IJSE

Rogers Yang, Nanyu Luo, Feng Ji

May 2025

1 Introduction

Statistical inference often relies on convenience assumptions, for instance a regression model with normal errors or a specific likelihood in Bayesian analysis. Model misspecification refers to the common situation where these assumptions are at least slightly wrong. For example, the true data-generating process might have heavier tails than assumed, heteroscedastic variance, omitted nonlinear effects, or other deviations from the data generating model (Huber, 1967). When the model is misspecified, standard inferential formulas (like model-based standard errors and confidence intervals) can misrepresent the true uncertainty (Knight, 2000). The model-based standard errors that are derived under the assumption that the model is correct may underestimate or miscalculate the sampling variance of estimators if those model assumptions do not hold (e.g., White, 1980; Imbens and Rubin, 2015).

This issue matters because in practice models are almost never exactly correct (Box, 1976). If we ignore model misspecification, we risk reporting confidence intervals or p-values that are not valid in reality. For example, an ordinary least squares regression assumes constant error variance and independent, normally distributed errors; if in reality the errors are heteroskedastic or nonnormal, the usual $\mathrm{SE}(\hat{\beta})$ formula from the regression output is no longer reliable. The consequences are that our inferential statements, such as "coefficient β is significantly different from 0 at 95% confidence," might be wrong more often than advertised. Thus, robust uncertainty quantification is needed to account for the discrepancy between our model and reality. Tools like robust variance estimators (Huber, 1967; White, 1980) and the bootstrap (Efron, 1982) were developed to address this gap.

Model misspecification can manifest in several distinct ways, each affecting inference. **Distributional assumption errors** arise when we posit a specific distributional form for errors (e.g. normality) that the data do not actually follow. Coefficient estimates in a regression (such as OLS) remain unbiased under these violations, but the usual standard error formulas become incorrect, which can make t-tests and confidence intervals misleading (Kutner and Neter, 2004). For example, one might fit $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ assuming $\varepsilon_i \sim N(0, \sigma^2)$, but if the residuals are heavily skewed or have longer tails (as with income data), the nominal 95% intervals will not have 95% coverage.

Another common problem is **omitted variable bias**: if an important predictor is left out of the regression, its effect is absorbed by the error term. This can bias the estimator $\hat{\beta}$ even as $n \to \infty$, because the omitted factor induces correlation between X and ε (Kutner and Neter, 2004). In such cases, robust standard errors may adjust the variance but cannot correct the bias in the coefficient estimates. For instance, regressing wage on education while omitting work experience will bias the education coefficient, since experience is related to both education and wage.

A related issue is **functional form misspecification**. If we force an incorrect functional form (say, a linear model when the true relationship is nonlinear), the result is systematic patterns in the residuals and biased, inconsistent estimates (Kutner and Neter, 2004). Continuing the previous example, if the true model is $Y = \alpha + \beta_1 X + \beta_2 X^2 + u$ (quadratic in X) but we fit a linear model $Y = \alpha + \beta_1 X + u$, the fitted line will systematically miss the curvature, leading to biased $\hat{\beta}_1$ and residuals that violate model assumptions.

Even when the mean structure is correct, distributional assumptions about variability can go wrong. **Heteroskedasticity** refers to the case when $\mathrm{Var}(\varepsilon_i)$ changes with X_i, violating the constant-variance assumption. OLS coefficient estimates remain unbiased but are no longer maximally efficient, and the usual variance formula $\widehat{\mathrm{Var}}(\hat{\beta})$ (which assumes homoscedasticity) is inconsistent. Consequently, naive standard errors are too small or too large; tests and intervals based on them will be invalid unless we use a heteroskedasticity-robust (sandwich) estimator. A typical example is a regression of house prices on size: larger homes tend to show greater variability in prices, so the residuals from a simple linear model will have increasing variance with X.

Similarly, **dependence or clustering** of observations breaks the independence assumption. If correlated observations are treated as independent, the standard error will be underestimated. Coefficient estimates $\hat{\beta}$ may still be unbiased (assuming exogenous regressors), but the inference is overly optimistic with inflated Type I error. For example, consider test scores of students where multiple students come from the same school (clusters): their outcomes share unobserved school-level effects, so a naive regression ignoring school clusters will underestimate the true uncertainty of estimated effects. Cluster-robust variance estimators are needed in this case to obtain valid inference.

In summary, there are myriad ways in which a model can be misspecified, and in all these scenarios the usual inference procedures break down. This is why robust alternatives like the sandwich variance estimator and the bootstrap have become essential tools for empirical researchers. In the classical (frequentist) framework, these methods provide a safety net: they yield standard errors that remain consistent for the true variability of $\hat{\beta}$ even if the model is wrong.

In the Bayesian context, model misspecification is also a concern. Bayesians interpret standard deviations of the posterior as a measure of uncertainty, but this measure is conditional on the model being correct. If the model is wrong, the posterior spread might not reflect how variable the estimator would be under the true data-generating process (Sriram et al., 2015; Yang and He, 2016). In other words, the Bayesian posterior variance does not necessarily coincide with

the frequentist variance of an estimator when the model is misspecified or the target parameter is non-smooth (Syring and Martin, 2019). Recent studies have highlighted the importance of evaluating the frequentist variability of Bayesian estimators under possible model misspecification. For example, Giordano and Broderick (2024) argue that the frequentist variability of Bayesian posterior expectations can provide a meaningful uncertainty measure even when the model is misspecified. This perspective treats a Bayesian analysis as a black-box estimator and asks: if we repeated the data collection and analysis many times, even with a slightly wrong model, how variable would our posterior-based estimates be? Answering this question requires methods beyond the standard posterior variance, leading us toward more robust approaches such as the infinitesimal jackknife (described below).

2 Robust Standard Errors and the Bootstrap

Two widely used approaches for obtaining more reliable standard error estimates under model misspecification are robust variance estimators and the bootstrap.

Robust standard errors. In classical statistics, the robust (sandwich) variance estimator provides valid inference for parameter estimates even when standard model assumptions (e.g. homoscedasticity or correct likelihood specification) are violated. This approach was formalized by White (1980) for linear regression models and generalized through the theory of M-estimators by Huber (1967). It is often referred to as the "sandwich" estimator.

The term "sandwich" derives from the matrix expression of this variance estimator:

$$\widehat{\mathrm{Var}}(\hat{\theta}) = A^{-1}BA^{-1},$$

where A is the estimated Hessian (information matrix) and B is the empirical outer product of the gradient contributions. A reflects the curvature of the objective function under the assumed model, while B captures the actual observed variability of the scores. By "sandwiching" B between the inverse information matrices A^{-1}, one obtains a variance estimator that remains consistent even if the likelihood is misspecified, provided the model yields consistent parameter estimates and standard regularity conditions hold.

This sandwich estimator is thus model-agnostic to first order: rather than relying entirely on parametric assumptions, it leverages the empirical distribution of the residuals to correct the naive model-based uncertainty. Robust standard errors are widely adopted—for example, in ordinary least squares with heteroskedastic errors, in generalized linear models, and in generalized estimating equations for longitudinal data (Zeger and Liang, 1986).

However, robust standard errors are not without limitations. First, they rely on large-sample approximations; in small samples, the sandwich variance estimator can be biased or unstable, sometimes even overestimating variability (MacKinnon and White, 1985). Second, robust SEs primarily address variance misspecification and are not designed to correct estimator bias. For instance,

if a regression omits a key covariate, robust SEs may inflate the variance but cannot remove the underlying bias in the coefficient estimate. Finally, implementing robust SEs can require nontrivial analytical work, such as computing gradients or Jacobians of estimating equations. In complex models, especially those defined implicitly or fit via numerical optimization, this can be infeasible or computationally expensive.

These limitations motivate the search for alternative uncertainty quantification methods that retain robustness but are easier to implement. This leads to approaches such as the bootstrap and, more recently, the infinitesimal jackknife.

Bootstrap. The bootstrap is a general and powerful resampling technique that can estimate the sampling distribution of almost any statistic without requiring an analytic formula. To get a bootstrap standard error, we repeatedly resample datasets from the observed data (with replacement, typically of the same size n), recalculate the estimator on each resample, and then take the standard deviation of those replicate estimates. The bootstrap approximates what would happen if we drew new samples from the true population by instead drawing samples from the empirical distribution (the observed data). This method is appealing because it is largely model-free and works for very complex estimators where deriving an analytic variance might be impractical. The bootstrap is often taught as a go-to method for variance estimation when one is unsure about parametric assumptions.

While very general, the bootstrap also has downsides. The most obvious is computational cost. We may need dozens, hundreds, or even thousands of resampled analyses to get a stable estimate of the variance. If each analysis is expensive, the bootstrap can be prohibitively slow. For example, in a Bayesian analysis via MCMC, running a full bootstrap might require re-running the entire MCMC for many different bootstrap datasets, which is extremely time-consuming. In large data settings or with complex machine learning models, retraining the model for each bootstrap sample can be an order of magnitude slower than a single fit. Another limitation is that the bootstrap, like any estimation method, relies on sufficient sample size; for very small n, the bootstrap distribution may not approximate the true sampling distribution well. Nonetheless, the bootstrap remains a staple for inference in many scenarios because of its flexibility and generally reliable performance as n grows.

In summary, robust estimators and the bootstrap aim to account for model misspecification or to avoid reliance on strict model assumptions. Each has trade-offs: sandwich formulas are analytical and fast but require derivations and large-sample justification, while the bootstrap is easy to apply conceptually and works for virtually any statistic but can be computationally intensive. In practice, one often seeks methods that combine the strengths of both—methods that are general and robust like the bootstrap, but faster and more analytical like the sandwich. One such method is the *infinitesimal jackknife* (IJ), which we introduce next. Notably, in Bayesian settings where re-fitting models many times is especially costly, a method like the IJ is particularly attractive.

3 The Infinitesimal Jackknife (IJ)

The infinitesimal jackknife is an analytic technique for estimating variance (and other sampling error measures) that builds on the classical jackknife and the concept of influence functions from robust statistics. To unpack this, recall the classical jackknife: given an estimator $\hat{\theta}$ computed on n observations, the jackknife method involves systematically leaving out one observation at a time and recomputing the estimate (Efron, 1982). That is, for each $i = 1, \ldots, n$, one calculates $\hat{\theta}_{(-i)}$ using all data except observation i. These leave-one-out estimates can be used to assess the estimator's variability or bias. For instance, the jackknife variance estimate is proportional to the sample variance of the $\hat{\theta}_{(-i)}$ values. The jackknife is a deterministic approximation to the bootstrap (it doesn't involve random resampling) and often yields reasonable variance estimates for smooth estimators. However, it too can be tedious for large n (requiring n re-computations of the estimator) and, for some non-smooth estimators, the jackknife variance can be inconsistent.

The infinitesimal jackknife (IJ) takes the jackknife idea to the limit of an "infinitesimally" small leave-one-out. Instead of removing one whole observation, we imagine down-weighting an observation by an infinitesimal amount and seeing how the estimate changes in response. The IJ formalizes this by considering the derivative of the estimator with respect to observation weights. This derivative is exactly what robust statistics calls the *influence function* (formally defined in the Appendix). The influence function $\mathrm{IF}(x; T, F)$ measures the impact of an infinitesimal contamination at point x on the estimator $T(F)$, which represents our target parameter as a functional of the population distribution F. Roughly speaking, if $\mathrm{IF}(x)$ is large in magnitude for some x, it means that observation has a strong influence on the estimate. The IJ method uses these influence function values for each data point to compute the variance of the estimator, instead of actually re-fitting the model multiple times.

In effect, the IJ provides an analytic approximation to what the bootstrap or jackknife would yield. It has been described as a "linear approximation" to those resampling procedures. Historically, the infinitesimal jackknife was treated as a theoretical tool to derive asymptotic variance formulas Efron (1982), but recent developments have turned it into a practical method for large-scale problems. For example, Giordano and Broderick (2024) dub it the "Swiss Army infinitesimal jackknife," showing that with modern automatic differentiation one can apply IJ ideas to a variety of machine learning models to drastically speed up cross-validation and bootstrap-type analyses. The key point is that the IJ is model-agnostic like the bootstrap (since it uses empirical derivatives rather than assuming the model form) and it avoids repeated re-fitting by leveraging a single fitted model and its local sensitivities. By computing the influence of each data point on the estimate, one can get the variance from one run of the model, as opposed to many runs for many resamples.

The IJ thus bridges between the robust sandwich estimator and the bootstrap. Like the sandwich method, it relies on analytical derivatives and thus is fast once those derivatives are obtained. Like the bootstrap, it does not assume

the model is correct—it operates even under misspecification, aiming to capture the true sampling variance of the estimator. Next, we describe how the IJ works mathematically and illustrate its use in practice.

Mathematically, the IJ can be derived as a first-order Taylor approximation to the ordinary jackknife. Suppose our estimator $\hat{\theta}$ arises as the solution to an estimating equation

$$\sum_{i=1}^{n} \psi(x_i, \hat{\theta}) = 0,$$

where $\psi(x_i, \theta)$ is the contribution of the ith observation to the estimating equations (for instance, $\psi(x_i, \theta) = x_i(y_i - x_i^\top \theta)$ for OLS). Denote

$$A = \frac{1}{n} \sum_{i=1}^{n} \partial_\theta \psi(x_i, \hat{\theta}),$$

the empirical Jacobian (derivative of the estimating equations with respect to θ), and

$$\psi_i = \psi(x_i, \hat{\theta})$$

the score contributions at the solution. Under regularity conditions, the Gateaux derivative of the functional defining $\hat{\theta}$ yields the influence function:

$$\mathrm{IF}(x_i; \hat{\theta}) = -\, A^{-1}\, \psi_i,$$

which measures the first-order effect on $\hat{\theta}$ of an infinitesimal downweighting of observation i (Huber, 1967). Intuitively, if $|IF(x_i)|$ is large, then observation i exerts a strong pull on the estimate.

Once the n influence values $\{IF(x_i)\}_{i=1}^n$ are computed, the IJ variance estimator follows by taking their empirical variance:

$$\widehat{\mathrm{Var}}_{\mathrm{IJ}}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \{\mathrm{IF}(x_i)\}^2,$$

and hence the standard error is $\widehat{\mathrm{SE}}_{\mathrm{IJ}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(x_i)^2}$. This formula is a first-order Taylor approximation to the leave-one-out jackknife variance and, under mild conditions, is asymptotically equivalent to both the full jackknife and the bootstrap (differing only at order $O(n^{-2})$) (Efron, 1982; Giordano and Broderick, 2024).

A small adjustment is needed to apply the IJ in the clustered setting. The standard IJ formulas assume independent observations; under clustering, one can derive a cluster-robust IJ by treating entire clusters as the units of resampling. In practice, this means aggregating the influence functions at the cluster level. We implement a cluster-IJ by computing

$$\mathbf{IF}_j = -A^{-1} \sum_{i \in C_j} \psi(x_{ij}, \hat{\theta}),$$

where C_j is the set of observations in cluster j, and then taking $\sum_{j=1}^{G} \mathbf{IF}_j \mathbf{IF}_j^\top$ as the variance estimator. This is analogous to how one computes cluster-robust sandwich variances by summing cluster-wise score contributions.

From a practical standpoint, the IJ is appealing in modern statistical work: once a single model fit has produced $\hat{\theta}$ along with the gradients ψ_i and the Jacobian A, no further refitting is necessary. In computational environments that support automatic differentiation, these derivatives can be obtained transparently, making the IJ a scalable and nearly turnkey solution for robust variance estimation in high-dimensional or complex models (Giordano and Broderick, 2024).

To illustrate the use of the IJ in a familiar setting, consider an ordinary least squares regression. After fitting

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2,$$

we obtain the residuals $r_i = y_i - x_i^\top \hat{\beta}$ and the design matrix $X = [x_1^\top, \ldots, x_n^\top]^\top$. The score contributions for this model are

$$\psi(x_i, \hat{\beta}) = x_i \, r_i,$$

and the empirical Hessian (Jacobian of the score equations) is

$$A = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top.$$

By inverting A, we can form the influence values for each observation:

$$\mathrm{IF}(x_i; \hat{\beta}) = -A^{-1} \, \psi(x_i, \hat{\beta}) = -A^{-1} x_i r_i.$$

The IJ variance estimator for $\hat{\beta}$ is then obtained by aggregating these influences:

$$\widehat{\mathrm{Var}}_{\mathrm{IJ}}(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(x_i; \hat{\beta}) \, \mathrm{IF}(x_i; \hat{\beta})^\top,$$

so the IJ standard error for each component of $\hat{\beta}$ is the square root of the corresponding diagonal entry of this matrix.

In more complex settings—such as generalized linear models or penalized estimators—the same recipe applies. One fits the model once, records any fitted weights or intermediate derivatives needed to construct A, and then evaluates the influence-function approximation. Modern automatic differentiation software can compute these derivatives with little effort from the user (Giordano and Broderick, 2024). For example, in a Poisson regression for count data, the weight matrix $W = \mathrm{diag}(\hat{\mu}_i)$ (where $\hat{\mu}_i$ are fitted means) arises naturally in the computation of A, and the score contributions take the form $\psi_i = (y_i - \hat{\mu}_i) x_i$. After assembling A and $\{\psi_i\}$, one again obtains the IJ variance by a simple matrix-vector multiplication and summation.

To summarize, the practical steps for using the IJ are: first, fit the model to obtain the point estimate and any auxiliary quantities (like weights or gradients); second, compute the Jacobian matrix of the estimating equations at the solution; third, form the influence values for each observation by multiplying the negative inverse Jacobian by each score contribution; finally, aggregate the influence values (e.g. by taking an average of their outer products, as in the formula above) to get the variance estimate. This procedure is both fast and robust to moderate misspecification, sharing the model-agnostic advantages of the bootstrap while avoiding its computational burden (Efron, 1982; Huber, 1967).

4 Bayesian Infinitesimal Jackknife Standard Errors

Bayesian analyses often rely on a convenient (working) likelihood to obtain posterior draws for a target functional of interest. Under model misspecification, posterior spreads need not match the frequentist sampling variability of the posterior-based estimator, so credible intervals can be miscalibrated. It is therefore useful to quantify uncertainty by the frequentist variability of $\hat{\theta} = \mathbb{E}(\theta \mid \mathcal{D})$ even when the likelihood is only approximate.

In Bayesian analysis, uncertainty is usually summarized by the posterior. Given S draws $\{\theta^{(s)}\}_{s=1}^{S}$ from an MCMC run, the posterior mean $\hat{\theta} = S^{-1}\sum_{s=1}^{S}\theta^{(s)}$ is a point estimate and the posterior covariance is

$$\widehat{\mathrm{Var}}[\theta] = \frac{1}{S-1}\sum_{s=1}^{S}\big(\theta^{(s)} - \hat{\theta}\big)\big(\theta^{(s)} - \hat{\theta}\big)^{\top}.$$

Under correct specification, this (after appropriate scaling) coincides asymptotically with the frequentist variance of $\hat{\theta}$. Under misspecification, however, the posterior variance may not reflect the true sampling variability of $\hat{\theta}$.

The IJ method provides a robust alternative for estimating the sampling variance of $\hat{\theta}$ from a single MCMC run. The idea is to study how $\hat{\theta} = \mathbb{E}(\theta \mid \mathcal{D})$ changes under small perturbations to the data contribution in the log posterior. Let $\ell_i(\theta)$ be the (working) log-likelihood contribution of observation i, and $\pi(\theta)$ the prior. Introduce weights $w = (w_1, \ldots, w_n)$ in the weighted log-posterior $\sum_{i=1}^{n} w_i \ell_i(\theta) + \log \pi(\theta)$, with $w_i \equiv 1$ at the observed data. Define $\hat{\theta}(w) = \mathbb{E}(\theta \mid \mathcal{D}, w)$. The influence of observation i on $\hat{\theta}$ is the Gateaux derivative at $w = \mathbf{1}$, which admits the posterior covariance identity

$$I_i = n \cdot \mathrm{Cov}_{\mathrm{post}}\Big(\theta,\ \nabla_{w_i} \log p(\mathcal{D} \mid \theta, w)\Big)\Big|_{w=\mathbf{1}} = n \cdot \mathrm{Cov}_{\mathrm{post}}\big(\theta,\ \ell_i(\theta)\big).$$

In practice, I_i is estimated from the MCMC output via the sample covariance between $\{\theta^{(s)}\}_{s=1}^{S}$ and $\{\ell_i(\theta^{(s)})\}_{s=1}^{S}$. Aggregating these influences yields the IJ

variance estimator

$$\widehat{V}^{\mathrm{IJ}}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (I_i - \bar{I})(I_i - \bar{I})^{\top}, \qquad \bar{I} = n^{-1} \sum_{i=1}^{n} I_i,$$

which, for a univariate θ, reduces to

$$\widehat{\mathrm{Var}}^{\mathrm{IJ}}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (I_i - \bar{I})^2.$$

This estimator targets the frequentist sampling variance of $\hat{\theta}$ without additional refitting, providing a sandwich-like correction based on posterior sensitivities. In applications, this framework covers a wide range of Bayesian procedures that use working likelihoods; specific models (e.g., quantile regression with an asymmetric Laplace working likelihood) are special cases obtained by plugging in their $\ell_i(\theta)$.

5 Simulation Study

5.1 Simulation Process

We conduct a Monte Carlo study to evaluate the finite-sample performance of the IJ under two common misspecification scenarios: (i) *clustered dependence* and (ii) *heteroskedastic errors*. The simulation is designed to test the IJ's robustness in situations where conventional methods fail, while also benchmarking it against three competitors: the naive (model-based) standard error, the robust (sandwich) standard error, and the bootstrap.

In **Scenario 1** (clustered dependence), data follow a hierarchical structure:

$$y_{ij} = 1 + 2x_{ij} + u_j + \varepsilon_{ij}, \qquad u_j \sim N(0, \tau^2), \;\; \varepsilon_{ij} \sim N(0, 1),$$

for clusters $j = 1, \ldots, G$ and observations $i = 1, \ldots, m$ per cluster. Here $\rho = \tau^2/(\tau^2 + 1)$ is the intra-class correlation, which we vary at levels $\{0.1, 0.3, 0.5\}$. The true coefficients are $\beta_0 = 1$ and $\beta_1 = 2$, so that any bias in estimates will be due solely to model violations. Covariates x_{ij} are generated as $N(0, 1)$ with a within-cluster correlation $\mathrm{Cor}(x_{ij}, x_{ik}) = 0.3$ to mimic shared cluster-level features.

In **Scenario 2** (heteroskedastic errors), the model is:

$$y_i = 1 + 2x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2), \;\; \sigma_i^2 = 1 + \gamma x_i^2,$$

with $\gamma \in \{0, 0.5, 1.0\}$ controlling the severity of heteroskedasticity. The case $\gamma = 0$ corresponds to homoskedastic errors (for validating coverage under correct specification), while $\gamma > 0$ introduces a variance pattern increasing with $|x|$, similar to what arises in many economic data (e.g. larger income leads to more variable expenditure). We generate $x_i \sim N(0, 1)$ independently.

For the clustered designs, we vary the number of clusters $G \in \{10, 50, 100\}$ and cluster size $m \in \{5, 20\}$. This allows us to examine performance with a small number of large clusters versus many smaller clusters. For the heteroskedastic designs, we consider sample sizes $n \in \{100, 500, 2000\}$ to assess how quickly methods approach their asymptotic behavior. In all simulations, we use $R = 2000$ independent replications to stabilize the Monte Carlo estimates.

Our analysis emulates an applied workflow where the form of misspecification is unknown and must be detected. First, we perform diagnostic tests: a likelihood ratio test for $\tau^2 = 0$ in the clustered scenario (testing independence), and a Breusch–Pagan test for heteroskedasticity. If a test significantly rejects ($\alpha = 0.05$), we apply the corresponding robust method (cluster-robust or heteroskedasticity-robust) in addition to the naive method, and also compute the IJ and bootstrap for comparison. If a test does not reject, we proceed with the naive method and IJ to see if IJ is "safe" to use even when no obvious misspecification is detected.

We evaluate the methods on four metrics for the slope estimate $\hat{\beta}_1$ (with true value 2): (1) coverage probability of nominal 95% confidence intervals (ideally 0.95 if the method is unbiased); (2) the average 95% CI width (a measure of interval precision); (3) the mean squared error of the standard error estimates, $\mathbb{E}\big[(\mathrm{SE}_{\mathrm{method}} - \mathrm{SE}_{\mathrm{true}})^2\big]$, where $\mathrm{SE}_{\mathrm{true}}$ is the Monte Carlo standard deviation of $\hat{\beta}_1$ across simulations; and (4) the computation time of each method, to highlight computational efficiency.

5.2 Simulation Results

In both scenarios, the IJ delivers a favorable trade-off between statistical efficiency and computation. Under clustered dependence (Scenario 1), runtime summaries and distributions show that the naive approach is fastest but ignores dependence; the cluster bootstrap scales poorly as the number of clusters G increases; and the cluster-IJ remains nearly flat in G because it requires only a single fit. Consistent with this, the mean 95% interval width decreases with G, and cluster-IJ yields the narrowest intervals in small-sample regimes, whereas bootstrap and (cluster-robust) sandwich intervals are slightly wider. Coverage results indicate that all methods approach nominal coverage as G grows; however, naive inference under-covers when G is small or intra-class correlation ρ is high, while cluster-robust methods (including IJ) stay close to 0.95.

Under heteroskedasticity (Scenario 2), IJ again achieves competitive efficiency with lower computational cost than the bootstrap at larger n, while the naive method remains fastest by construction. Interval widths contract with sample size, with IJ producing the narrowest intervals for moderate to large n. Coverage is well-controlled by IJ and the bootstrap under mild to moderate heteroskedasticity, and remains comparatively robust even when heteroskedasticity is strong; in contrast, the naive method severely under-covers at small n. Overall, the results support IJ as a practical default for robust uncertainty quantification: it attains near-nominal coverage with tighter intervals than standard

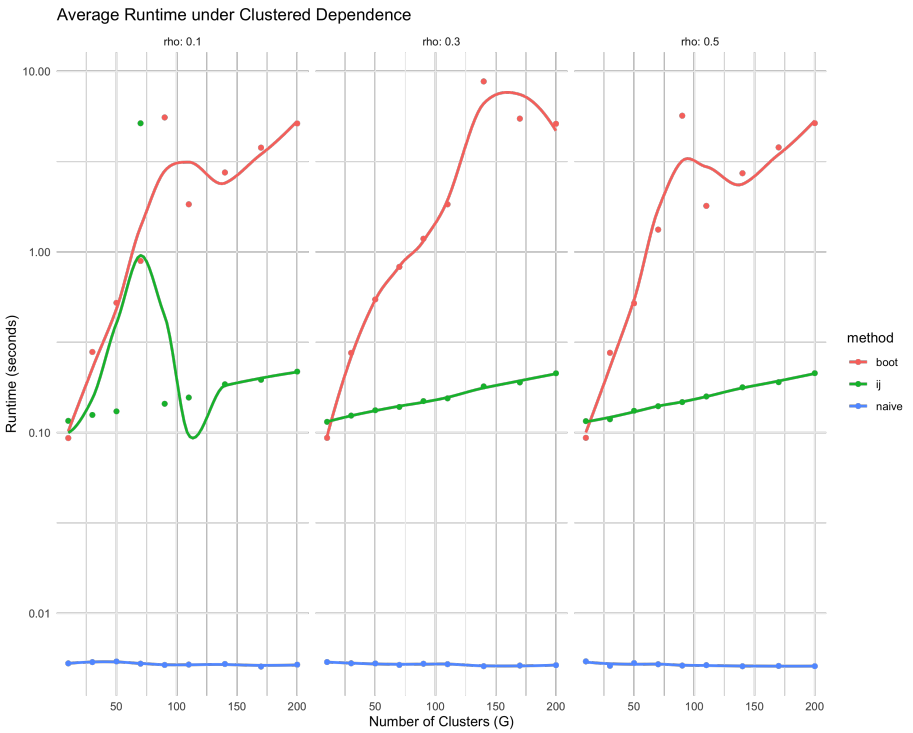

Figure 1: **Average runtime (log scale) in Scenario 1.** The naive method (blue) is fastest as it involves no resampling; cluster bootstrap (red) scales poorly, with runtime increasing sharply as the number of clusters G grows; the cluster-IJ (green) has nearly constant runtime regardless of G (one model fit), providing a huge speed advantage for large G.

robust alternatives and markedly better computational scaling than bootstrap-based procedures.

6 Discussion

Our results suggest that the infinitesimal jackknife (IJ) is a practical option for uncertainty quantification under model misspecification. It preserves the model-agnostic spirit of the bootstrap but requires only a single fit, which is useful when re-fitting is costly, such as with MCMC. In both the clustered and heteroskedastic settings, IJ produced intervals with coverage close to the nominal level and widths that were comparable to, and often smaller than, those from the sandwich and the bootstrap, while keeping runtime low and stable. The method is, however, a first-order linear approximation. Its accuracy can degrade with severe nonlinearity, very influential observations, a very small

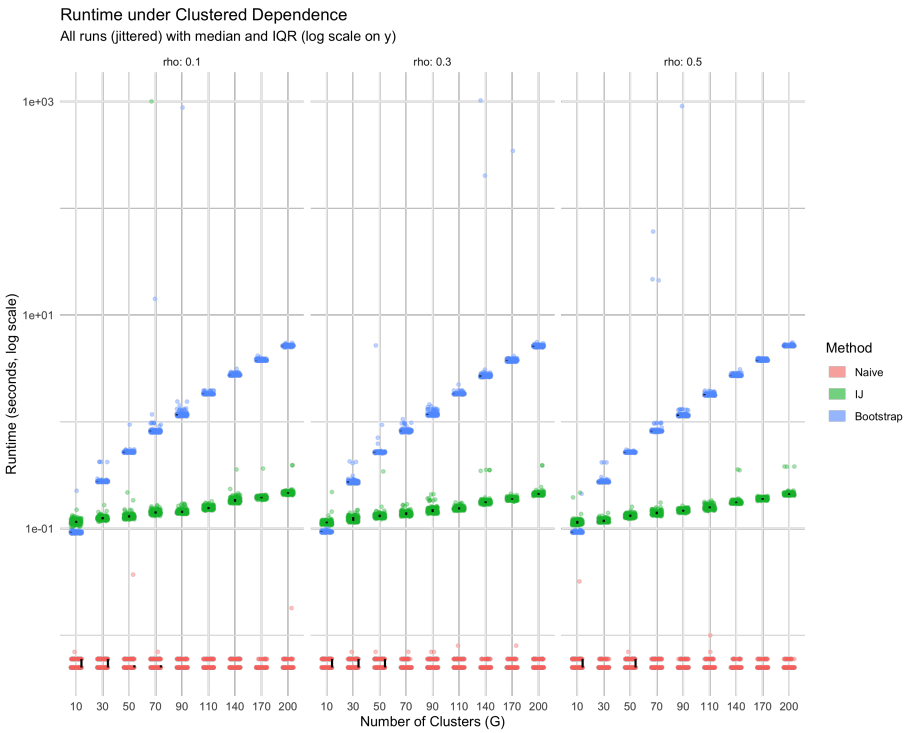

Figure 2: **Runtime distribution (log scale) in Scenario 1.** Points (jittered) show individual runs; short horizontal bars mark the median and vertical lines mark the interquartile range. The naive method (blue) remains fastest; the cluster-IJ (green) scales nearly flat in G (single fit); the cluster bootstrap (red) grows rapidly with G.

number of clusters, or highly non-smooth targets (e.g., extreme quantiles) (Huber, 1967; Efron, 1982). For Bayesian analyses based on a working likelihood, such as asymmetric Laplace quantile regression, IJ standard errors tended to be better calibrated than raw posterior spreads under misspecification, in line with recent recommendations to assess the frequentist variability of posterior functionals (Sriram et al., 2015; Syring and Martin, 2019; Giordano and Broderick, 2024). In practice, we recommend fitting once, computing IJ (with cluster aggregation when needed), and, for critical applications, validating with a small pilot bootstrap. Extensions to time-series dependence (HAC-style IJ), penalized and high-dimensional estimators, and small-sample calibrations are natural next steps.

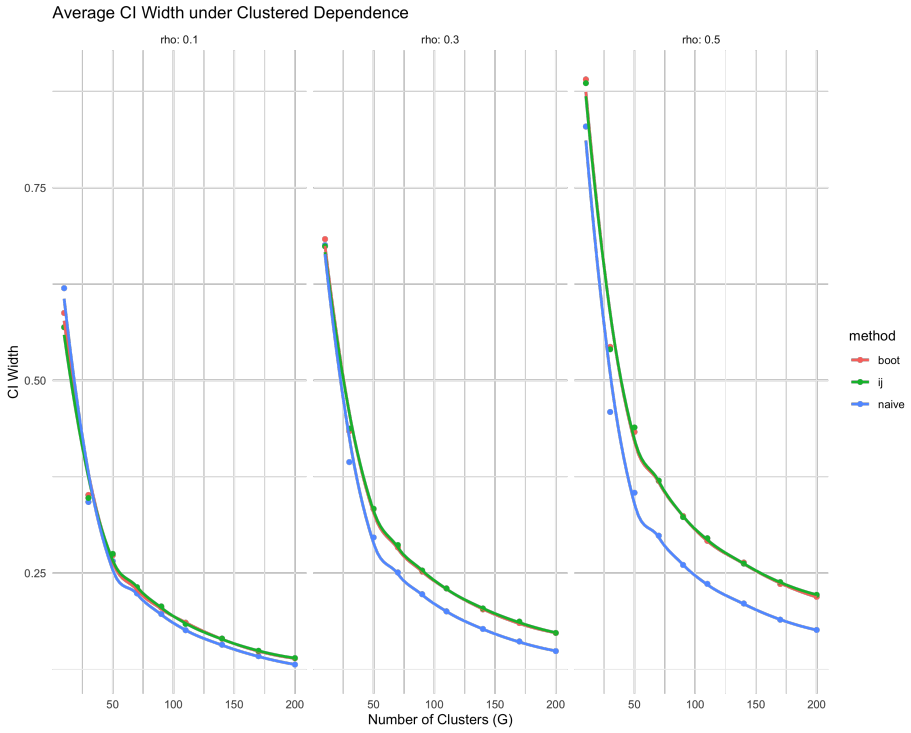

Figure 3: **Mean 95% CI width in Scenario 1.** Confidence intervals shrink as the number of clusters G increases (more information). The cluster-IJ (green) yields the narrowest intervals in small-sample settings, indicating higher efficiency, while the bootstrap (red) and sandwich (blue) intervals are slightly wider.

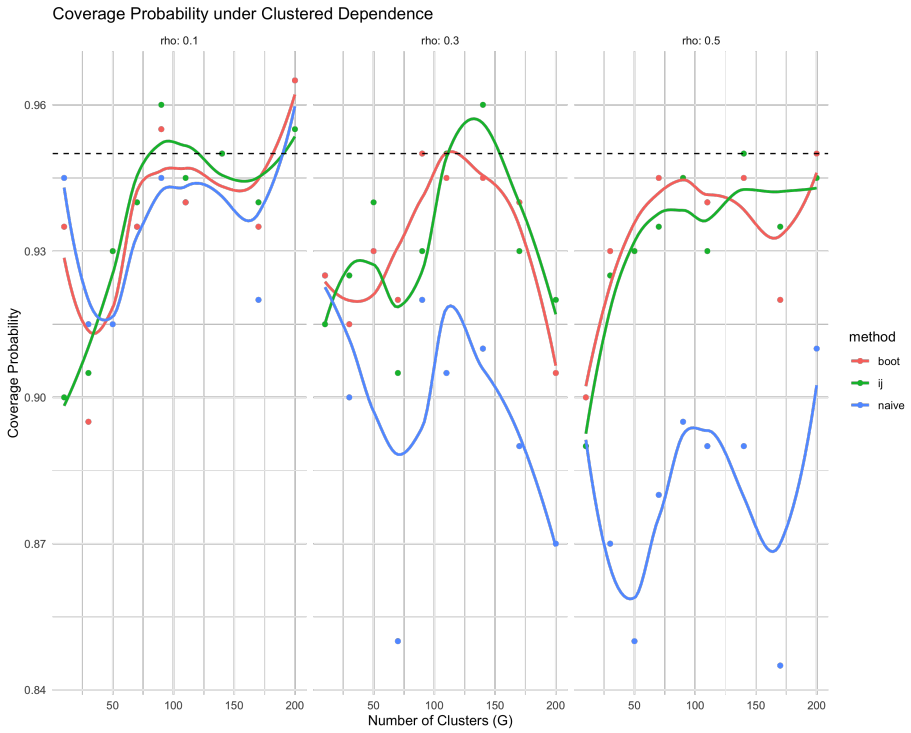

Figure 4: **Coverage probability in Scenario 1.** The dashed line marks the nominal 0.95 level. As G increases, all methods approach nominal coverage. The naive method (blue) under-covers when the number of clusters is small or the intra-class correlation ρ is high, whereas the cluster-robust methods (IJ in green, bootstrap in red, and sandwich in blue when adjusted) maintain coverage closer to 95%.

Figure 5: **Average runtime (log scale) in Scenario 2.** Here, the IJ (green) involves computing derivatives but still requires only one model fit, making it faster than the bootstrap (red) for larger n. The naive (blue) is fastest as it assumes homoskedasticity. Runtime is relatively insensitive to sample size n for all methods shown (note the log scale).

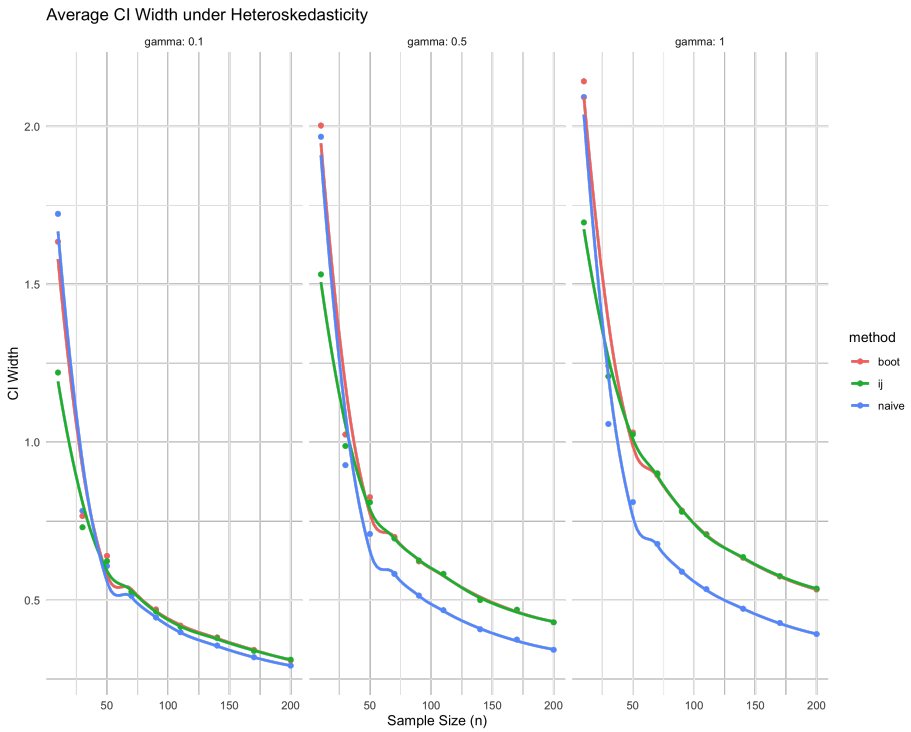

Figure 6: **Mean 95% CI width in Scenario 2.** Interval width contracts with larger n. The IJ (green) yields the narrowest intervals for moderate to large samples; bootstrap (red) and robust (blue) intervals are slightly wider on average, reflecting the IJ's greater efficiency in utilizing the data.

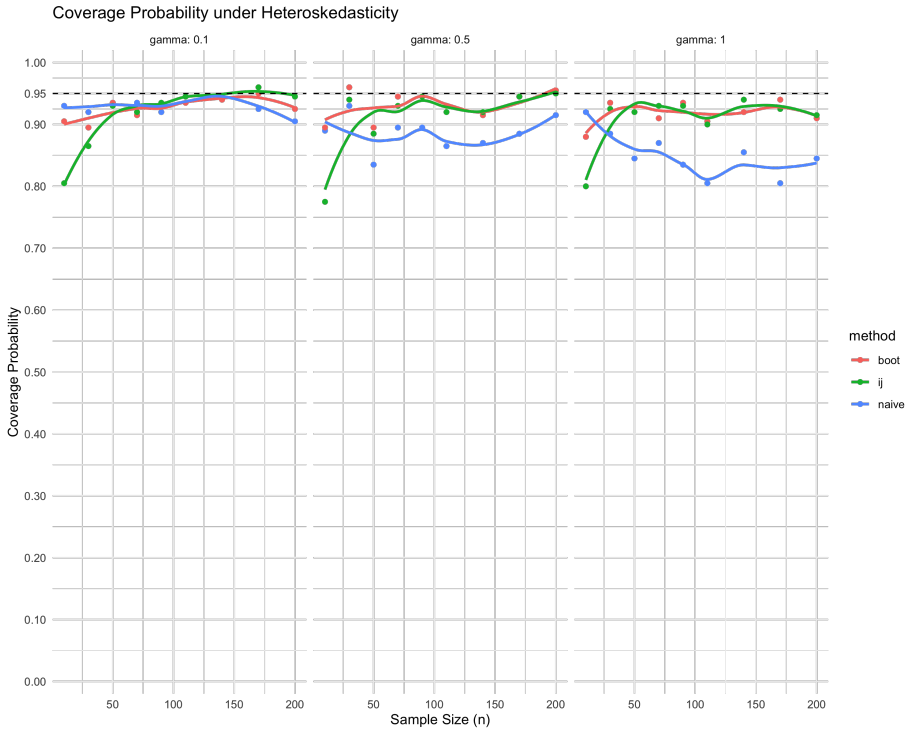

Figure 7: **Coverage probability in Scenario 2.** For mild to moderate heteroskedasticity ($\gamma = 0.5$), both IJ (green) and bootstrap (red) maintain coverage close to 0.95 even at smaller n. When heteroskedasticity is strong ($\gamma = 1.0$), the naive method (blue) under-covers severely for small samples, whereas the robust methods control coverage relatively well, especially as n grows.

References

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

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

Giordano, R. and Broderick, T. (2024). A swiss army infinitesimal jackknife: Variance estimation for bayesian posterior statistics. *Journal of the American Statistical Association*. arXiv:2209.07426.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:221–233.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall/CRC.

Koenker, R. W. and Bassett, Gilbert, J. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Kutner, M. H. and Neter, J. (2004). *Applied Linear Regression Models*. Irwin/McGraw-Hill.

MacKinnon, J. G. and White, H. (1985). Some critical values for hypothesis tests with strong instruments. *Econometric Reviews*, 4(1):131–150.

Sriram, K., Ramamoorthi, R. V., and Ghosh, J. K. (2015). Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, 10(4):937–962.

Syring, N. and Martin, R. (2019). Calibrated bayes: A bayes/frequentist compromise that improves posterior interval coverage. *Electronic Journal of Statistics*, 13(1):2269–2309.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Yang, Y. and He, X. (2016). Posterior consistency of bayesian quantile regression using misspecified asymmetric laplace likelihood. *Statistica Sinica*, 26(2):667–684.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130.

Appendix: Technical Definitions and Derivations

A Concepts

A.1 Influence Functions

Definition: An influence function provides a way to measure how an individual data point affects a statistical estimator. Given a distribution F, the influence function of an estimator $T(F)$ at the point x is defined as:

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon},$$

where δ_x is a distribution that puts all its mass at x. Intuitively, $IF(x; T, F)$ describes the first-order effect on the estimate $T(F)$ of placing an infinitesimal proportion of probability at the point x.

Influence functions are closely tied to the concept of robust statistics. In robust statistics, an estimator is considered robust if its influence function is bounded—meaning that no single observation can have an arbitrarily large influence on the estimate. Moreover, influence functions allow for direct estimation of an estimator's variance using the relationship:

$$V_{\hat{\theta}} = \frac{1}{n} E\Big[IF(x; T, F)^2\Big],$$

where the expectation is taken with respect to the true distribution F. This forms the basis of robust (sandwich) standard errors, which are estimated using the empirical variance of the influence function values across the sample.

Why are influence functions needed? They provide several key benefits. First, they allow estimation of standard errors without relying on strict parametric assumptions (such as homoscedasticity or normality of errors). Second, they offer a clear measure of estimator sensitivity, helping to identify outliers or influential points. Third, they underpin robust standard error formulas, which remain valid even in the presence of heteroskedasticity or certain other forms of misspecification.

Relationship to robust standard errors: Influence functions are the theoretical foundation for robust standard errors. By calculating the influence of each observation on the estimator, we can measure the variability of the estimator without assuming constant variance. Specifically, the robust (sandwich) variance estimator is obtained by essentially taking the empirical variance of the influence function values:

$$SE^2_{\text{robust}} = \frac{1}{n} \sum_{i=1}^{n} IF(x_i; T, \hat{F})^2,$$

where $\hat{F}$ is the empirical distribution of the observed data. This is exactly the form of the infinitesimal jackknife variance estimator described in the main text.

A.2 Robust Standard Errors

Definition: A robust standard error is an uncertainty estimate for an estimator that remains valid even when certain model assumptions are violated. For example, in a regression $Y = X\beta + \varepsilon$, the conventional standard error formula for $\hat{\beta}$ assumes homoscedastic and uncorrelated errors. A robust standard error formula, by contrast, does not assume $\mathrm{Var}(\varepsilon_i)$ is constant.

Formally, the robust (White–Huber) variance estimator for $\hat{\beta}$ in the linear model can be written as:

$$\widehat{\mathrm{Var}}_{\mathrm{robust}}(\hat{\beta}) = (X^\top X)^{-1}\Big(\sum_{i=1}^{n} x_i x_i^\top \hat{r}_i^2\Big)(X^\top X)^{-1},$$

where $\hat{r}_i$ are the OLS residuals. This is a special case of the sandwich formula $A^{-1}BA^{-1}$ mentioned in the main text.

Why are robust SEs needed? They are essential when model assumptions (like equal variance or independence of errors) are questionable. Using standard errors derived under false assumptions can lead to underestimating uncertainty—yielding overly optimistic (too narrow) confidence intervals and overly liberal hypothesis tests (Type I error inflation).

Relation to the bootstrap: The bootstrap offers a non-parametric way to estimate standard errors by resampling the data. Interestingly, robust standard errors and the bootstrap are often in close agreement. Both aim to approximate the true sampling variability of an estimator without relying on strong parametric assumptions. In fact, robust SEs can be seen as a linear approximation (via the influence function) to the full resampling distribution that the bootstrap would generate. The robust SE formula effectively linearizes the estimator and then assesses variability, whereas the bootstrap empirically builds the distribution of the estimator by resampling. When n is large and the estimator is relatively smooth, the robust SE and bootstrap SE usually coincide (up to simulation error in the bootstrap).

A.3 Jackknife and Infinitesimal Jackknife

Jackknife: The jackknife is a resampling technique that estimates the variability of a statistic by recomputing it on leave-one-out subsets of the data. Given an estimator $\hat{\theta} = T(\{x_1, \ldots, x_n\})$, the jackknife produces n replicates $\hat{\theta}_{(-i)} = T(\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\})$. The jackknife estimate of variance is

$$\widehat{\mathrm{Var}}_{\mathrm{jackknife}}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}\Big(\hat{\theta}_{(-i)} - \bar{\theta}_{(\cdot)}\Big)^2,$$

where $\bar{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}$.

Infinitesimal jackknife: The infinitesimal jackknife takes the idea of the jackknife to the infinitesimal limit. Instead of removing one observation entirely, it considers down-weighting each observation by an infinitesimal amount. The

influence function formalism (see above) is used to derive the effect of this infinitesimal change. The IJ variance estimator is then

$$\widehat{\mathrm{Var}}_{\mathrm{IJ}}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} IF(x_i; T, \hat{F})^2,$$

which as noted is asymptotically equivalent to the jackknife and bootstrap under regularity conditions. The IJ is computationally appealing because one can often obtain $IF(x_i; T, \hat{F})$ in closed form (or via automatic differentiation) without performing n separate fits.

A.4 Bayesian Quantile Regression

Background: Quantile regression (Koenker and Bassett, 1978) seeks to estimate conditional quantiles of Y given $X = x$. The τ-quantile function is defined implicitly by $P(Y \leq Q_\tau(Y|x) \mid X = x) = \tau$. In a linear quantile regression model, one assumes $Q_\tau(Y \mid X = x) = x^\top \beta(\tau)$ for some unknown coefficient vector $\beta(\tau)$. The classical (frequentist) estimator of $\beta(\tau)$ is obtained by solving:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta),$$

where $\rho_\tau(u) = u[\tau - \mathbb{I}\{u < 0\}]$ is the check loss.

Bayesian approach: Yu and Moyeed (2001) noted that if one assumes $y_i \mid x_i, \beta, \sigma \sim \mathrm{AL}(x_i^\top \beta, \sigma, \tau)$ (an asymmetric Laplace distribution with location $x_i^\top \beta$, scale σ, and asymmetry τ), then the maximum likelihood estimate coincides with the quantile regression solution. This makes the AL likelihood a convenient "working" likelihood for Bayesian quantile regression. One puts a prior on (β, σ) and uses the AL likelihood, obtaining a posterior that can be sampled via MCMC. Importantly, however, the AL likelihood is not true generative model for $y_i \mid x_i$ in general—it is chosen for computational convenience. Thus, the model is typically misspecified (unless the true $y_i \mid x_i$ distribution just happens to be AL).

Posterior vs. sampling variance: Under this misspecification, the posterior for $\beta(\tau)$ will still concentrate around the "quasi-true" value (the minimizer of the check loss), but the posterior credible intervals can be poorly calibrated as confidence intervals. The IJ procedure described in the main text provides a way to get frequentist-valid standard errors from the Bayesian output. Essentially, it acknowledges that the AL likelihood was just a device to get an estimate, and uses the variability of the score (influence) contributions to assess uncertainty.