstrap distribution for that particular sample as a light blue histogram. These last two are for one particular realization $(\hat{\theta}_x, \hat{\beta}_m, s_v^2, s_u^2) = (0.2216, 0.2477, 0.9668, 1.0913)$ which is the purple star in Figure 5. When $(\theta_x, \beta_m) = (0, 0)$, the true distribution of $\hat{\gamma}$ is symmetric, but one will always estimate $(\hat{\theta}_x, \hat{\beta}_m) \neq (0, 0)$. This will lead to an asymmetric bootstrap distribution with skewness as in formula (7) with $(\theta_x, \beta_m) = (\hat{\theta}_x, \hat{\beta}_m)$.

The red lines in Figure 5 are determined such that the appropriate limit of the parametric bootstrap confidence interval equals γ (=0). For $\hat{\gamma} > 0$, these lines represent the values for $(\hat{\theta}_x, \hat{\beta}_m)$ such that the 2.5%-quantile of $\hat{\gamma}^*$ is $\gamma = 0$ based on numerical

integration of equation (6) with $(\theta_x, \beta_m, \sigma_v^2, \sigma_u^2) = (\hat{\theta}_x, \hat{\beta}_m, 1, 1)$, i.e. $(\hat{\theta}_x, \hat{\beta}_m)$ such that

$$\mathbb{P}^*[\hat{\gamma}^* \leq \gamma = 0 \mid (\hat{\alpha}_x, \hat{\beta}_m)] = 2.5\%.$$

We expect all dots outside the red lines, away from the origin, to not include the true value $\gamma = 0$, since the distribution used to determine these boundaries could be interpreted as the parametric bootstrap distribution with knowledge of the nuisance parameters (σ_v^2, σ_u^2). If the true parametric bootstrap distribution does not significantly depend on the values of the estimated nuisance parameters, we expect the bootstrap distribution based on the true nuisance values to be an accurate approximation. This is indeed the case, since most dots outside the red lines are colored red. The difference between the green and the red lines is that the green line is based on the quantiles of $\hat{\gamma}$ for the single point $(\theta_x, \beta_m) = (0, 0)$ and the red lines are based on quantiles of the parametric bootstrap distribution of $\hat{\gamma}^*$ for all possible values of $(\hat{\theta}_x, \hat{\beta}_m)$. The root of the coverage problem of the percentile method is the extreme dependence of the shape, rather than location, of the bootstrap distribution of $\hat{\gamma}^*$ on $(\hat{\theta}_x, \hat{\beta}_m)$.

We continue with the results for the percentile interval for $\gamma \neq 0$: the ncRFs are asymmetric, such that the ncRF is higher on the left of γ than on the right of γ. This asymmetry becomes less as the sample size increases. Even for $n = 100$ (500), most (half) of the ncRFs are significantly different from 2.5%. The sum of ncRFs, referred to as total ncRFs, should be around 5%, i.e. inside $(4.573\%, 5.423\%)$ based on $10,000$ simulations (with 95% confidence). We see that only for $n = 500$ that ncRFs are not significantly above 5%. Comparing the residual with the paired bootstrap, we observe that the ncRFs for the residual bootstrap are somewhat closer to 2.5% than the paired bootstrap. Hence, exploiting the correct structure as done by the residual bootstrap is noticeable, but the improvement is marginal.

Next, we discuss the results for the basic interval. The results for these intervals are worse than reported for the percentile intervals: in general the ncRFs are lower/higher for small/large values of γ and they are also more asymmetric. For instance, when $(\theta_x, \beta_m) = (.0, .59)$ and $n = 100$, the ncRFs of the basic interval are 1.0 and 1.3 compared to 2.3 and 3.0 of the percentile interval.

The BC and BC$_a$ intervals, advocated by inter alia MacKinnon et al. (2004), do not seem to perform much better than the percentile intervals, but we confirm their finding that they are liberal, i.e. coverage rates below the required 95%. When these intervals are used in testing, the overrejection is clear for θ_x small/medium and β_m medium: for $(\theta_x, \beta_m) = (0.14, 0.14)$ and $n = 100$, a ncRF larger than 8% is found for the paired bootstrap and 7.6% for the residual bootstrap. There is hardly any difference between the BC and BC$_a$ intervals, due to the estimated acceleration constants in a small interval

around 0.

Although the percentile-t intervals theoretically improve an order of magnitude upon the accuracy of the percentile intervals, the ncRFs in the simulation vary substantially with γ: e.g. for $n = 100$, when γ is small, the ncRFs are significantly smaller than 2.5% (but conservative intervals do not violate the stated 95%), for large values of γ ncRFs are close to 2.5%, but for the intermediate values $(\theta_x, \beta_m) = (0.14, 0.14)$ the total ncRFs are over 18% for the paired bootstrap and 17.7% for the residual bootstrap. These coverage rates worse than 82.3% (instead of the required 95%) disqualify the percentile-t method. No substantial difference is observed between the percentile-t based on Sobel's or Jackknife standard errors. Apparently neither one is able to appropriately standardize $(\hat{\gamma}^* - \hat{\gamma})$ and turn it into a proper pivotal root.

Finally, the double bootstrap results show that this method, in spite of the promising results in other applications reported in the literature, see for instance Shi (1992), Letson and McCullough (1998), McKnight et al. (2000), Chronopoulos et al. (2015), and Montoya and Hayes (2017), is not able to make the required adjustments. In fact, for medium values of γ, the second-level bootstrap seems to aggravate the high left ncRFs for the percentile method. The effect on the basic method appears to be even larger.

To investigate this unexpected behavior, Figure 5 shows the double-bootstrap correction as function of $(\hat{\theta}_x, \hat{\beta}_m)$ for $n = 100$. The double bootstrap percentile interval can be written as $(\hat{\gamma}^*_{\alpha_1}, \hat{\gamma}^*_{\alpha_2})$, where α_1 and α_2 denote the 2.5% and 97.5% percentiles of the p-values u^* based on the double bootstrap. For a grid of $(\hat{\theta}_x, \hat{\beta}_m)$-values, the parametric bootstrap approximation, assuming $(\sigma_v^2, \sigma_u^2) = (1, 1)$, is used to determine the values of α_1 and α_2. When $(\hat{\theta}_x, \hat{\beta}_m)$ is close to the origin, there is a substantial double bootstrap correction: α_1 is close to 15% and α_2 close to 85%. So, when $(\hat{\theta}_x, \hat{\beta}_m)$ is close to the origin, the 95% double-bootstrap percentile interval uses $(\hat{\gamma}^*_{0.15}, \hat{\gamma}^*_{0.85})$, which is much smaller than the single bootstrap percentile interval $(\hat{\gamma}^*_{0.025}, \hat{\gamma}^*_{0.975})$: the smaller the interval, the higher the probability that it excludes the true value γ leading to a severe increase in the ncRF.

This behavior can be seen in Figure 5 where values of $(\hat{\theta}_x, \hat{\beta}_m)$ that lead to bootstrap confidence intervals that exclude the true value γ are shown as red colored dots. The upper scatter plots are for $(\theta_x, \beta_m) = (0.14, 0.14)$, while the lower ones are for $(\theta_x, \beta_m) = (0.14, 0.39)$ and $n = 100$ in each case. In all scatter plots, the true value (θ_x, β_m) is represented by the blue diamond, while the blue lines represent all values $(\hat{\theta}_x, \hat{\beta}_m)$ such that their product equals this same true value $\gamma = \alpha_x \beta_m$. The intervals in the left/right scatter plots are based on single/double bootstrap. In all plots, we observe two clusters of red $(\hat{\theta}_x, \hat{\beta}_m)$-realizations away from the upper right blue line that lead to non-coverage as expected. However, for $(\theta_x, \beta_m) = (0.14, 0.14)$ there are (white) dots even further away

that do not lead to non-coverage. The reason is that $(-\theta_x, -\beta_m)$ results in the same γ value and and hence the lower left blue line. For the double bootstrap we observe far more red non-coverage points near the origin. The explanation is provided by Figure 5 which shows the large (over)correction close to the origin and leading to increased probability of excluding the true value as just described.

For comparison we also show the results for $(\theta_x, \beta_m) = (0.14, 0.39)$ when there are hardly any $(\hat{\theta}_x, \hat{\beta}_m)$-values close to the origin and therefore the double-bootstrap correction is negligible. So in conclusion, the double bootstrap is counter productive: it either overcorrects when $(\hat{\theta}_x, \hat{\beta}_m)$ is close to the origin, or hardly corrects when $(\hat{\theta}_x, \hat{\beta}_m)$ is further afield.

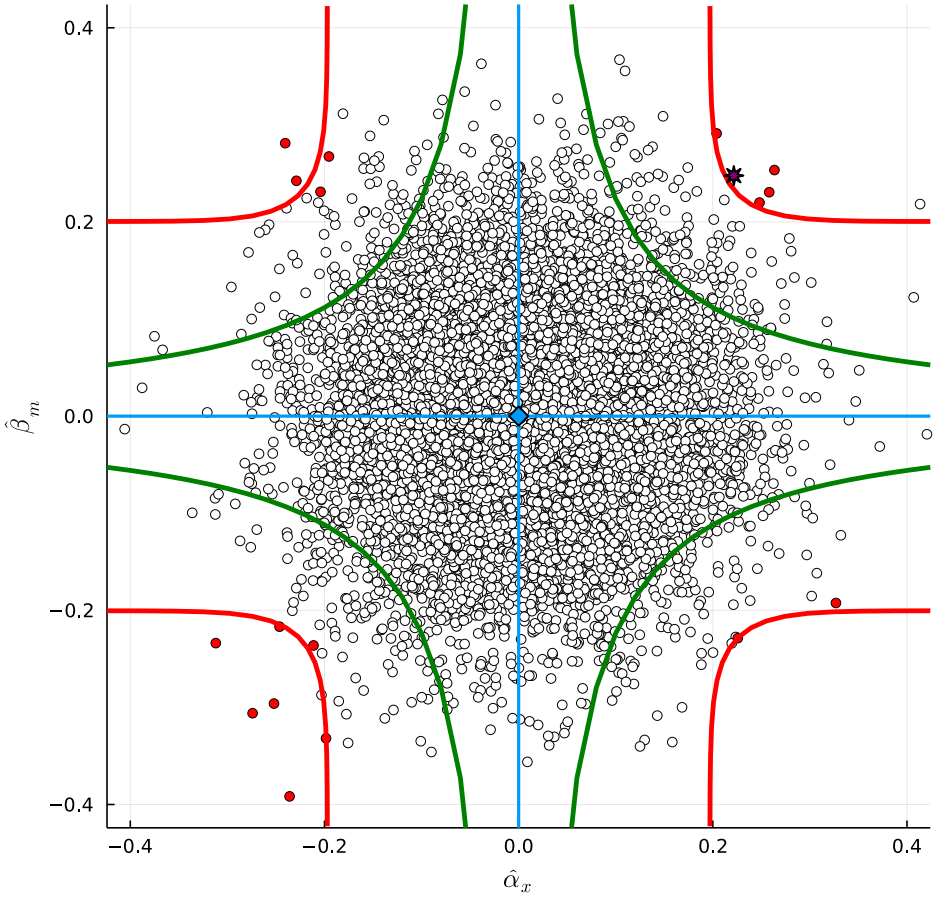

Figure 3: Scatter plot showing the 10,000 realizations $(\hat{\alpha}_x, \hat{\beta}_m)$ as dots in our simulation for $n = 100$ and $(\theta_x, \beta_m) = (0, 0)$. Intervals based on white-colored dots contain the true value $\gamma = 0$, while intervals based on red-colored dots do not include $\gamma = 0$. The true value (θ_x, β_m) is represented by the blue diamond at the origin, while the blue lines represent the true value $\gamma = 0$. The four green lines represents the 2.5% and 97.5% percentile of the distribution of $\hat{\gamma}$ assuming $(\sigma_v^2, \sigma_u^2) = (1, 1)$: ± 0.02236. A perfect ncRF of 95% would result if all dots outside the green lines were red, but the majority are white leading a severe conservative non-coverage. The boundary for red colored dots can be accurately approximated by the parametric bootstrap assuming $(\sigma_v^2, \sigma_u^2) = (1, 1)$: the red lines are determined such that the parametric bootstrap confidence interval does not include 0.

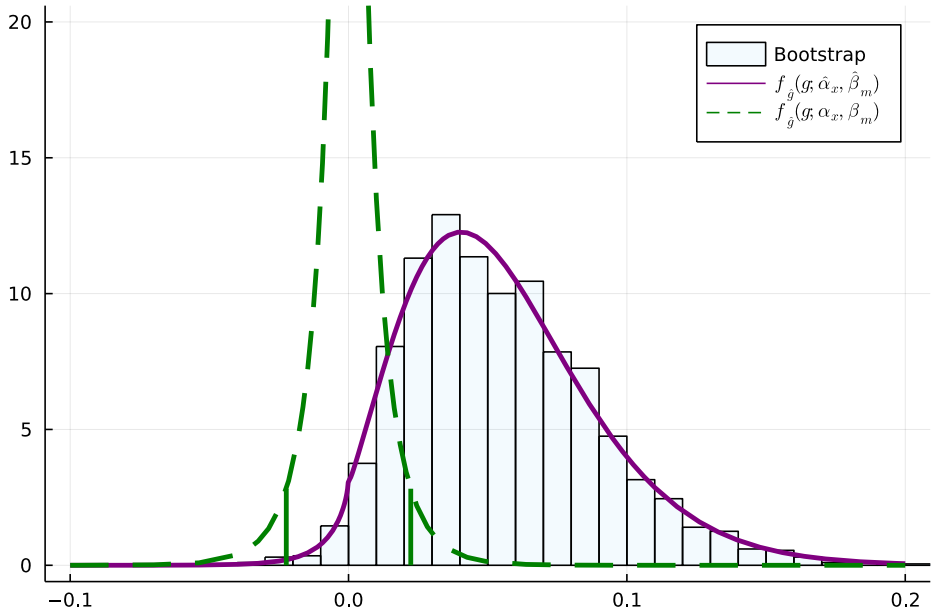

Figure 4: Histogram showing the non-parametric bootstrap distribution of $\hat{\gamma}^*$ for the realization of $(\hat{\alpha}_x, \hat{\beta}_m)$ indicated by the purple star in Figure 2. The non-parametric bootstrap distribution is accurately approximated by the purple density showing the parametric bootstrap distribution of $\hat{\gamma}^*$ assuming $(\sigma_v^2, \sigma_u^2) = (1, 1)$. Note that the area under the non-parametric bootstrap density to the left of zero is slightly smaller than 2.5% since the lower confidence limit is just above 0, i.e. the interval does not include 0. The dashed green lines represents the finite-sample distribution of $\hat{\gamma}$ based on the true values of $(\theta_x, \beta_m, \sigma_v^2, \sigma_u^2) = (0, 0, 1, 1)$. The solid green lines in this figure represent the quantiles corresponding to the green solid lines shown in Figure 5

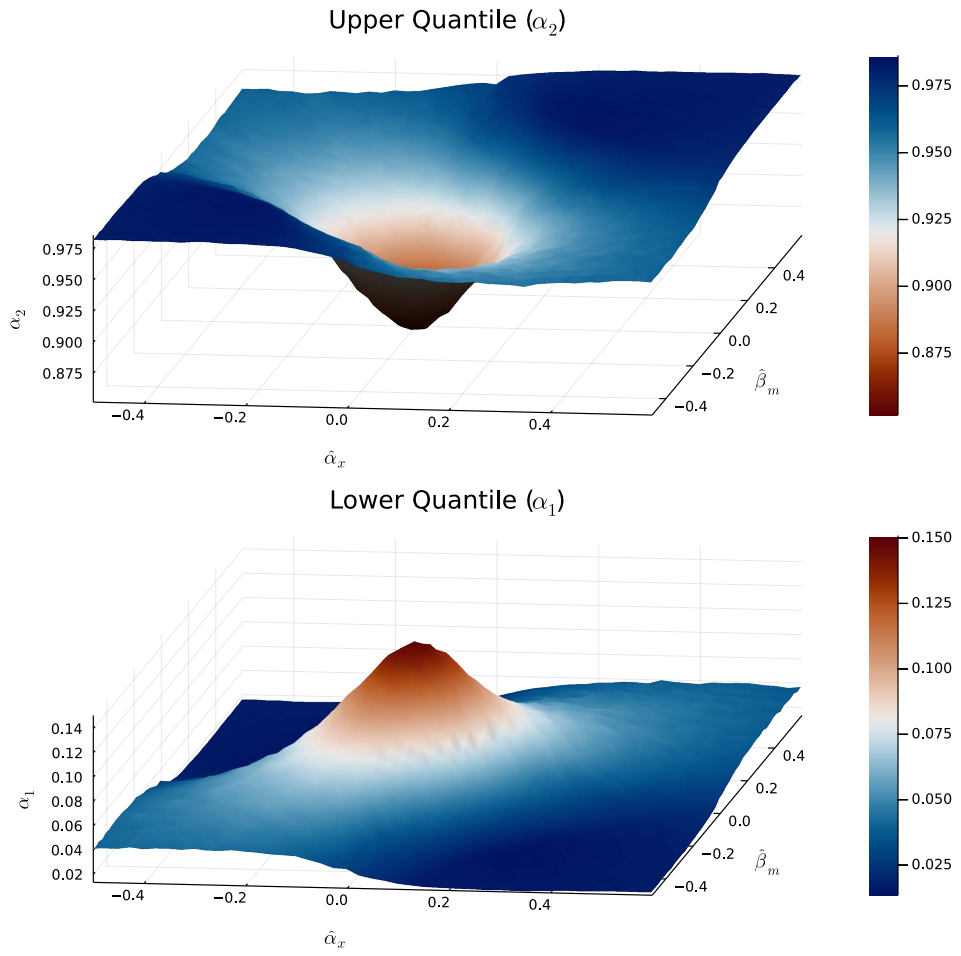

Figure 5: the double-bootstrap correction as function of $(\hat{\alpha}_x, \hat{\beta}_m)$ for $n = 100$.

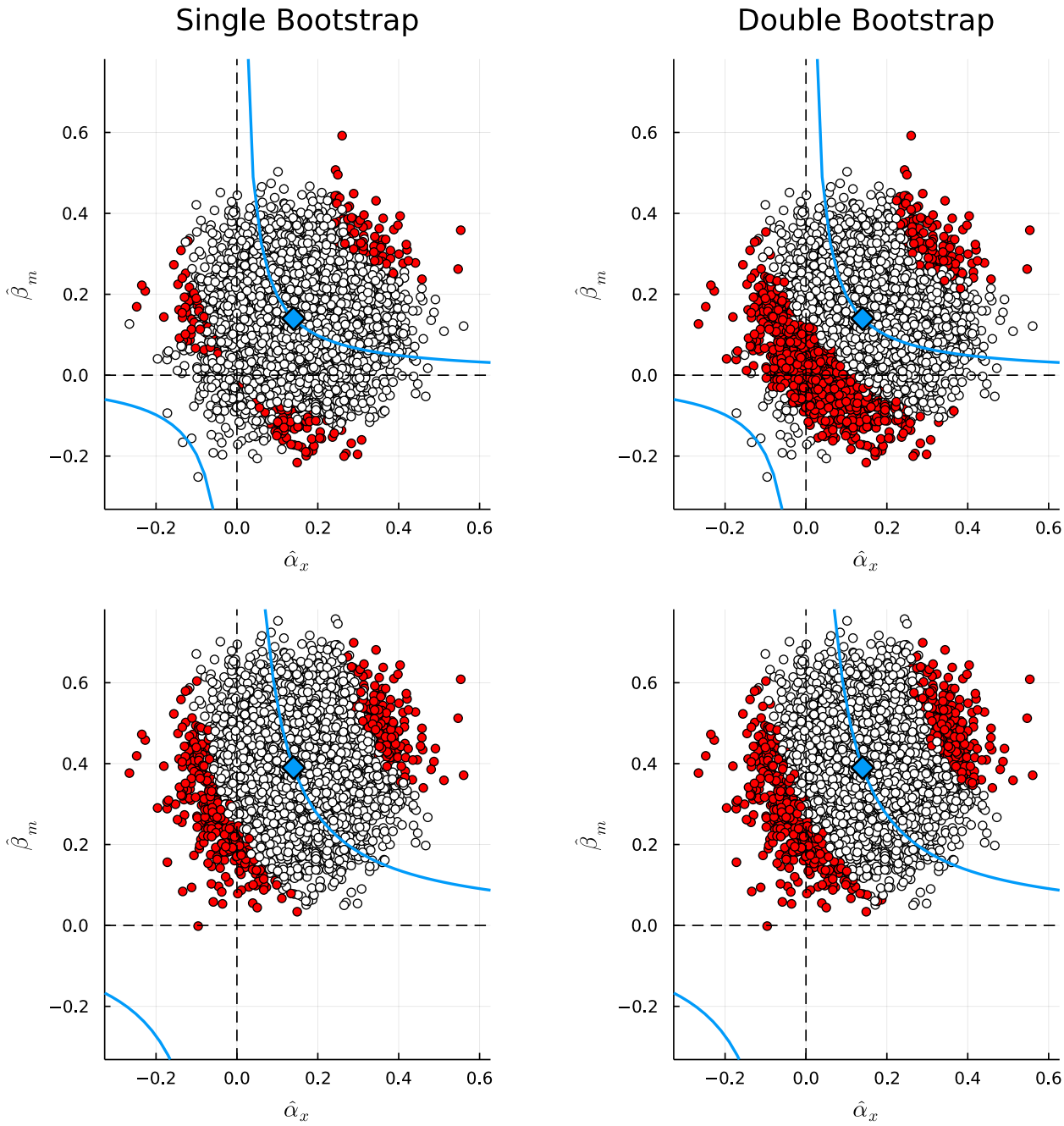

Figure 6: Scatter plot showing the 10,000 realizations $(\hat{\alpha}_x, \hat{\beta}_m)$ as dots in our simulation for $n = 100$. Confidence intervals based on red dots do not include γ. The upper scatter plots are for $(\theta_x, \beta_m) = (0.14, 0.14)$, while the lower ones are for $(\theta_x, \beta_m) = (0.14, 0.39)$. The true value (θ_x, β_m) is represented by the blue diamond, while blue lines represent the true value $\gamma = \theta_x \beta_m$. The intervals in the left/right scatter plots are based on single/double bootstrap.

Table 1: Percentage points (non-coverage frequencies×100%) that the 95% confidence interval is to the left or right of of true value $\gamma = \theta_x \beta_m$ for $n = 100$.

	Basic		Percentile		BC		BCa		Percentile-t		Perc.-t Jack		Basic-d		Percentile-d	
(θ_x, β_m)	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Residual																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.3∗	0.3∗	0.3∗	0.4∗	0.2∗	0.2∗	0.2∗	0.2∗	0.1∗	0.2∗	0.3∗	0.4∗
(.0,.14)	0.0∗	0.0∗	0.3∗	0.6∗	1.1∗	1.5∗	1.0∗	1.5∗	0.7∗	1.1∗	0.8∗	1.2∗	0.5∗	0.8∗	1.2∗	1.8∗
(.0,.39)	0.3∗	0.4∗	2.0∗	2.8∗	3.3∗	4.1∗	3.3∗	4.0∗	3.3∗	4.1∗	3.3∗	4.0∗	2.8∗	3.8∗	3.6∗	4.3∗
(.0,.59)	1.0∗	1.3∗	2.3	3.0∗	3.1∗	3.8∗	3.0∗	3.7∗	3.3∗	4.1∗	3.3∗	4.1∗	3.2∗	4.1∗	3.2∗	3.8∗
(.14,.14)	0.0∗	0.1∗	1.6∗	0.9∗	6.2∗	1.4∗	6.0∗	1.4∗	16.4∗	1.3∗	16.3∗	1.4∗	19.5∗	1.0∗	13.2∗	1.5∗
(.14,.39)	3.4∗	0.3∗	3.5∗	1.8∗	3.6∗	2.6	3.5∗	2.5	6.4∗	2.5	6.2∗	2.5	8.5∗	2.2	4.6∗	2.7
(.14,.59)	2.7	0.8∗	3.0∗	2.4	3.2∗	3.2∗	3.1∗	3.2∗	4.3∗	3.1∗	4.2∗	3.2∗	5.1∗	3.0∗	3.1∗	3.1∗
(.39,.39)	7.7∗	0.5∗	3.4∗	1.6∗	2.2∗	2.2∗	2.3	2.2∗	2.1∗	2.1∗	2.2	2.3	4.9∗	2.0∗	2.3	2.3
(.39,.59)	6.1∗	0.8∗	3.2∗	2.1∗	2.3	2.5	2.3	2.6	2.4	2.5	2.4	2.7	3.4∗	2.4	2.3	2.6
(.59,.59)	5.7∗	1.0∗	3.0∗	1.9∗	2.4	2.4	2.4	2.4	2.2	2.3	2.3	2.6	2.4	2.2∗	2.4	2.5
Paired																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.3∗	0.4∗	0.4∗	0.4∗	0.2∗	0.3∗	0.2∗	0.2∗	0.2∗	0.2∗	0.4∗	0.5∗
(.0,.14)	0.0∗	0.0∗	0.5∗	0.7∗	1.2∗	1.7∗	1.3∗	1.7∗	0.9∗	1.4∗	0.8∗	1.3∗	0.6∗	1.0∗	1.3∗	1.8∗
(.0,.39)	0.4∗	0.7∗	2.2	3.0∗	3.5∗	4.4∗	3.5∗	4.4∗	3.8∗	4.5∗	3.4∗	4.0∗	3.2∗	3.7∗	3.6∗	4.4∗
(.0,.59)	1.2∗	1.6∗	2.5	3.2∗	3.2∗	3.9∗	3.2∗	4.0∗	3.8∗	4.3∗	3.4∗	4.0∗	3.4∗	4.0∗	3.1∗	3.9∗
(.14,.14)	0.7∗	0.1∗	2.8	0.9∗	6.5∗	1.7∗	6.5∗	1.6∗	16.6∗	1.6∗	16.2∗	1.5∗	18.4∗	1.3∗	12.7∗	1.8∗
(.14,.39)	3.9∗	0.6∗	3.7∗	2.0∗	3.7∗	2.8∗	3.7∗	2.8	6.6∗	2.9∗	6.3∗	2.6	8.2∗	2.5	4.5∗	2.8∗
(.14,.59)	3.1∗	1.2∗	3.2∗	2.7	3.4∗	3.5∗	3.3∗	3.5∗	4.5∗	3.6∗	4.3∗	3.3∗	5.1∗	3.1∗	3.0∗	3.3∗
(.39,.39)	7.8∗	0.8∗	4.0∗	2.0∗	2.6	2.6	2.6	2.5	2.5	2.6	2.5	2.4	4.9∗	2.2	2.4	2.5
(.39,.59)	6.3∗	1.1∗	3.5∗	2.5	2.6	3.1∗	2.8	2.9∗	2.7	3.0∗	2.5	2.7	3.5∗	2.6	2.5	2.8
(.59,.59)	6.1∗	1.3∗	3.7∗	2.3	2.8	2.8	2.7	2.7	2.6	2.8∗	2.5	2.6	2.9∗	2.5	2.5	2.6

6 Conclusions

Reliable inference on mediation effects is crucial in empirical research and non-trivial from a statistical perspective. The mediation effect γ can be estimated as the product of two estimators $\hat{\theta}_x$ and $\hat{\beta}_m$, with a distribution that is highly dependent on the values of the two parameters θ_x and β_m, in particular when one of them is 0. Currently the preferred method of inference is based on resampling joint observations, i.e. the paired bootstrap. It is well known that this method has various problems. Confidence intervals are very conservative when both θ_x and β_m are small, i.e. their coverage probability is much larger than the nominal 95% coverage. This implies that the intervals are (much) too large. In terms of hypothesis testing, this leads to low power, implying that it is harder to establish statistical significance of mediation effects. It should be noted that, even in the idealized parametric setting, mediation tests such as the Likelihood Ratio and Wald (Sobel) test suffer from extremely low power when the mediation effects are small.

There are many alternatives to the paired bootstrap that we include and investigate in this paper. Some have been investigated previously, but we refine some of the results by detailing different constellations of α_x and β_m. In certain cases this leads to opposite conclusions from aggregate analysis where various values combined. In particular when $\gamma = 0$, e.g. when $\theta_x = 0$, the value of β_m has a very large impact on the various distributions (see section 2), simulation results, and even which method is preferred. For instance, although on the whole, bias correcting the bootstrap may appear a good idea, for medium values of β_x this actually renders confidence intervals invalid since the coverage is smaller than the stated level. Even smaller than would be acceptable under the Bradley (1978) liberal robustness criterion that is sometimes employed.

In this paper, the main question addressed is whether the double bootstrap is able to solve the problem of conservative coverage of the single bootstrap. This iterated bootstrap seems a logical solution, but had not yet been investigated. We show that it overcorrects when parameter estimates are small, which can lead to undercoverage, and for large estimates, hardly corrects at all. Hence it provides no solution. This holds for both the paired- and the residual bootstrap. In a single bootstrap setting the residual bootstrap performs slightly better than the paired bootstrap. The explanation we offer is that it exploits the structure of the model, but that makes it susceptible to misspecification, whereas the paired bootstrap is robust against e.g. heteroskedasticity; see Shao and Tu (1996).

To analyze and explain different simulation results, the finite-sample distribution of $\hat{\gamma} = \theta_x \beta_m$ is derived assuming normality of the errors. This distribution is used as a benchmark and is very useful in explaining various findings in the simulation results.

The result that $\hat{\beta}_m$ conditional on x has a student-t distribution in the simple mediation model is new to the literature.

The research turns out to be a cautionary tale about the appropriate choice of bootstrap to use. The simulations results suggest that only the percentile method based on the single bootstrap is able to control the coverage probability. The bias correction methods seem to introduce unnecessary randomness leading to conservative coverage probability for moderate non-zero value of γ. Comparing the residual to the paired bootstrap, there is not much difference between the two. The double-bootstrap correction seems to be large, but unfortunately in the wrong direction. Based on graphical methods, it appears that this correction is substantial when $\hat{\theta}_x$ and $\hat{\beta}_m$ are small. However, small values of $\hat{\theta}_x$ and $\hat{\beta}_m$ only lead to non-coverage for larger values of θ_x and β_m that do not require a correction. Stated otherwise, the double-bootstrap correction is large when it should be small and vice versa.

Overall, not every bootstrap provides the panacea that current practice seems to suggest. Moreover, none of the bootstrap methods solves the well-known problem of extreme conservative coverage that leads to extremely low rejection probabilities when testing for mediation when effects are small or inaccurately estimated. A different non-bootstrap solution for this problem is given by van Garderen and van Giersbergen (2021).

Appendix A Proofs and Further Simulation Results

This appendix first summarizes some statistical properties of estimators and the sum of squared residuals in the Gaussian linear regression model. Next, we state a lemma to find the unconditional distribution of a normal distributed variable with a stochastic variance. Finally, the remaining text contains the proofs of the propositions stated in the paper.

Statistical Properties in the Classical Gaussian Linear Regression

In the classical Gaussian linear regression model:

$$y = X\theta + \varepsilon = X_1\theta_1 + X_2\theta_2 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n),$$

where X_1 is $n \times k_1$, X_2 is $n \times k_2$, and $X = [X_1 : X_2]$ is $n \times k$, we have the results

$$\hat{\theta} \mid X \sim N(\theta, \sigma^2(X'X)^{-1}), \qquad \hat{\theta} = (X'X)^{-1}X'y, \tag{A.1}$$

$$\frac{SS_{\hat{\varepsilon}}}{\sigma^2} \equiv \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} \mid X \sim \chi^2(n-k), \qquad \hat{\varepsilon} = M_X y = M_X \varepsilon, \tag{A.2}$$

$$s^2 = \frac{1}{n-k}\hat{\varepsilon}'\hat{\varepsilon} \mid X \sim Gamma\left(\frac{n-k}{2}, \frac{2\sigma^2}{n-k}\right) \tag{A.3}$$

$$\hat{\theta}_2 \mid X \sim N(\theta_2, \sigma^2(X_2'M_{X_1}X_2)^{-1}), \qquad \hat{\theta}_2 = (X_2'M_{X_1}X_2)^{-1}X_2'M_{X_1}y \tag{A.4}$$

$$\hat{\varepsilon}'\hat{\varepsilon} \quad \text{is independent of } \hat{\beta}_m, \tag{A.5}$$

with $M_A = I - A(A'A)^{-1}A'$ the projection onto the orthogonal complement of the space spanned by A. The result in (A.4) is known as the Frisch-Waugh-Lovell (FWL) Theorem; see Davidson and MacKinnon (2004, p. 69). For result (A.5), see Davidson and MacKinnon (2004, p. 141). Note that the ML estimator of θ equals the OLS estimator $\hat{\theta}$, while the ML estimator of σ^2 equals $\hat{\sigma}^2 = n^{-1}\hat{\varepsilon}'\hat{\varepsilon}$ differs from the OLS estimator s^2.

Lemma 1 *Suppose Q has a $\chi^2(df)$-distribution and P conditional on $Q = q$ has a normal distribution with zero mean and variance df/q, then the unconditional distribution of P has a Student's $t(df)$-distribution:*

$$\begin{aligned} Q &\sim \chi^2(df), \\ P|q &\sim N\left(0, \frac{df}{q}\right), \\ P &\sim t(df). \end{aligned} \tag{A.6}$$

The result in Lemma 1 is known in the Bayesian literature as the marginal prior distribution of an unknown mean of a normal distribution when the unknown variance that is assumed to have a prior inverse χ^2-distribution, has been marginalized out; see for instance Gelman et al. (2013, p. 66). Since the proof is usually in the notation of the distributions of the sample mean and sample variance, we provide a straightforward proof using the lemma's notation below.

Proof of Lemma 1: The density of Q is given by

$$f(q) = \frac{1}{2^{df/2}\Gamma(df/2)} q^{df/2-1} \exp(-q/2),$$

while the density of $P|q$ equals

$$f(p|q) = \frac{1}{\sqrt{2\pi}\sqrt{df/q}} \exp\left(-\frac{q}{2df}p^2\right).$$

Hence, the joint density $f(p,q) = f(p|q)f(q)$ is

$$f(p,q) = \frac{1}{2^{df/2}\Gamma(df/2)\sqrt{2\pi}\sqrt{df/q}} q^{df/2-1} \exp\left(-\frac{df+p^2}{2df}q\right).$$

Employing the transformation $t = \frac{df+p^2}{2df}q$ such that the last term turns into $\exp(-t)$, leads to the following change of variable:

$$q = \frac{2df}{df+p^2}t \qquad \text{and} \qquad dq = \frac{2df}{df+p^2}\ \mathrm{d}t.$$

Using this transformation, the joint density can be rewritten as

$$f(p,t) = \frac{1}{2^{df/2}\Gamma(df/2)\sqrt{\pi}\sqrt{df+p^2}} \left(\frac{2df}{df+p^2}\right)^{df/2} t^{df/2-1/2} \exp(-t).$$

Finally, the marginal distribution is obtained by integrating t out of the joint density and using the definition of the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)\ \mathrm{d}t$, we get

$$\begin{aligned} f(p) &= \int_0^\infty f(p,t)\ \mathrm{d}t = \frac{1}{\sqrt{\pi}\sqrt{df+p^2}\Gamma(df/2)} \left(\frac{1}{1+p^2/df}\right)^{df/2} \int_0^\infty t^{(df+1)/2-1}\exp(-t)\ \mathrm{d}t \\ &= \frac{1}{\sqrt{\pi}\sqrt{df}\sqrt{(1+p^2/df)}\Gamma(df/2)} \left(\frac{1}{1+p^2/df}\right)^{df/2} \Gamma(\tfrac{df+1}{2}) \\ &= \frac{\Gamma(\frac{df+1}{2})}{\sqrt{\pi}\Gamma(\frac{df}{2})\sqrt{df}} \left(1+p^2/df\right)^{-\frac{df+1}{2}}, \end{aligned}$$

which can be identified as the density of the Student t-distribution with df degrees of freedom.

Proof of Proposition 1

Since the parameter β_x is not of interest, the first equation $y = \beta_x x + \beta_m m + u$ is considered after partialling x out. Using the FWL theorem, we have the following expression of $\hat{\beta}_m$:

$$\hat{\beta}_m = (m'M_x m)^{-1} m'M_x y.$$

The joint distribution of $(\hat{\theta}_x, SS_{\hat{v}}, \hat{\beta}_m, SS_{\hat{u}})$, with $SS_{\hat{v}} \equiv \hat{v}'\hat{v}$ and $SS_{\hat{u}} \equiv \hat{u}'\hat{u}$, can be decomposed without loss of generality into a marginal and a conditional distribution:

$$f(\hat{\theta}_x, SS_{\hat{v}}, \hat{\beta}_m, SS_{\hat{u}}|x) = f(\hat{\alpha}_x, SS_{\hat{v}}|x) f(\hat{\beta}_m, SS_{\hat{u}}|\hat{\theta}_x, SS_{\hat{v}}, x).$$

Conditional on $(\hat{\theta}_x, SS_{\hat{v}}, x)$, we have using (A.4)

$$\hat{\beta}_m|(\hat{\theta}_x, SS_{\hat{v}}, x) \sim N\left(\beta_m, \frac{\sigma_u^2}{m'M_x m} = \frac{\sigma_u^2}{SS_{\hat{v}}}\right), \tag{A.7}$$

where the equality in the variance is due to the observation that in the second equation $m = \theta_x x + v$ we have $M_x m = M_x v$, so that

$$\hat{v}'\hat{v} = (M_x m)'(M_x m) = m'M_x m$$

using (A.2). Note that the conditional distribution in (A.7) is independent of $\hat{\theta}_x$ since $\hat{\alpha}_x$ is independent of $SS_{\hat{v}}$ due to (A.5). Furthermore, $\hat{\beta}_m$ is also independent of $SS_{\hat{u}}$ again due to (A.5). Hence, the distribution of $\hat{\beta}_m$ shown in equation (A.7) only dependents on $SS_{\hat{v}}$ through the variance. Given all these independence results, the joint distribution can be written as

$$f(\hat{\theta}_x, SS_{\hat{v}}, \hat{\beta}_m, SS_{\hat{u}}|x) = f(\hat{\alpha}_x|x)f(SS_{\hat{v}}|x)f(\hat{\beta}_m|SS_{\hat{v}}, x)f(SS_{\hat{u}}).$$

Now, the joint distribution of $\hat{\theta}_x$ and $\hat{\beta}_m$ is obtained by integrating out $SS_{\hat{v}}$:

$$f(\hat{\theta}_x, SS_{\hat{v}}, \hat{\beta}_m|x) = f(\hat{\alpha}_x|x) \int_{SS_{\hat{v}}>0} f(\hat{\beta}_m|SS_{\hat{v}}, x)\ \mathrm{d}f(SS_{\hat{v}}). \tag{A.8}$$

To determine the second term in equation (A.8), note that (A.7) implies that

$$\sqrt{n-2}\frac{\sigma_v}{\sigma_u}(\hat{\beta}_m - \beta_m)|(SS_{\hat{v}}, x) \sim N\left(0, \frac{n-2}{SS_{\hat{v}}/\sigma_v^2}\right),$$

with $SS_{\hat{v}}/\sigma_v^2|x \sim \chi^2(n-2)$ due to (A.2). Using Lemma 1 with $Q = SS_{\hat{v}}/\sigma_v^2$ and $df = n-2$, we find that $P = \sqrt{n-2}\frac{\sigma_v}{\sigma_u}(\hat{\beta}_m - \beta_m)|(SS_{\hat{v}}, x)$ has a $t(n-2)$-distribution. Such a distribution is also known as a non-standardized Student's t-distribution with location parameter β_m and scale parameter $\sigma_u/(\sigma_v\sqrt{n-2})$; see e.g. Jackman (2009, Def. B.37). The joint distribution of $(\hat{\theta_x}, \hat{\beta}_m)$ conditional on only x can therefore be written as

$$\hat{\theta}_x, \hat{\beta}_m|x \sim f_N\left(\theta_x, \frac{\sigma_v^2}{x'x}\right) f_{t(n-2)}\left(\beta_m, \frac{\sigma_u^2}{\sigma_v^2(n-2)}\right), \tag{A.9}$$

where $f_N(\mu, \sigma^2)$ denotes the density of $N(\mu, \sigma^2)$ and $f_{t(\nu)}(\mu, \sigma^2)$ the density of a t-distribution with location and scale parameters μ and σ^2 and ν degrees of freedom. From (A.9) we see that the joint distribution of $(\hat{\theta}_x, \hat{\beta}_m)$ factorizes into a product of two distributions, so that $\hat{\theta}_x$ is independent of $\hat{\beta}_m$.

Proof of Proposition 2

Using the simple mediation model and classical OLS results, we can write

$$\hat{\theta}_x = \theta_x + (x'x)^{-1}x'v \tag{A.10}$$

$$\begin{aligned}\hat{\beta}_m &= \beta_m + (m'M_x m)^{-1}m'M_x u \\ &= \beta_m + (v'M_x v)^{-1}v'M_x u,\end{aligned} \tag{A.11}$$

where for the last equality we have used $m = \theta_x x + v$ post-multiplied with M_x. Since $\hat{\theta}_x$ and $\hat{\beta}_m$ are independent, we have

$$\mathbb{E}[\hat{\theta}_x^p \hat{\beta}_m^q | x] = \mathbb{E}[\hat{\alpha}_x^p | x] \mathbb{E}[\hat{\beta}_m^p | x].$$

The first three non-central moments for the $N(\mu, \sigma^2)$-distribution are given by (i) μ (ii) $\sigma^2 + \mu^2$ and (iii) $3\mu\sigma^2 + \mu^3$, while for the $t(\mu, \sigma^2, \nu)$-distribution we have (i) μ for $v > 1$ (ii) $\nu/(\nu-2)\sigma^2 + \mu^2$ for $v > 2$ and (iii) $3\mu\nu/(\nu-2)\sigma^2 + \mu^3$ for $v > 3$. Using these moments, the expectation of $\hat{\gamma}$ equals

$$\mathbb{E}[\hat{\gamma}|x] = \mathbb{E}[\hat{\theta}_x \hat{\beta}_m | x] = \mathbb{E}[\hat{\theta}_x | x] \mathbb{E}[\hat{\beta}_m | x] = \theta_x \beta_m = \gamma. \tag{A.12}$$

For the variance of $\hat{\gamma}$, we use the first two non-central moments stated before:

$$\begin{aligned} Var(\hat{\gamma}|x) &= \mathbb{E}[\hat{\theta}_x^2 \hat{\beta}_m^2 | x] - \mathbb{E}[\hat{\theta}_x \hat{\beta}_m | x]^2 \\ &= (\sigma^2_{\hat{\theta}_x} + \mu^2_{\hat{\theta}_x})(\sigma^2_{\hat{\beta}_m} + \mu^2_{\hat{\beta}_m}) - \mu^2_{\hat{\alpha}_x} \mu^2_{\hat{\beta}_m} \\ &= \left(\frac{\sigma_v^2}{x'x} + \theta_x^2 \right) \left(\frac{n-2}{n-4} \frac{1}{n-2} \frac{\sigma_u^2}{\sigma_v^2} + \beta_m^2 \right) - \theta_x^2 \beta_m^2 \\ &= \theta_x^2 \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{n-4} + \beta_m^2 \frac{\sigma_v^2}{x'x} + \frac{\sigma_u^2}{x'x} \frac{1}{n-4}, \end{aligned} \tag{A.13}$$

which can be formulated as $\theta_x^2 Var(\hat{\beta}_m|x) + \beta_m^2 Var(\hat{\theta_x}|x) + Var(\hat{\theta}_x|x) Var(\hat{\beta}_m|x)$. Note that $\theta_x^2 Var(\hat{\beta}_m|m, x) + \beta_m^2 Var(\hat{\theta}_x|x)$ is the well-known variance used in the Sobel test.

Finally, the skewness of $\hat{\gamma}$ is derived. Since the skewness is the third standardized moment, we first rewrite the third central moment as:

$$\begin{aligned} \mathbb{E}[(\hat{\gamma} - \mathbb{E}[\hat{\gamma}|x])^3 | x] &= \mathbb{E}[\hat{\gamma}^3 | x] - 3\mathbb{E}[\hat{\gamma}|x] Var(\hat{\gamma}|x) - \mathbb{E}[\hat{\gamma}|x]^3 \\ &= \mathbb{E}[\hat{\theta}_x^3 \hat{\beta}_m^3 | x] - 3\theta_x \beta_m Var(\hat{\gamma}|x) - (\theta_x \beta_m)^3. \end{aligned} \tag{A.14}$$

The third non-central moments is given by:

$$\begin{aligned} \mathbb{E}[\hat{\theta}_x^3 \hat{\beta}_m^3 | x] &= \left(3\theta_x \frac{\sigma_v^2}{x'x} + \theta_x^3 \right) \left(3\beta_m \frac{n-2}{n-4} \frac{1}{n-2} \frac{\sigma_u^2}{\sigma_v^2} + \beta_m^3 \right) \\ &= 9\theta_x \beta_m \frac{\sigma_u^2}{x'x} \frac{1}{n-4} + 3\theta_x \beta_m^3 \frac{\sigma_v^2}{x'x} + 3\alpha_x^3 \frac{1}{n-4} \frac{\sigma_u^2}{\sigma_v^2} + \alpha_x^3 \beta_m^3. \end{aligned} \tag{A.15}$$

Substitution of (A.13) and (A.15) into (A.14), and simplifying gives

$$\mathbb{E}[(\hat{\gamma} - \mathbb{E}[\hat{\gamma}|x])^3 | x] = \frac{6\alpha_x \beta_m \sigma_u^2}{(n-4)x'x}. \tag{A.16}$$

Hence, the skewness is obtained as the ratio of (A.16) and (A.13) to the power 3/2:

$$\frac{\mathbb{E}[(\hat{\gamma} - \mathbb{E}[\hat{\gamma}|x])^3 | x]}{Var(\hat{\gamma}|x)^{3/2}} = \frac{6\theta_x \beta_m \sigma_u^2}{(n-4)x'x} / \left(\theta_x^2 \frac{\sigma_u^2}{\sigma_v^2} \frac{1}{n-4} + \beta_m^2 \frac{\sigma_v^2}{x'x} + \frac{\sigma_u^2}{x'x} \frac{1}{n-4} \right)^{3/2}.$$

Proof of Proposition 3

Equation (3) implies that

$$\hat{\theta}_x - \theta_x = (x'x)^{-1}x'v|x \sim N\left(\theta_x, \sigma_v^2(x'x)^{-1}\right).$$

Hence, the t-statistic for θ_x under the null conditional on x is distributed as:

$$\begin{aligned} t_{\theta_x} &= \frac{\hat{\theta}_x - \theta_x}{\sqrt{s_v^2(x'x)^{-1}}} = \frac{(x'x)^{-1/2}x'v/\sigma_v}{\sqrt{s_v^2/\sigma_v^2}} \\ &= \frac{(x'x)^{-1/2}x'v/\sigma_v}{\sqrt{(v'M_xv/\sigma_v^2)/(n-1)}} \stackrel{d}{=} \frac{N(0,1)}{\sqrt{\chi^2/(n-1)}} \sim t_{n-1}. \end{aligned} \tag{A.17}$$

The FWL theorem shows that the estimation error for β_m can be written as

$$\hat{\beta}_m - \beta_m = (m'M_xm)^{-1}m'M_xu.$$

Note that the regression model for m implies that $M_xm = M_xv = \hat{v}$, so conditional on (x, m), or equivalently conditional on (x, v), we have

$$\hat{\beta}_m|(x,v) \sim N\left(\beta_m, \frac{\sigma_u^2}{m'M_xm} = \frac{\sigma_u^2}{SS_{\hat{v}}}\right).$$

This leads to the following distribution conditional on (x, v):

$$\begin{aligned} t_{\beta_m} &= \frac{\hat{\beta}_m - \beta_x}{\sqrt{s_u^2(m'M_xm)^{-1}}} = \frac{(m'M_xm)^{-1/2}m'M_xu/\sigma_u}{\sqrt{s_u^2/\sigma_u^2}} \\ &= \frac{(v'M_xv)^{-1/2}v'M_xu/\sigma_u}{\sqrt{(u'M_{[x:m]}u/\sigma_u^2)/(n-2)}} \stackrel{d}{=} \frac{N(0,1)}{\sqrt{\chi^2/(n-2)}} \sim t_{n-2}. \end{aligned} \tag{A.18}$$

The expression in (A.17) shows that the conditional distribution of t_{θ_x} given x is a t_{n-1}-distribution that is independent of x. This implies that the unconditional distribution of t_α is also independent of x. The conditional distribution in (A.18) given (x, u) indicates that t_{β_m} is conditionally independent of (x, u), so that the unconditional distribution of t_{β_m} is independent of (x, u) as well. Because t_{α_x} is a function of (x, u), the previous observation also implies that t_{β_m} is independent of t_α.

Table 2: Percentage points (non-coverage frequencies×100%) that the 95% confidence interval is to the left or right of of true value $\gamma = \theta_x \beta_m$ for $n = 25$.

	Basic		Percentile		BC		BCa		Percentile-t		Perc.-t Jack		Basic-d		Percentile-d	
(θ_x, β_m)	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Residual																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.4∗	0.4∗	0.4∗	0.4∗	0.2∗	0.2∗	0.3∗	0.5∗	0.1∗	0.1∗	0.4∗	0.4∗
(.0,.14)	0.0∗	0.0∗	0.2∗	0.2∗	0.6∗	0.5∗	0.6∗	0.5∗	0.4∗	0.4∗	0.6∗	0.7∗	0.3∗	0.3∗	0.6∗	0.6∗
(.0,.39)	0.0∗	0.1∗	1.3∗	0.8∗	6.7∗	1.4∗	5.8∗	1.3∗	15.1∗	1.2∗	12.8∗	1.5∗	13.2∗	0.9∗	12.1∗	1.4∗
(.0,.59)	0.2∗	0.2∗	3.3∗	1.2∗	6.9∗	2.0∗	6.6∗	1.9∗	10.4∗	1.8∗	9.3∗	2.2∗	11.0∗	1.5∗	8.8∗	2.1∗
(.14,.14)	0.0∗	0.0∗	0.1∗	0.3∗	0.8∗	0.8∗	0.7∗	0.7∗	6.3∗	0.6∗	5.8∗	0.9∗	0.2∗	0.5∗	1.0∗	0.9∗
(.14,.39)	0.0∗	0.1∗	1.3∗	0.8∗	6.7∗	1.4∗	5.8∗	1.3∗	15.1∗	1.2∗	12.8∗	1.5∗	13.2∗	0.9∗	12.1∗	1.4∗
(.14,.59)	0.2∗	0.2∗	3.3∗	1.2∗	6.9∗	2.0∗	6.6∗	1.9∗	10.4∗	1.8∗	9.3∗	2.2∗	11.0∗	1.5∗	8.8∗	2.1∗
(.39,.39)	5.6∗	0.1∗	4.3∗	1.0∗	4.0∗	1.5∗	4.1∗	1.4∗	11.4∗	1.4∗	11.3∗	1.7∗	16.9∗	1.2∗	9.7∗	1.6∗
(.39,.59)	7.4∗	0.2∗	4.2∗	1.1∗	2.9∗	1.8∗	3.2∗	1.7∗	7.2∗	1.7∗	6.9∗	1.9∗	12.8∗	1.5∗	5.9∗	1.8∗
(.59,.59)	9.1∗	0.4∗	3.7∗	1.2∗	2.1∗	1.8∗	2.3	1.7∗	3.7∗	1.7∗	3.8∗	2.0∗	9.4∗	1.6∗	3.4∗	1.8∗
Paired																
(.0,.0)	0.0∗	0.0∗	0.2∗	0.2∗	0.5∗	0.5∗	0.5∗	0.6∗	0.5∗	0.6∗	0.5∗	0.7∗	0.6∗	0.7∗	0.5∗	0.7∗
(.0,.14)	0.0∗	0.1∗	0.3∗	0.3∗	0.7∗	0.9∗	0.7∗	0.8∗	0.7∗	1.0∗	0.9∗	1.1∗	0.9∗	1.1∗	0.8∗	1.0∗
(.0,.39)	0.3∗	0.3∗	1.2∗	1.2∗	2.3	2.6	2.3	2.4	2.2	2.4	2.3	2.6	2.3	2.6	2.4	2.6
(.0,.59)	0.6∗	0.8∗	2.1∗	2.2	3.8∗	4.0∗	3.5∗	3.7∗	3.6∗	3.9∗	3.4∗	3.7∗	3.5∗	3.8∗	3.7∗	4.1∗
(.14,.14)	0.0∗	0.1∗	0.3∗	0.5∗	1.4∗	1.0∗	1.3∗	1.0∗	8.8∗	1.2∗	6.9∗	1.2∗	3.9∗	1.3∗	2.3	1.1∗
(.14,.39)	0.6∗	0.4∗	2.7	1.0∗	7.7∗	1.8∗	6.7∗	1.8∗	15.0∗	2.0∗	12.9∗	1.8∗	11.8∗	2.1∗	10.9∗	1.9∗
(.14,.59)	1.5∗	0.7∗	4.4∗	1.6∗	7.4∗	2.7	7.0∗	2.6	10.9∗	2.7	9.4∗	2.6	9.9∗	2.7	8.5∗	2.6
(.39,.39)	7.3∗	0.5∗	5.9∗	1.3∗	4.7∗	2.1∗	4.8∗	2.0∗	11.5∗	2.2	10.7∗	2.0∗	14.5∗	2.3	9.0∗	2.1∗
(.39,.59)	8.4∗	0.8∗	5.6∗	1.5∗	3.6∗	2.3	3.7∗	2.2∗	7.7∗	2.5	7.0∗	2.3	10.9∗	2.5	5.6∗	2.2
(.59,.59)	9.6∗	0.9∗	5.1∗	1.8∗	3.0∗	2.5	3.1∗	2.4	4.5∗	2.7	4.3∗	2.4	8.1∗	2.6	3.4∗	2.3

Table 3: Percentage points (non-coverage frequencies×100%) that the 95% confidence interval is to the left or right of of true value $\gamma = \theta_x \beta_m$ for $n = 50$.

	Basic		Percentile		BC		BCa		Percentile-t		Perc.-t Jack		Basic-d		Percentile-d	
(θ_x, β_m)	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Residual																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.3∗	0.2∗	0.3∗	0.2∗	0.2∗	0.1∗	0.3∗	0.3∗	0.1∗	0.1∗	0.3∗	0.3∗
(.0,.14)	0.0∗	0.0∗	0.3∗	0.3∗	0.7∗	0.7∗	0.7∗	0.7∗	0.5∗	0.5∗	0.6∗	0.6∗	0.3∗	0.3∗	0.8∗	0.8∗
(.0,.39)	0.1∗	0.1∗	1.8∗	1.7∗	3.3∗	3.1∗	3.2∗	3.0∗	2.7	2.7	2.9∗	2.8	2.2	2.1∗	3.4∗	3.5∗
(.0,.59)	0.4∗	0.5∗	2.7	2.6	4.0∗	3.8∗	4.0∗	3.6∗	3.8∗	3.7∗	3.9∗	3.6∗	3.6∗	3.4∗	4.3∗	4.0∗
(.14,.14)	0.0∗	0.0∗	0.3∗	0.5∗	3.3∗	0.9∗	3.0∗	1.0∗	17.5∗	0.8∗	15.7∗	0.9∗	11.9∗	0.6∗	10.7∗	1.0∗
(.14,.39)	0.9∗	0.1∗	3.9∗	1.1∗	6.0∗	1.9∗	5.9∗	1.9∗	10.8∗	1.7∗	10.6∗	1.8∗	13.2∗	1.4∗	8.8∗	2.1∗
(.14,.59)	1.8∗	0.3∗	3.5∗	1.8∗	4.5∗	2.5	4.5∗	2.4	6.8∗	2.4	6.5∗	2.5	7.9∗	2.1∗	5.2∗	2.5
(.39,.39)	8.8∗	0.3∗	4.0∗	1.2∗	2.2∗	1.8∗	2.4	1.7∗	4.5∗	1.7∗	4.5∗	1.7∗	9.9∗	1.6∗	4.0∗	1.9∗
(.39,.59)	7.7∗	0.4∗	3.7∗	1.5∗	2.4	2.2	2.5	2.3	3.1∗	2.2∗	3.2∗	2.1∗	6.6∗	1.8∗	2.6	2.3
(.59,.59)	7.5∗	0.5∗	3.5∗	1.5∗	2.3	2.1∗	2.4	2.0∗	2.2∗	2.0∗	2.2	2.1∗	4.5∗	1.8∗	2.4	2.1∗
Paired																
(.0,.0)	0.0∗	0.0∗	0.2∗	0.1∗	0.3∗	0.4∗	0.3∗	0.4∗	0.4∗	0.4∗	0.3∗	0.4∗	0.3∗	0.3∗	0.4∗	0.4∗
(.0,.14)	0.0∗	0.0∗	0.4∗	0.3∗	0.8∗	0.9∗	0.7∗	0.9∗	0.8∗	0.8∗	0.7∗	0.8∗	0.6∗	0.6∗	0.9∗	0.9∗
(.0,.39)	0.4∗	0.3∗	2.1∗	2.1∗	3.5∗	3.6∗	3.5∗	3.5∗	3.5∗	3.4∗	3.2∗	3.1∗	2.9∗	2.8	3.5∗	3.6∗
(.0,.59)	1.1∗	0.9∗	3.1∗	3.0∗	4.4∗	4.2∗	4.3∗	4.2∗	4.5∗	4.6∗	4.2∗	3.9∗	4.2∗	3.8∗	4.3∗	4.1∗
(.14,.14)	0.0∗	0.1∗	0.8∗	0.6∗	4.6∗	1.1∗	4.2∗	1.1∗	18.1∗	1.2∗	15.8∗	1.1∗	12.0∗	0.9∗	10.5∗	1.1∗
(.14,.39)	2.3	0.4∗	4.5∗	1.3∗	6.3∗	2.2∗	6.3∗	2.2∗	11.3∗	2.4	10.6∗	2.0∗	12.3∗	1.9∗	8.7∗	2.2
(.14,.59)	2.6	0.7∗	4.1∗	2.1∗	4.8∗	2.9∗	4.9∗	2.8∗	7.3∗	3.0∗	6.6∗	2.7	7.7∗	2.6	5.2∗	2.8
(.39,.39)	9.3∗	0.6∗	4.8∗	1.5∗	2.7	2.1∗	3.0∗	2.1∗	4.9∗	2.2	4.8∗	1.9∗	9.2∗	1.9∗	4.0∗	2.0∗
(.39,.59)	8.0∗	0.8∗	4.2∗	1.8∗	3.0∗	2.5	3.1∗	2.5	3.7∗	2.6	3.5∗	2.2	6.3∗	2.3	2.8	2.3
(.59,.59)	7.8∗	0.9∗	4.3∗	1.9∗	2.8	2.5	3.0∗	2.4	2.7	2.6	2.6	2.3	4.8∗	2.3	2.5	2.3

Table 4: Percentage points (non-coverage frequencies×100%) that the 95% confidence interval is to the left or right of of true value $\gamma = \theta_x \beta_m$ for $n = 500$.

	Basic		Percentile		BC		BCa		Percentile-t		Perc.-t Jack		Basic-d		Percentile-d	
(θ_x, β_m)	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
Residual																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.3∗	0.3∗	0.3∗	0.3∗	0.1∗	0.2∗	0.2∗	0.2∗	0.1∗	0.2∗	0.3∗	0.4∗
(.0,.14)	0.1∗	0.1∗	1.9∗	1.9∗	3.5∗	3.4∗	3.4∗	3.4∗	3.0∗	3.2∗	3.1∗	3.1∗	2.5	2.7	3.7∗	3.8∗
(.0,.39)	1.8∗	1.8∗	2.6	2.5	2.9∗	2.8	2.9∗	2.8∗	3.3∗	3.1∗	3.2∗	3.1∗	3.3∗	3.3∗	2.9∗	2.7
(.0,.59)	2.2∗	2.1∗	2.6	2.5	2.7	2.6	2.7	2.6	2.8∗	2.8	2.9∗	2.8	2.9∗	2.8∗	2.6	2.5
(.14,.14)	9.1∗	0.3∗	3.8∗	1.3∗	2.1∗	1.9∗	2.1∗	1.9∗	3.1∗	1.9∗	3.2∗	1.9∗	7.8∗	1.7∗	2.9∗	2.0∗
(.14,.39)	3.7∗	1.1∗	2.9∗	2.1∗	2.7	2.5	2.7	2.5	3.0∗	2.6	3.0∗	2.5	3.5∗	2.6	2.6	2.6
(.14,.59)	3.1∗	1.7∗	2.8	2.3	2.7	2.5	2.7	2.5	2.8∗	2.6	2.7	2.6	3.0∗	2.6	2.5	2.5
(.39,.39)	4.8∗	1.2∗	3.0∗	2.0∗	2.5	2.3	2.5	2.3	2.4	2.4	2.4	2.4	2.3	2.3	2.5	2.4
(.39,.59)	4.0∗	1.4∗	2.8∗	2.1∗	2.5	2.5	2.5	2.5	2.6	2.6	2.5	2.5	2.6	2.5	2.5	2.4
(.59,.59)	4.0∗	1.5∗	2.8∗	2.2	2.5	2.4	2.5	2.4	2.5	2.5	2.5	2.5	2.4	2.4	2.5	2.5
Paired																
(.0,.0)	0.0∗	0.0∗	0.1∗	0.1∗	0.3∗	0.3∗	0.3∗	0.3∗	0.2∗	0.2∗	0.2∗	0.2∗	0.1∗	0.1∗	0.3∗	0.4∗
(.0,.14)	0.2∗	0.2∗	1.9∗	2.0∗	3.4∗	3.4∗	3.4∗	3.4∗	3.2∗	3.1∗	3.0∗	3.0∗	2.5	2.5	3.7∗	3.5∗
(.0,.39)	1.8∗	1.7∗	2.7	2.5	2.9∗	2.8∗	2.9∗	2.8	3.3∗	3.1∗	3.2∗	3.0∗	3.3∗	3.1∗	2.8∗	2.7
(.0,.59)	2.2	2.2∗	2.7	2.5	2.8∗	2.6	2.8∗	2.6	2.9∗	2.8∗	2.9∗	2.7	2.9∗	2.8	2.5	2.5
(.14,.14)	9.0∗	0.3∗	3.9∗	1.3∗	2.2∗	2.1∗	2.2∗	2.1∗	3.2∗	1.9∗	3.3∗	1.8∗	7.8∗	1.6∗	2.8∗	2.0∗
(.14,.39)	3.8∗	1.2∗	3.0∗	2.1∗	2.8	2.6	2.8	2.6	3.0∗	2.5	3.0∗	2.4	3.4∗	2.3	2.5	2.4
(.14,.59)	3.1∗	1.7∗	2.7	2.3	2.6	2.6	2.6	2.6	2.8	2.6	2.8	2.5	3.0∗	2.5	2.5	2.5
(.39,.39)	4.9∗	1.1∗	3.2∗	2.1∗	2.6	2.5	2.6	2.5	2.5	2.5	2.4	2.3	2.2	2.2	2.5	2.4
(.39,.59)	4.2∗	1.4∗	3.0∗	2.2∗	2.5	2.5	2.5	2.5	2.5	2.6	2.4	2.5	2.4	2.4	2.3	2.5
(.59,.59)	4.1∗	1.5∗	2.9∗	2.3	2.6	2.6	2.6	2.5	2.5	2.5	2.5	2.4	2.4	2.4	2.4	2.4

References

Aroian, L. A. (1947). The probability function of the product of two normally distributed variables. *The Annals of Mathematical Statistics 18*(2), 265–271.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology 51*(6), 1173.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017). Julia: A fresh approach to numerical computing. *SIAM review 59*(1), 65–98.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*, 1196–1217.

Bollen, K. and R. Stine (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological methodology 20*, 115–140.

Booth, J. and B. Presnell (1998). Allocation of monte carlo resources for the iterated bootstrap. *Journal of Computational and Graphical Statistics 7*(1), 92–112.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology 31*(2), 144–152.

Chronopoulos, D. K., C. Girardone, and J. C. Nankervis (2015). Double bootstrap confidence intervals in the two-stage dea approach. *Journal of Time Series Analysis 36*(5), 653–662.

Craig, C. C. (1936). On the frequency function of xy. *The Annals of Mathematical Statistics 7*(1), 1–15.

Davidson, R. and J. G. MacKinnon (2004). *Econometric theory and methods*. Oxford University Press New York.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application*. Number 1 in Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge university press.

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

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics 9*, 139–172.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association 82*, 171–200.

Fritz, M. S., A. B. Taylor, and D. P. MacKinnon (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research 47*(1), 61–87.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Glonek, G. F. V. (1993). On the behaviour of wald statistics for the disjunction of two regular hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(3), 749–755.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag.

Hall, P., S. M.-S. Lee, and G. A. Young (2000). Importance of interpolation when constructing double-bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(2), 479–491.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

Lee, S. M. and G. A. Young (1996). Sequential iterated bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 235–251.

Letson, D. and B. McCullough (1998). Better confidence intervals: The double bootstrap with no pivot. *American journal of agricultural economics 80*(3), 552–559.

MacKinnon, D. P., A. J. Fairchild, and M. S. Fritz (2007). Mediation analysis. *Annu. Rev. Psychol. 58*, 593–614.

MacKinnon, D. P., C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods 7*(1), 83.

MacKinnon, D. P., C. M. Lockwood, and J. Williams (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate behavioral research 39*(1), 99–128.

McKnight, S. D., J. W. McKean, and B. E. Huitema (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological methods 5*(1), 87.

Meeker, W. Q., L. W. Cornwell, and L. A. Aroian (1981). The product of two normally distributed random variables. In W. J. Kennedy and R. E. Odeh (Eds.), *Selected tables in mathematical statistics*, Volume VII. Providence: RI: American Mathematical Society.

Montoya, A. K. and A. F. Hayes (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods 22*(1), 6.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the theory of statistics* (3rd ed ed.). McGraw-Hill Series in Probability and Statistics. London: McGraw-Hill.

Nankervis, J. C. (2005). Computational algorithms for double bootstrap confidence intervals. *Computational statistics & data analysis 49*(2), 461–475.

Preacher, K. J., D. D. Rucker, and A. F. Hayes (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate behavioral research 42*(1), 185–227.

Rothenberg, T. J. and C. T. Leenders (1964). Efficient estimation of simultaneous equation systems. *Econometrica 32*(1–2), 57–76.

Rozeboom, W. W. (1956). Mediation variables in scientific theory. *Psychological Review 63*(4), 249.

Shao, J. and D. Tu (1996). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Shi, S. G. (1992). Accurate and efficient double-bootstrap confidence limit method. *Computational statistics & data analysis 13*(1), 21–32.

Shrout, P. E. and N. Bolger (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods 7*(4), 422.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology 13*, 290–312.

van Garderen, K. J. and N. van Giersbergen (2021). A nearly similar powerful test for mediation. Technical report, arxiv.org/abs/2012.11342v2.

Woodworth, R. (1928). Dynamic psychology in: Murchison c.(ed.), psychologies of 1925.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America 6*(6), 320.