Using gradients to check sensitivity of MCMC-based analyses to removing data

Tin Nguyen [1] **Ryan Giordano** [2] **Rachael Meager** [3] **Tamara Broderick** [1]

Abstract

If the conclusion of a data analysis is sensitive to dropping very few data points, that conclusion might hinge on the particular data at hand rather than representing a more broadly applicable truth. To check for this sensitivity, one idea is to consider every small data subset, drop it, and re-run our analysis. But the number of re-runs needed is combinatorially large. Recent work proposes a differentiable relaxation to find the worst-case subset, but that work was developed for conclusions based on estimating equations — and does not directly handle Bayesian posterior approximations using MCMC. We make two principal contributions. We adapt the existing data-dropping relaxation to estimators computed via MCMC; in particular, we re-use existing MCMC draws to estimate the necessary derivatives via a covariance relationship. Observing that Monte Carlo errors induce variability in the estimates, we use a variant of the bootstrap to quantify this uncertainty. Empirically, our method is accurate in simple models, such as linear regression. In models with complex structure, such as hierarchies, the performance of our method is mixed.

1. Introduction

Consider this motivating example. Angelucci et al. (2015) conducted a randomized controlled trial (RCT) in Mexico to study whether microcredit loans improve business profits. To analyze this data, one might use a simple Bayesian model and Markov chain Monte Carlo (MCMC). Based on the posterior mean MCMC estimate, one might conclude that microcredit actually reduces profit in this study. Next, if a policymaker wants to advocate against microcredit deployment outside of Mexico, they need to know if microcredit remains detrimental beyond the data gathered in Angelucci et al. (2015). More broadly, many researchers analyze data with Bayesian models and MCMC (Senf et al., 2020; Meager, 2022; Jones et al., 2021; Porter et al., 2022) and want to know if their conclusions generalize beyond their data.

If one is interested in generalization, an intuitive idea is checking if a conclusion changes after dropping very few data points. For instance, we show in Section 4.1 that after removing less than 0.1% of the RCT data (Angelucci et al., 2015), microcredit is estimated to increase, rather than decrease, profit. So, the conclusion hinges on a small number of businesses. We do not know if these data points appear elsewhere. Hence, we worry about generalization. For more motivation on why small-data sensitivity is related to generalization, see Appendix A.1.

One might want to know if a similar sensitivity exists for other analyses. A natural idea is to enumerate over every data subset and re-analyze. Unfortunately, the number of subsets to search over is prohibitively large. See Appendix A.2 for an estimate on brute-force runtime. So we turn to approximations. While there has been previous work on approximating dropping worst-case data subsets (Broderick et al., 2023; Kuschnig et al., 2021; Shiffman et al., 2023; Moitra and Rohatgi, 2022; Freund and Hopkins, 2023), none directly apply to MCMC: the focus of past works have been either estimating equations (Kosorok, 2008)[Chapter 13] or ordinary least squares (OLS).

Our work extends Broderick et al. (2023) to conclusions based on MCMC. In Section 3.1, similar to Broderick et al. (2023), we use a first-order Taylor series. We observe that the first derivatives can be interpreted as posterior covariances, and we use MCMC to estimate the covariances (Section 3.2). Recognizing that Monte Carlo errors induce variability in our approximation, in Section 3.3 we use a variant of the bootstrap Efron (1979) to quantify this uncertainty. We provide a longer discussion of related work in Appendix A.3. Experimentally, in Section 4, while our approximation performs well in simple models such as linear regression, it is less reliable in complex models.

2. Background

We introduce notation in two parts. First, we cover the notation and concepts involved in Bayesian data analysis.

[1]MIT [2]University of California, Berkeley [3]University of New South Wales. Correspondence to: Tin Nguyen <tdn@mit.edu>.

Published at the 2^{nd} Differentiable Almost Everything Workshop at the 41^{st} International Conference on Machine Learning, Vienna, Austria. July 2024.

Second, we extend the notation to dropping data.

Bayesian data analysis. Suppose we have a dataset $\{d^{(n)}\}_{n=1}^N$. Consider a parameter $\beta \in \mathbb{R}^V$ of interest. To estimate β, we take a Bayesian approach. First, we model the link between β and the data through a likelihood $L(d^{(n)} \mid \beta)$. Secondly, we specify a prior distribution over β, and use $p(\beta)$ to denote the prior density. Then, the density of the posterior distribution of β given the data is $p(\beta \mid \{d^{(n)}\}_{n=1}^N) \propto p(\beta) \prod_{n=1}^N \exp(L(d^{(n)} \mid \beta))$.

In practice, an analyst uses a posterior functional to make conclusions. One example is the posterior mean $\mathbb{E}g(\beta)$, where g is a mapping from $\mathbb{R}^V$ to $\mathbb{R}$. In linear regression, commonly a practitioner will make a decision based on the sign of the posterior mean of a particular regression coefficient. Other decisions are made with uncertainty intervals: see Appendix B.1 for details.

Computationally, posterior functionals need not have closed forms. To approximate posterior functionals, practitioners frequently use MCMC methods. Let $(\beta^{(1)}, \ldots, \beta^{(S)})$ denote the MCMC draws that target the posterior distribution. We estimate expectations using $(\beta^{(1)}, \ldots, \beta^{(S)})$, and make a decision based on such estimates.

Dropping data. A Bayesian analyst might be worried if the substantive decision arising from their data analysis changed after removing some small fraction α of the data. For instance, if their decision were based on the sign of the posterior mean, they would be worried if that sign changed. Other decision changes are given in Appendix B.1.

To describe non-robustness precisely and to develop our approximation, we need to indicate the dependence of posterior functionals on the presence of data points. We introduce data weights $w = (w_1, w_2, \ldots, w_N)$. This vector defines the so-called *weighted* posterior distribution.

Definition 2.1. *Let $Z(w)$ be the normalizing constant for $p(\beta) \prod_{n=1}^N \exp(w_n L(d^{(n)} \mid \beta))$. If $Z(w) < \infty$, the density of the weighted posterior distribution associated with w is denoted by $p(\beta \mid w, \{d^{(n)}\}_{n=1}^N)$, and is equal to $\frac{1}{Z(w)} p(\beta) \exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)$.*

If $w_n = 0$, the n-th observation is ignored; if $w_n = 1$, the n-th observation is fully included. We recover the regular posterior by setting all weights to 1: $w = \mathbf{1}_N = (1, 1, \ldots, 1)$. It is possible that $p(\beta) \exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)$ is not integrable. For example, the prior $p(\beta)$ is improper and all weights have been set to zero: $w = \mathbf{0}_N = (0, 0, \ldots, 0)$. We assume that any likelihood contribution is enough to define a proper posterior.

Assumption 2.1. $\forall w \in [0,1]^N \setminus \{\mathbf{0}_N\}$, $Z(w) < \infty$.

This assumption is immediate in the case of proper prior and standard likelihoods.

To indicate that an expectation is taken with respect to the randomness $\beta \sim p(\beta \mid w, \{d^{(n)}\}_{n=1}^N)$, we use $\mathbb{E}_w$. The value of a posterior functional depends on w. For instance, the posterior mean under the weighted posterior is $\mathbb{E}_w g(\beta)$.

The non-robustness concern can be formalized as follows. For $\alpha \in (0,1)$, let W_α denote the set of all weight vectors that correspond to dropping no more than $100\alpha\%$ of the data i.e. $W_\alpha := \left\{ w \in \{0,1\}^N : \frac{1}{N} \sum_{n=1}^N (1 - w_n) \leq \alpha \right\}$. We say the analysis is non-robust if there exists a weight $w \in W_\alpha$ that changes the conclusion.

We focus on decisions satisying the following simplifying assumption: there exists a posterior functional ($\phi(w)$) such that $\phi(\mathbf{1}_N) < 0$ and the conclusion changes if and only if $\phi(w) > 0$. Such a functional will be called a "quantity of interest" (QoI). The decision based on sign fits this framework. To change this conclusion, if the full-data posterior mean were positive, we take $\phi(w) = -\mathbb{E}_w g(\beta)$. Appendix B.1 shows how other decision changes fit this framework.

Checking non-robustness is equivalent to a) finding the maximum value of $\phi(w)$ subject to $w \in W_\alpha$ and b) checking its sign. The outcome of this comparison remains the same if we maximize the objective function $\phi(w) - c$ and compare the optimal value with $-c$, for any constant c. Out of later convenience, we set $c = \phi(\mathbf{1}_N)$. As in Broderick et al. (2023, Section 2), we define the Maximum Influence Perturbation (MIP) to be the largest change induced by dropping no more than $100\alpha\%$ of the data. In our notation, it is

$$\max_{w \in W_\alpha} \left(\phi(w) - \phi(\mathbf{1}_N)\right). \tag{1}$$

In general, the brute-force approach to find the MIP takes a prohibitively long time; recall Appendix A.2.

3. Methods

We turn to relaxations of the MIP. We focus on $\phi(w) = \mathbb{E}_w g(\beta)$; Appendices B.1 and B.3 discusses other QoIs.

3.1. Taylor series

Our first approximation relies on the first-order Taylor series of the quantity of interest $\phi(w)$. We require that the quantity of interest $\phi(w)$ is differentiable with respect to w. The combination of Assumption 2.1 and Assumption B.2 from Appendix B.2 forms a set of mild regularity conditions that ensure this differentiability.

Theorem 3.1. *Assume Assumption 2.1 and Assumption B.2. For any $\delta \in (0,1)$, $\phi(w)$ is continuously differentiable on $\{w \in [0,1]^N : \max_n w_n \geq \delta\}$. The n-th partial derivative is equal to $Cov_w\left(g(\beta), L(d^{(n)} \mid \beta)\right)$.*

See Proof D.1 for the proof, and connections to previous

works on sensitivity analysis of posterior expectations.

We define the n-th *influence* as the partial derivative of $\phi(w)$ at $w = \mathbf{1}_N$: $\psi_n := \frac{\partial \phi(w)}{\partial w_n}\big|_{w=\mathbf{1}_N}$. Then, the first-order Taylor series approximation of $\phi(w) - \phi(\mathbf{1}_N)$ is $\sum_{n=1}^N \psi_n(w_n - 1)$. We approximately solve Equation (1) by replacing its objective function:

$$\max_{w \in W_\alpha} \quad \sum_{n=1}^N (w_n - 1)\psi_n \tag{2}$$

Solving Equation (2) involves a sort. For any $w \in W_\alpha$, the objective function is equal to $\sum_{n:w_n=0}(-\psi_n)$. Let $r_1, r_2, \ldots, r_N$ sort the ψ_n in increasing order: $\psi_{r_1} \leq \psi_{r_2} \leq \ldots \leq \psi_{r_N}$. Then, the optimal value is equal to the negative of $\sum_{m=1}^{\lfloor N\alpha \rfloor} \psi_{r_m} \mathbb{I}\{\psi_{r_m} < 0\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. We denote the optimal value by $\Delta(\alpha)$, and the data points to be removed by $U(\alpha)$.

3.2. Estimating the influence

To solve Equation (2), we need to compute the influence ψ_n. Each ψ_n is a covariance under the full-data posterior. Therefore, the MCMC draws, which are already used to estimate $\phi(\mathbf{1}_N)$, can be used to estimate ψ_n: $\psi_n \approx \hat{\psi}_n$ where $\hat{\psi}_n$ is the sample covariance between $L(d^{(n)} \mid \beta^{(s)})$ and $g(\beta^{(s)})$ (the sample being $(\beta^{(1)}, \ldots, \beta^{(S)})$).

Since $\hat{\psi}_n$ is only an approximation of ψ_n, we are not able to solve Equation (2) exactly, but only solve an approximation of it. Namely, we replace all instances of ψ_n with $\hat{\psi}_n$ in Equation (2), and solve: $\max_{w \in W_\alpha} \sum_{n=1}^N (w_n - 1)\hat{\psi}_n$. The algorithm to solve this is analogous of that for Equation (2). For instance, let the ranks $v_1, v_2, \ldots, v_N$ be such that $\hat{\psi}_{v_1} \leq \hat{\psi}_{v_2} \leq \ldots \leq \hat{\psi}_{v_N}$. Then the optimal value is $-\sum_{m=1}^{\lfloor N\alpha \rfloor} \hat{\psi}_{v_m} \mathbb{I}\{\hat{\psi}_{v_m} < 0\}$. We denote this value by $\widehat{\Delta}$, and the data points to be removed by $\widehat{U}$.

3.3. Confidence intervals

$\widehat{\Delta}$ is a noisy point estimate of $\Delta(\alpha)$. One concern regarding the quality of $\widehat{\Delta}$ is noise due to sampling uncertainty of $(\beta^{(1)}, \ldots, \beta^{(S)})$. So, we design confidence intervals.

Exact sampling. In certain cases, such as conjugate models (Diaconis and Ylvisaker, 1979), we can generate exact Monte Carlo draws from the posterior distribution. Then, conceptually, $\widehat{\Delta}$ is an estimator constructed from an i.i.d. sample. We appeal to the bootstrap (Efron, 1979), a general-purpose technique to quantify the sampling uncertainty.

Our confidence interval construction proceeds in three steps. First, we define the so-called *bootstrap distribution* of $\widehat{\Delta}$. Second, we approximate this distribution with an empirical distribution based on Monte Carlo draws. Finally, we use an interquantile range of this empirical distribution as our confidence interval for $\Delta(\alpha)$. For details, see Appendix B.4.

General MCMC. We now handle the case in which $(\beta^{(1)}, \ldots, \beta^{(S)})$ exhibit dependence. One idea is to use the previous section's construction without modification. Theoretically, it is known that the bootstrap struggles on non-i.i.d. samples, for even simple estimators. Intuitively, the bootstrap fails because the bootstrap draws do not have any dependence, while the original draws do. Appendix B.5 explains this in detail.

To improve upon the bootstrap, one option is to resample in a way that respects the original sample's dependence structure. In particular, we use the non-overlapping block bootstrap (Lahiri, 2003; Carlstein, 1986): instead of resampling individual Markov chain states, we first divide the Markov chain into blocks, and then resample blocks. For details, see Appendix B.5. The outcome of this resampling is an interval, denoted by $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$, which is our confidence interval for $\Delta(\alpha)$.

4. Experiments

Now, we check the quality of our approximations empirically on real data analyses. For a particular MCMC run, we estimate sensitivity for a range of α values: see Appendix E for details. For each α, our method proposes an influential data subset ($\widehat{U}$) and a change in the quantity of interest, represented by a confidence interval ($[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$). We plot how the change from re-running minus the proposed data compares to the confidence interval.

4.1. Linear model

We consider a slight variation of Meager (2019)'s analysis of the Angelucci et al. (2015) RCT. For context and experimental details, see Appendix F.1. In this simple linear model, our prediction typically contains the rerun.

The most interesting parameter, θ, is the treatment effect: the difference in profit between treatment and control. Figure 1 plots the histogram of the treatment effect draws as well as key sample summaries. The sample mean is equal to -4.55. The sample standard deviation is 5.79. The uncertainty interval is $(-16.10, 6.99)$. An analyst might conclude that while the posterior mean of the effect of microcredit is negative, the uncertainty interval covers zero, so they cannot confidently conclude that microcredit either helps or hurts.

In Figure 2, we plot our confidence intervals and the result after removing the proposed data. For changing sign, our method predicts there exists a data subset at most 0.1% such that if we remove it, we change the posterior mean's sign. Refitting after removing the proposed data confirms this prediction. For changing significance, our method predicts

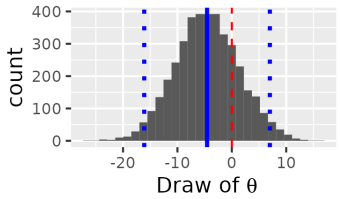

Figure 1. (Linear model) Histogram of treatment effect MCMC draws. Blue line is sample mean. Dashed red line is zero threshold. Dotted blue lines are uncertainty interval endpoints.

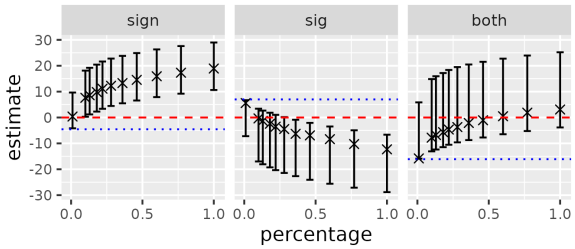

Figure 2. (Linear model) Confidence interval and refit. At maximum, we remove 1% of the data. Each panel corresponds to a target conclusion change: 'sign' is the change in sign, 'sig' is change in significance, and 'both' is the change to a significant effect of the opposite sign. Error bars are confidence interval for refit after removing the most extreme data subset. Each 'x' is the refit after removing the proposed data and re-running MCMC. The dotted blue line is the fit on the full data.

there exists a data subset of relative size at most 0.36% such that if we remove it, we change the sign of the uncertainty interval's right endpoint: refitting confirms this prediction. Our method is not able to predict whether the result can be changed to significant effect of the opposite sign for these α values and this number of draws: we recommend more MCMC draws, which should decrease interval width.

4.2. Hierarchical model

We consider a slight variation of the analysis of European tree mortality from Senf et al. (2020). For details, see Appendix F.2. Our approximation struggles: we predict more extreme changes than realized by re-running.

The most interesting parameter, θ, is the association between water availability and the amount of tree canopy death. Figure 3 plots the histogram of the association effect draws and sample summaries. The sample mean is equal to -1.88. The sample standard deviation is 0.48. The uncertainty interval is $(-2.81, -0.94)$. One might decide that water balance has a negative relationship with canopy mortality, since the posterior mean is negative, and this relationship is significant, since the uncertainty interval omits zero.

Figure 4 plots our confidence intervals and the reruns. In

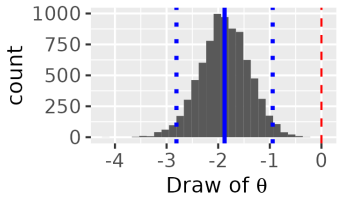

Figure 3. (Hierarchical model for tree mortality) Histogram of slope MCMC draws. See the caption of Figure 1 for the meaning of the distinguished vertical lines.

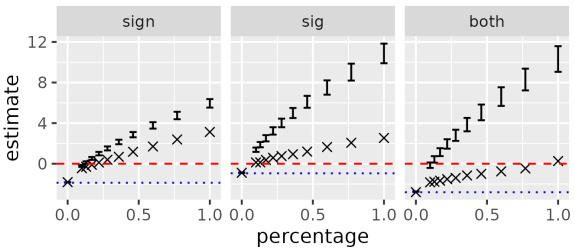

Figure 4. (Hierarchical model for tree mortality) Confidence interval and refit. See the caption of Figure 2 for the meaning of the panels and the distinguished lines.

general, our interval is not conservative. The overestimation is particularly severe for the 'both' QoI and the 'sig' QoI. For changing sign, our method predicts there exists a data subset of relative size at most 0.17% such that if we remove it, we change the posterior mean's sign; refitting does not confirm this prediction, however. The smallest α whose refit's posterior mean actually changes sign is 0.22%. For changing significance, our method predicts there exists a data subset of relative size at most 0.10% such that if we remove it, we change the sign of the right endpoint; refitting confirms this prediction. For generating a significant result of the opposite sign, our method predicts there exists a data subset of relative size at most 0.17% such that if we remove it, we change the sign of the left endpoint; refitting does not confirm this prediction, however. The smallest α whose refit's left endpoint actually changes sign is 1.0%.

5. Next Steps

We have provided a fast approximation to what happens to conclusions made with MCMC in Bayesian models when a small percentage of data is removed. In real data experiments, our approximation is accurate in simple models, such as linear regression. In complicated models, such as hierarchical ones, our methods are less accurate. Naturally, we would like to understand *why* the performance in mixed. In Appendix E, we describe the checks that could be run to identify error sources. Appendix F.1 and Appendix F.2 show the outcomes of running such checks.

6. Acknowledgments

The authors are grateful to Hannah Diehl for useful discussions and comments. This work was supported in part by an ONR Early Career Grant (N000142012532), an NSF CAREER Award (1750286), and SystemsThatLearn@CSAIL Ignite Grant.

References

M. Angelucci, D. Karlan, J. Zinman, K. Brennan, E. Degnan, A. Fishbane, A. Hillis, H. Koizumi, E. Safran, R. Strohm, B. Torres, A. Troychansky, I. Velez, G. Startz, S. Swamy, M. White, A. York, and C. Banco, "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco," *American Economic Journal: Applied Economics*, vol. 7, pp. 151–82, 2015.

C. Senf, A. Buras, C. S. Zang, A. Rammig, and R. Seidl, "Excess forest mortality is consistently linked to drought across europe," *Nature Communications*, vol. 11, 12 2020.

R. Meager, "Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature," *American Economic Review*, vol. 112, no. 6, pp. 1818–47, June 2022.

T. C. Jones, G. Biele, B. Mühlemann, T. Veith, J. Schneider, J. Beheim-Schwarzbach, T. Bleicker, J. Tesch, M. L. Schmidt, L. E. Sander, F. Kurth, P. Menzel, R. Schwarzer, M. Zuchowski, J. Hofmann, A. Krumbholz, A. Stein, A. Edelmann, V. M. Corman, and C. Drosten, "Estimating infectiousness throughout SARS-CoV-2 infection course," *Science*, vol. 373, 7 2021.

T. Porter, D. C. Molina, A. Cimpian, S. Roberts, A. Fredericks, L. S. Blackwell, and K. Trzesniewski, "Growth-Mindset Intervention Delivered by Teachers Boosts Achievement in Early Adolescence," *Psychological Science*, vol. 33, pp. 1086–1096, 7 2022.

T. Broderick, R. Giordano, and R. Meager, "An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?" 2023.

N. Kuschnig, G. Zens, and J. C. Cuaresma, "Hidden in Plain Sight: Influential Sets in Linear Models," 2021.

M. Shiffman, R. Giordano, and T. Broderick, "Could dropping a few cells change the takeaways from differential expression?" 2023.

A. Moitra and D. Rohatgi, "Provably Auditing Ordinary Least Squares in Low Dimensions," 2022.

D. Freund and S. B. Hopkins, "Towards Practical Robustness Auditing for Linear Regression," 2023.

M. R. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.

B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

P. Diaconis and D. Ylvisaker, "Conjugate Priors for Exponential Families," *The Annals of Statistics*, vol. 7, pp. 269–281, 1979.

S. N. Lahiri, *Resampling Methods for Dependent Data*. Springer, 2003.

E. Carlstein, "The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence," *Annals of Statistics*, vol. 14, pp. 1171–1179, 1986.

R. Meager, "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, vol. 11, pp. 57–91, 2019.

G. Arya, M. Schauer, F. Schäfer, and C. Rackauckas, "Automatic differentiation of programs with discrete randomness," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 435–10 447, 2022.

G. Arya, R. Seyer, F. Schäfer, K. Chandra, A. K. Lew, M. Huot, V. K. Mansinghka, J. Ragan-Kelley, C. Rackauckas, and M. Schauer, "Differentiating Metropolis-Hastings to Optimize Intractable Densities," 2023.

R. Seyer, "Differentiable Monte Carlo samplers with piecewise deterministic Markov processes," 2023.

J. P. Kleijnen and R. Y. Rubinstein, "Optimization and sensitivity analysis of computer simulation models by the score function method," *European Journal of Operational Research*, vol. 88, no. 3, pp. 413–427, 1996.

M. C. Fu and J.-Q. Hu, "Conditional Monte Carlo gradient estimation," in *Conditional Monte Carlo: Gradient estimation and optimization applications*. Springer Science & Business Media, 2012, vol. 392.

B. Heidergott and F. Vázquez-Abad, "Measure-valued differentiation for Markov chains," *Journal of Optimization Theory and Applications*, vol. 136, no. 2, pp. 187–209, 2008.

R. Giordano, T. Broderick, and M. I. Jordan, "Covariances, Robustness, and Variational Bayes," *Journal of Machine Learning Research*, vol. 19, pp. 1–49, 2018.

S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih, "Monte Carlo Gradient Estimation in Machine Learning," *Journal of Machine Learning Research*, vol. 21, pp. 1–62, 2020.

R. Giordano and T. Broderick, "The Bayesian Infinitesimal Jackknife for Variance," 2023.

R. Giordano, R. Liu, M. I. Jordan, and T. Broderick, "Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics (with Discussion)," *Bayesian Analysis*, vol. 18, pp. 287–366, 2023.

P. Diaconis and D. Freedman, "On the Consistency of Bayes Estimates," *The Annals of Statistics*, pp. 1–26, 1986.

F. Ruggeri and L. Wasserman, "Infinitesimal sensitivity of posterior distributions," *The Canadian Journal of Statistic*, vol. 21, pp. 195–203, 1993.

P. Gustafson, "Local Sensitivity of Posterior Expectations," *The Annals of Statistics*, vol. 24, p. 195, 1996.

W. Johnson and S. Geisser, "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis," *Journal of the American Statistical Association*, vol. 78, pp. 137–144, 1983.

R. E. Mcculloch, "Local Model Influence," *Journal of the American Statistical Association*, vol. 84, pp. 473–478, 1989.

M. Lavine, "Local predictive influence in bayesian linear models with conjugate priors," *Communications in Statistics - Simulation and Computation*, vol. 21, pp. 269–283, 1 1992.

B. P. Carlin and N. G. Polson, "An Expected Utility Approach to Influence Diagnostics," *Journal of the American Statistical Association*, vol. 86, pp. 1013–1021, 1991.

E. C. Marshall and D. J. Spiegelhalter, "Identifying outliers in Bayesian hierarchical models: a simulation-based approach," *Bayesian Analysis*, vol. 2, pp. 409–444, 2007.

R. B. Millar and W. S. Stewart, "Assessment of Locally Influential Observations in Bayesian Models," *Bayesian Analysis*, vol. 2, pp. 365–384, 2007.

A. van der Linde, "Local Influence on Posterior Distributions under Multiplicative Modes of Perturbation," *Bayesian Analysis*, vol. 2, pp. 319–332, 2007.

Z. M. Thomas, S. N. MacEachern, and M. Peruggia, "Reconciling Curvature and Importance Sampling Based Procedures for Summarizing Case Influence in Bayesian Models," *Journal of the American Statistical Association*, vol. 113, pp. 1669–1683, 10 2018.

M. T. Pratola, E. I. George, and R. E. McCulloch, "Influential Observations in Bayesian Regression Tree Models," *Journal of Computational and Graphical Statistics*, 2023.

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick, "Validated variational inference via practical posterior error bounds," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1792–1802.

S. Basu, S. R. Jammalamadaka, and W. Liu, "Local posterior robustness with parametric priors: Maximum and average sensitivity," in *Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993*, G. R. Heidbreder, Ed. Dordrecht: Springer Netherlands, 1996, pp. 97–106.

N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya-Gamma latent variables," *Journal of the American Statistical Association*, vol. 108, pp. 1339–1349, 2013.

W. H. Fleming, *Functions of Several Variables*, 2nd ed. Springer, 1977.

A. W. van der Vaart, *Asymptotic Statistics*. University of Cambridge, 1998.

C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, vol. 76, pp. 1–32, 2017.

A. Reuther, J. Kepner, C. Byun, S. Samsi, W. Arcand, D. Bestor, B. Bergeron, V. Gadepally, M. Houle, M. Hubbell, M. Jones, A. Klein, L. Milechin, J. Mullen, A. Prout, A. Rosa, C. Yee, and P. Michaleas, "Interactive supercomputing on 40,000 cores for machine learning and data analysis," in *2018 IEEE High Performance extreme Computing Conference (HPEC)*. IEEE, 2018, pp. 1–6.

S. N. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003.

A. Other Intro Details

A.1. More motivation of dropping data as generalization check

Standard tools to assess generalization often make assumptions that are not realistic in practice. For instance, an analyst might use frequentist tools (confidence interval, p-values) to predict whether their inferences hold in the broader population. The validity of these methods technically depends on the assumption that the gathered data is an independent and identically distributed (i.i.d.) sample from the broader population. In practice, we have reason to suspect that this assumption is not met; for instance, it might not be reasonable to assume that data collected in Mexico and data collected in a separate country are i.i.d. from the same distribution.

As pointed out by Shiffman et al. (2023), an analyst might hope that deviations from the i.i.d. assumption are small enough that (a) their conclusions remain the same in the broader population and (b) standard tools accurately assess generalization. On the other hand, the analyst might worry that this hope is misplaced if small, realistic deviations from i.i.d.-ness could affect the substantive conclusions of an analysis. An often-realistic kind of deviation is the missingness of a small fraction of data; for instance, some percentage of the population might not respond to a survey. So, if it were possible to remove a small fraction of data and change conclusions, the analyst might worry about generalization.

A.2. Brute-force runtime

We need to enumerate every data subset that drops no more than $100\alpha\%$ of the original data. And, for each subset, we would need to re-run MCMC to re-estimate the quantity of interest. There are more than $\binom{N}{\lfloor N\alpha \rfloor}$ elements in W_α. The RCT data from Angelucci et al. (2015) has $N = 16{,}560$ observations; even for $\alpha = 0.001$, there are more than 10^{54} subsets to consider. Each Markov chain already takes a noticeable amount of time to construct; in this analysis, to generate $S = 4{,}000$ draws, we need to run the chain for 1 minute. The total time to find the worst-case data subset is on the order of 10^{48} years.

A.3. Related work

Our work arguably fits into the intersection of three lines of work.

The first is papers on detecting sensitivity to small-data removal. Broderick et al. (2023) were the first to formulate sensitivity to dropping a small fraction of data as a check on generalization. Along with the formulation, one contribution of this work is a fast approximation to detect sensitivity when the analysis in question is based on estimating equations (Kosorok, 2008)[Chapter 13]. Regardless of how estimators are constructed, in general, the brute-force approach to finding an influential small fraction of data is computationally intractable. One would need to enumerate all possible data subsets of a given cardinality and re-analyze on each subset: even when the fraction of data removed is small and each analysis takes little time, there are too many such subsets to consider; see the discussion at the end of Section 3. For estimating equations, Broderick et al. (2023) approximate the effect of dropping data with a first-order Taylor series approximation; this approximation can be optimized very efficiently, while the brute-force approach is not at all practical. Neither Broderick et al. (2023) nor subsequent existing work on small-data removals (Kuschnig et al., 2021; Shiffman et al., 2023; Moitra and Rohatgi, 2022; Freund and Hopkins, 2023) can be immediately applied to determine sensitivity in MCMC. In particular, since MCMC cannot be cast as the root of an estimating equation or the solution to an optimization problem, neither Broderick et al. (2023) nor Shiffman et al. (2023) apply to our situation. As Kuschnig et al. (2021); Moitra and Rohatgi (2022); Freund and Hopkins (2023) focus on ordinary least squares (OLS), their work does not address our problem, either.

The second line of work estimates the changes that happen to a posterior expectation because of small perturbations to the total log likelihood. There are two conceptually distinct approaches to this sensitivity analysis.

- One approach (e.g. Arya et al., 2022, 2023; Seyer, 2023) applies to when the posterior is approximated with a Metropolis-Hastings algorithm. In particular, this approach computes the gradient of the Metropolis-Hastings sampler to small perturbations in the total log likelihood. More broadly, there is a literature on estimating gradients for random processes with discrete components (Kleijnen and Rubinstein, 1996; Fu and Hu, 2012; Heidergott and Vázquez-Abad, 2008).

- The other approach does not compute the gradient of the MCMC algorithm or steps within it. Instead, it directly computes (and then estimates) the gradient of the posterior expectation. Recent works in this literature include Giordano et al. (2018); Mohamed et al. (2020); Giordano and Broderick (2023); Giordano et al. (2023), while foundational works

include Diaconis and Freedman (1986); Ruggeri and Wasserman (1993); Gustafson (1996).

In our work, we take the second approach. A priori, it is not clear which approach is superior. Two reasons to prefer the second approach over the first approach are the following. While the discrete operations in Metropolis-Hastings, e.g. the accept/reject steps, pose a key challenge in the first approach, they do not cause any issues in the second approach; the second approach is "oblivious" to details regarding how the posterior is approximated. In addition, suppose that an analyst wishes to compute gradients of multiple quantities of interest. If they follow the first approach, for each quantity of interest, they would need to re-run the sampling algorithm to estimate the gradient. Taking the second approach, the analyst only needs to run the sampling algorithm once, and use the resulting draws to simultaneously estimate the gradient of multiple quantities of interest. On the other hand, the first approach might be better than the second approach in the following way. Our experiments later show that gradient estimates coming from the second approach can be noisy. The first approach, with the promise of variance reduction through a good choice of Markov chain coupling, might produce more accurate gradient estimates. It is an interesting direction for future work to apply the first approach to our problem and compare the performance of the two approaches.

While papers taking the second approach have already mentioned how to estimate the effect of dropping an individual observation, these estimates have not been used to assess whether conclusions based on MCMC are sensitive to the removal of a small data fraction. Some works (e.g. Gustafson, 1996; Giordano et al., 2018, 2023) generate perturbations by varying prior or likelihood choice. Giordano and Broderick (2023) estimate the frequentist variability of Bayesian procedures, a task that can be seen as equivalent to the goal of bootstrap resampling. No existing work aims to find a small fraction of data that, if dropped, would change conclusions.

The third set of works, in the Bayesian case influence literature, quantifies the importance of individual observations to a Bayesian analysis. As we will explain, existing works do not tackle our problem. Early works in this area include Johnson and Geisser (1983); Mcculloch (1989); Lavine (1992); Carlin and Polson (1991), while recent works include Marshall and Spiegelhalter (2007); Millar and Stewart (2007); van der Linde (2007); Thomas et al. (2018); Pratola et al. (2023). Such papers focus on the identification of outliers, rather than predictions about whether the conclusion changes after removing a small amount of data. Generally, this literature defines an observation to be an outlier if the Kullback–Leibler (KL) divergence between the posterior after removing the observation and the original posterior is large. For conclusions based on posterior functionals, such as the mean, we are not aware of how to systematically connect the KL divergence to the sensitivity of the decision-making process; in fact, recent work (Huggins et al., 2020) has shown that comparing probability distributions based on the KL divergence can be misleading if an analyst really cared about the comparison between the distributions' means or variances.

B. Other Methods Details

B.1. Other decisions and quantities of interest

In Section 2 and Section 3, we focus on decisions based on the sign of the posterior mean. Here, we describe other decisions, which are based on uncertainty intervals, and the corresponding quantities of interest.

An econometrician might declare that an intervention is helping some population if the vast majority of the posterior mass for a particular coefficient lies above zero. That is, the practitioner checks if the lower bound of an uncertainty interval lies above zero. This decision might be considered to reflect a Bayesian notion of *significance*.

A Bayesian analyst might be worried if the substantive decision arising from their data analysis changed after removing some small fraction α of the data. For instance,

- If their decision were based on zero falling outside a credible interval, they would be worried if we can make the credible interval contain zero.
- If their decision was based on both the sign and the significance, they would be worried if we can both change the posterior mean's sign and put a majority of the posterior mass on the opposite side of zero.

To change the conclusion about significance, if uncertainty interval's left endpoint[1] ($\mathbb{E}_{\mathbf{1}_N} g(\beta) - z_{0.975}\sqrt{\text{Var}_{\mathbf{1}_N} g(\beta)}$) were

[1]Our uncertainty interval multiplies the posterior standard deviation by $z_{0.975}$, which is the 97.5% quantile of the standard normal, but we can replace this with other scaling without undue effort.

positive, we take

$$\phi(w) = -(\mathbb{E}_w g(\beta) - z_{0.975}\sqrt{\text{Var}_w g(\beta)}).$$

$\phi(w) > 0$ is equivalent to moving the left endpoint below zero, thus changing from a significant result to a non-significant one. Finally, to change to a significant result of the opposite sign, if the uncertainty interval's left endpoint were positive, we take

$$\phi(w) = -(\mathbb{E}_w g(\beta) + z_{0.975}\sqrt{\text{Var}_w g(\beta)}).$$

On the full data, the right endpoint is above zero. On weight such that $\phi(w) > 0$, the right endpoint has been moved below zero: the conclusion has changed from a positive result to a significant negative result.

B.2. Regularity conditions

To be able to form a Taylor series, we require that the quantity of interest $\phi(w)$ is differentiable with respect to the weight w. We are not aware of a complete theory (necessary and sufficient conditions) for this differentiability. However, through Assumption B.1 and Assumption B.2, we state a set of sufficient conditions.

Assumption B.1. *Let g be a function from $\mathbb{R}^V$ to the real line. $\phi(w)$ is a linear combination of posterior mean and posterior standard deviation i.e. there exists constants c_1 and c_2, which are independent of w, such that*

$$\phi(w) = c_1\mathbb{E}_w g(\beta) + c_2\sqrt{Var_w g(\beta)}.$$

A typical choice of g is the function that returns the v-th coordinate of a V-dimensional vector.

It might appear that constraining $\phi(w)$ to be a linear combination of the posterior mean and standard deviation is overly restrictive. However, this choice encompasses many cases of practical interest: recall from Appendix B.1 that the quantities of interest for changing sign, changing significance, and producing a significant result of the opposite sign, take the form of Assumption B.1. Furthermore, the choice of constraining $\phi(w)$ to be a linear combination of the posterior mean and standard deviation in Assumption B.1 is done out of convenience. Our framework can also handle quantities of interest that involve higher moments of the posterior distribution, and the function that combines these moments need not be linear, but we omit these cases for brevity. However, we note that posterior quantiles in general do not satisfy Assumption B.1 and leave to future work the question of how to diagnose the sensitivity of such quantities of interest.

Assumption B.2. *For any $w \in [0,1]^N \setminus \{\mathbf{0}_N\}$, the following functions have finite expectations under the weighted posterior: $|g(\beta)|$, $g(\beta)^2$, $|L(d^{(n)} \mid \beta)|$ (for all n), $|g(\beta)L(d^{(n)} \mid \beta)|$ (for all n) and $|g(\beta)^2 L(d^{(n)} \mid \beta)|$ (for all n).*

The assumption is mild. It is satisfied by for instance, linear regression under Gaussian likelihood and $g(\beta) = \beta_v$.

Under Assumption 2.1, Assumption B.1, and Assumption B.2, $\phi(w)$ is continuously differentiable with respect to w.

Theorem B.1. *Assume Assumption 2.1, Assumption B.1, and Assumption B.2. For any $\delta \in (0,1)$, $\phi(w)$ is continuously differentiable with respect to w on $\{w \in [0,1]^N : \max_n w_n \geq \delta\}$. The n-th partial derivative[2] at w is equal to $c_1 f + c_2 s$ where*

$$f = Cov_w\left(g(\beta), L(d^{(n)} \mid \beta)\right), \tag{3}$$

and

$$s = \frac{Cov_w\left(g(\beta)^2, L(d^{(n)} \mid \beta)\right) - 2\mathbb{E}_w g(\beta) \times Cov_w\left(g(\beta), L(d^{(n)} \mid \beta)\right)}{\sqrt{Var_w g(\beta)}}. \tag{4}$$

See Proof D.2 for the proof. This theorem is a specific instance of the sensitivity of posterior expectations with respect to log likelihood perturbations: for further reading, we recommend Diaconis and Freedman (1986); Basu et al. (1996); Gustafson (1996). Theorem 3.1 establishes both the existence of the partial derivatives and their formula. Equation (3) is the partial derivative of the posterior mean with respect to the weights, while Equation (4) is that for the posterior standard deviation, with the understanding that the derivative is one-sided.

B.3. Estimating the influence for other QoI

Algorithm 1 describes the algorithm to estimate the influence of a quantity of interest $\phi(w)$ that satisfies Assumption B.1 and Assumption B.2.

[2] If w_n lies on the boundary, the partial derivative is understood to be one-sided.

Algorithm 1 Influence Estimate (EI)

Input: $\phi(w)$-defining constants c_1, c_2, Markov chain $(\beta^{(1)}, \ldots, \beta^{(S)})$
$m = \frac{1}{S}\sum_{s=1}^{S} g(\beta^{(s)})$, $k = \frac{1}{S}\sum_{s=1}^{S} g(\beta^{(s)})^2$
$v = k - m^2$
$\hat{\psi} = (0, 0, \ldots, 0)$ {N-dimensional vector}
for $n = 1$ **to** N **do**
  $a = \frac{1}{S}\sum_{s=1}^{S} g(\beta^{(s)}) L(d^{(n)} \mid \beta^{(s)})$
  $b = \frac{1}{S}\sum_{s=1}^{S} g(\beta^{(s)})^2 L(d^{(n)} \mid \beta^{(s)})$
  $u = \frac{1}{S}\sum_{s=1}^{S} L(d^{(n)} \mid \beta^{(s)})$
  $f = a - mu$ {Estimate of Equation (3)}
  $g = b - ku$
  $\hat{s} = (g - 2mf)/(\sqrt{v})$ {Estimate of Equation (4)}
  $\hat{\psi}_n = c_1 f + c_2 s$ {Estimate of ψ_n }
end for
Return: $\hat{\psi}$

Algorithm 2 Sum of Sorted Influence Estimate (SoSIE)

Input: $\phi(w)$-defining constants c_1, c_2, Markov chain $(\beta^{(1)}, \ldots, \beta^{(S)})$, fraction of data to drop α
$\hat{\psi} = \text{EI}(c_1, c_2, (\beta^{(1)}, \ldots, \beta^{(S)}))$
Find ranks $v_1, v_2, \ldots, v_N$ such that $\hat{\psi}_{v_1} \leq \hat{\psi}_{v_2} \leq \ldots \leq \hat{\psi}_{v_N}$
Find the smallest p such that $\hat{\psi}_{v_{p+1}} \geq 0$. If none exists, set p to N.
$\widehat{w}$ is the N-vector where $\widehat{w}_n = 1$ for $n \in \{v_1, \ldots, v_{\min(p, \lfloor N\alpha \rfloor)}\}$ and $\widehat{w}_n = 0$ otherwise
If $p \geq 1$, $\widehat{U} = \{d_{v_1}, \ldots, d_{v_{\min(p, \lfloor N\alpha \rfloor)}}\}$. Otherwise, $\widehat{U} = \emptyset$
$\widehat{\Delta} = -\sum_{m=1}^{\lfloor N\alpha \rfloor} \hat{\psi}_{v_m} \mathbb{I}\{\hat{\psi}_{v_m} < 0\}$
Return: $\widehat{\Delta}, \widehat{U}$

Algorithm 2 describes the algorithm to estimate the worst-case change in quantity of interest $\phi(w)$ that satisfies Assumption B.1 and Assumption B.2.

B.4. Confidence interval under exact sampling

For certain prior and likelihoods, we are able to draw exact Monte Carlo samples from the posterior distribution i.e. $(\beta^{(1)}, \ldots, \beta^{(S)})$ is an i.i.d. sample of size S drawn from the full-data posterior distribution. This happens for conjugate models (Diaconis and Ylvisaker, 1979), or for models in which convenient augmentation schemes have been discovered, such as Bayesian logistic regression with Polya-Gamma augmentation (Polson et al., 2013). Conceptually, $\widehat{\Delta}$ can be thought of as an estimator constructed from an i.i.d. sample. However, the sample in question is not the data $\{d^{(n)}\}_{n=1}^{N}$, but $(\beta^{(1)}, \ldots, \beta^{(S)})$. To highlight the dependence between $\widehat{\Delta}$ and $(\beta^{(1)}, \ldots, \beta^{(S)})$, we will use the notation $\widehat{\Delta}(\beta^{(1)}, \ldots, \beta^{(S)})$. The estimator $\widehat{\Delta}$ is a complex, non-smooth function of the sample: the act of taking the minimum across the estimated influences $\hat{\psi}_n$ is non-smooth. We do not attempt to prove distributional results for this estimator and use such results to quantify uncertainty. Instead, we appeal to the bootstrap (Efron, 1979), a general-purpose technique to quantify the sampling uncertainty of estimators.

Our confidence interval construction proceeds in three steps. First, we define the so-called *bootstrap distribution* of $\widehat{\Delta}$. Second, we approximate this distribution with an empirical distribution based on Monte Carlo draws. Finally, we use the range spanned by quantiles of this empirical distribution as our confidence interval for $\Delta(\alpha)$.

To define the bootstrap distribution, consider the empirical distribution of the sample $(\beta^{(1)}, \ldots, \beta^{(S)})$:

$$\frac{1}{S}\sum_{i=1}^{S} \delta_{\{\beta^{(i)}\}}(\cdot).$$

We denote one draw from this empirical distribution by $\beta^{*(s)}$. A *bootstrap sample* is a set of S draws:

$(\beta^{*(1)}, \beta^{*(1)}, \ldots, \beta^{*(S)})$. The bootstrap distribution of $\widehat{\Delta}$ is the distribution of $\widehat{\Delta}(\beta^{*(1)}, \beta^{*(1)}, \ldots, \beta^{*(S)})$, where the randomness is taken over the bootstrap sample but is conditional on the original sample $(\beta^{(1)}, \ldots, \beta^{(S)})$

Clearly, the bootstrap distribution is discrete with finite support. If we chose to, we can enumerate its support and compute its probability mass function, by enumerating all possible values a bootstrap sample can take. However, this is time-consuming. It suffices to approximate the bootstrap distribution with Monte Carlo draws. The draw $\widehat{\Delta}(\beta^{*(1)}, \beta^{*(1)}, \ldots, \beta^{*(S)})$ is abbreviated by $\widehat{\Delta}^*$: we generate a total number of B such draws. When B increases, the empirical distribution of $(\widehat{\Delta}_1^*, \widehat{\Delta}_2^*, \ldots, \widehat{\Delta}_B^*)$ becomes a better approximation of the bootstrap distribution. However, the computational cost scales up with B. In practice, B in the hundreds are commonplace: our numerical work uses $B = 200$.

We now define confidence intervals for $\Delta(\alpha)$. Each interval is parametrized by η, the nominal coverage level, which is valued in $(0, 1)$. We compute two quantiles of the empirical distribution over $(\widehat{\Delta}_1^*, \widehat{\Delta}_2^*, \ldots, \widehat{\Delta}_B^*)$, the $(1-\eta)/2$ and $(1+\eta)/2$ quantiles[3], and define the interval spanned by these two values as our confidence interval. By default, we set $\eta = 0.95$.

One limitation of our current work is that we do not make theoretical claims regarding the actual coverage of such confidence intervals. Although bootstrap confidence intervals can always be computed, whether the actual coverage matches the nominal coverage η depends on structural properties of the estimator and regularity conditions on the sample. To verify the quality of these confidence intervals, we turn to numerical simulation. We leave to future work the task of formulating reasonable assumptions and theoretically analyzing the actual coverage.

B.5. Confidence interval under general MCMC

In the previous section, we made the simplifying assumption that exact sampling were possible. We now lift this assumption and handle the case in which $(\beta^{(1)}, \ldots, \beta^{(S)})$ truly came from a Markov chain (such as the output of Hamiltonian Monte Carlo). This case is much more common in practice than the exact sampling case.

To construct confidence intervals, one idea is to use the previous section's construction without modification. In other words, apply the bootstrap to a non-i.i.d. sample: recall that the Markov chain states are not independent of each other. Theoretically, it is known that the bootstrap struggles on non-i.i.d. samples, for even simple estimators. For example, if the estimator in question is the sample mean and the draws exhibit positive autocorrelation, under mild regularity conditions, the bootstrap variance estimate seriously underestimates the true sampling variance, even in the limit of infinite sample size (Lahiri, 2003, Theorem 2.2). In our case, the bootstrap likely struggles on the sample means that are involved in the definition of $\widehat{\Delta}$: for instance, it is very common for some coordinate v that $(\beta_v^{(1)}, \beta_v^{(2)}, \ldots, \beta_v^{(S)})$ exhibits positive autocorrelation in practice. Therefore, we have reason to be pessimistic about the ability of bootstrap confidence intervals to adequately cover $\Delta(\alpha)$.

Fundamentally, the bootstrap fails in the non-i.i.d. case because the draws that form the bootstrap sample do not have any dependence, while the draws that form the original sample do. To improve upon the bootstrap, one option is to resample in a way that respects the original sample's dependence structure. We recognize that the sample in question, $(\beta^{(1)}, \ldots, \beta^{(S)})$, is a (multivariate) time series: we focus on methods that perform well under time series dependence. One such scheme is the non-overlapping block bootstrap (Lahiri, 2003; Carlstein, 1986).[4] The sample $(\beta^{(1)}, \ldots, \beta^{(S)})$ is divided up into a number of blocks: each block is a vector of contiguous draws. Let L be the number of elements in a block, and let $M := \lfloor S/L \rfloor$ denote the number of blocks. The m-th block is defined as

$$B_m := \left(\beta^{((m-1)L+1)}, \ldots, \beta^{(mL)}\right).$$

To generate one sample from the non-overlapping block bootstrap distribution, we first draw with replacement from the set of blocks M values: $B_1^*, \ldots, B_M^*$. Then, we write the elements of these drawn blocks in a contiguous series. For example, when $(\beta^{(1)}, \ldots, \beta^{(S)}) = (\beta^{(1)}, \beta^{(2)}, \beta^{(3)}, \beta^{(4)})$ and $L = 2$, the two blocks are $(\beta^{(1)}, \beta^{(2)})$, and $(\beta^{(3)}, \beta^{(4)})$. The set of possible samples from resampling include $(\beta^{(1)}, \beta^{(2)}, \beta^{(1)}, \beta^{(2)})$ and $(\beta^{(3)}, \beta^{(4)}, \beta^{(3)}, \beta^{(4)})$ but not $(\beta^{(1)}, \beta^{(3)}, \beta^{(1)}, \beta^{(3)})$.

The name "non-overlapping block bootstrap" comes from the fact that these blocks, viewed as sets, are disjoint from each other. While the name is needed in Lahiri (2003) to distinguish from other blocking rules, moving forward, as we only

[3] We use R's quantile() to compute the sample quantiles. When $(1+\eta)/2 \times B$ is not an integer, the $(1+\eta)/2$ quantile is defined by linearly interpolating the order statistics.

[4] The original paper, Carlstein (1986), did not use the term "non-overlapping block bootstrap" to describe the technique. The name comes from Lahiri (2003).

consider the above blocking rule, we will refer to the procedure as simply, block bootstrap. Intuitively, the block bootstrap sample is a good approximation of the original sample if the latter has short-term dependence: in such a case, the original sample itself can be thought of as the concatenation of smaller, i.i.d. subsamples, and the generation of a block bootstrap sample mimics that. In well-behaved probabilistic models with well-tuned algorithms, the MCMC draws can be expected to only have short-term dependence, and the block bootstrap is a good choice.

The block bootstrap has one hyperparameter: the block length L. We would like both L and M to be large: large L captures time series dependence at larger lags, and large M is close to having many i.i.d. subsamples. However, since their product is constrained to be S, the choice of L is a trade-off. In numerical studies, we set $L = 10$.

Our construction of confidence intervals for general MCMC proceeds identically to the previous section's construction, except for the step of generating the bootstrap sample: instead of drawing from the vanilla bootstrap, we draw from the block bootstrap. We will denote the endpoints of such an interval by $\Delta^{lb}(\alpha)$ (lower endpoint) and $\Delta^{ub}(\alpha)$ (upper endpoint).

Similar to the previous section, we do not make theoretical claims on the actual coverage of our block bootstrap confidence intervals: we verify the quality of the intervals through later numerical studies.

C. Theory

In this section, we theoretically quantify the approximation errors incurred by our methods. Namely, Appendix C.1 analyzes the error made by the first-order approximation, while Appendix C.2 analyzes the error made by using MCMC to estimate influences.

C.1. Accuracy of first-order approximation

In this section, we investigate the error incurred by replacing $\phi(w) - \phi(\mathbf{1}_N)$ with the Taylor series from Section 3.1. While the approximation applies to any model that satisfies Assumption 2.1, Assumption B.1, and Assumption B.2, our error analysis is limited to two models: a normal model and a normal means model. Their salient features are the following. Both are convenient to analyze and address the same statistical task: derive the population mean based on a finite sample $\{x^{(n)}\}_{n=1}^N$ where $x^{(n)} \in \mathbb{R}$. The normal means model is *hierarchical*: the observations are organized into disjoint groups. Each observation $d^{(n)}$ is $(x^{(n)}, g^{(n)})$, where $g^{(n)}$ is valued in $\{1, 2, \ldots, G\}$, with $g^{(n)} = g$ indicating that the n-th observation belongs to the g-th group. The normal model does not have this structure, as only $x^{(n)}$ is observed and used in modeling. We show that error in the normal model is qualitatively different from the error in the normal means model. Roughly speaking, the former depends on the ratio between the number of observations left out, $\lfloor N\alpha \rfloor$, and the total number of observations, N. Meanwhile, the later depends on three quantities a) $\lfloor N\alpha \rfloor$, b) the number of groups G and c) the number of observations in a group.

Before specializing to different models, we pin down the common notion of error. We define error to be the difference between $\phi(w) - \phi(\mathbf{1}_N)$ and $\sum_n (w_n - 1)\psi_n$. We mainly care when w encodes the full removal of certain observations and full inclusion of the remaining ones i.e. $w \in \{0, 1\}^N$. If we let q be the function that returns the zero indices of such a weight ($q(w) = \{n : w_n = 0\}$), then its inverse q^{-1} takes a set of observation indices ($I \subset \{1, 2, \ldots, N\}$) and produces a weight valued in $\{0, 1\}^N$. We reformulate the error as a function of I instead of w by replacing w with $q^{-1}(I)$ in the definition of error. This reformulation reads

$$\text{Err}_{\text{1st}}(I) = \phi(q^{-1}(I)) - \phi(\mathbf{1}_N) + \sum_{n \in I} \psi_n.$$

C.1.1. Normal model.

We detail the prior and likelihood of the normal model and the associated quantity of interest. The parameter of interest is the population mean μ. The likelihood of an observation is Gaussian with a known standard deviation σ. In other words, the n-th log-likelihood evaluated at μ is $L(d^{(n)} \mid \mu) = \frac{1}{2} \log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}[(x^{(n)})^2 - 2x^{(n)}\mu + \mu^2]$. We choose the uniform distribution over the real line as the prior for μ. The quantity of interest is the posterior mean of μ.

In this model, expectations under the weighted posterior have closed forms. We can derive an explicit expression for the error. To display the error, it is convenient to define the sample average of observations as a function of I: for any $I \subset \{1, 2, \ldots, N\}$, let $\bar{x}_I := (1/|I|) \sum_{n \in I} x^{(n)}$. The sample average of the whole dataset will be denote by $\bar{x}$.

Lemma C.1. *For the normal model, $Err_{1st}(I)$ is qual to*

$$\frac{|I|^2(\bar{x} - \bar{x}_I)}{N(N - |I|)}$$

We prove Lemma C.1 in Proof D.3. The error is a function of I through the a) the cardinality of the set $|I|$ and b) the difference between the whole dataset's sample mean, $\bar{x}$, and the sample mean for elements in I. Since the data is fixed, we can upper bound $|\bar{x} - \bar{x}_I|$ with the constant $2\|x\|_\infty$, where $\|x\|_\infty := \max_n |x^{(n)}|$. The rate at which the absolute value of the error goes to zero is $|I|^2$: as the ratio $|I|/N$ equals α, this means the error's absolute value goes to zero like α^2.

C.1.2. Normal means model.

We detail the prior and likelihood of the normal means model and the associated quantity of interest. The parameters of interest are the population mean μ and the group means $\theta = (\theta_1, \theta_2, \ldots, \theta_G)$. Observations in group g are modeled as Gaussian centered at the group mean θ_g with a known standard deviation σ. In other words, the n-th log-likelihood is $L(d^{(n)} \mid \mu, \theta) = \frac{1}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}[(x^{(n)})^2 - 2x^{(n)}\theta_{g^{(n)}} + \theta^2_{g^{(n)}}]$. The prior over (μ, θ) is the following. We choose the uniform distribution over the real line as the prior for μ. Conditioned on μ, the group means are Gaussian centered at μ, with a known standard deviation τ. The quantity of interest is the posterior mean of μ.

This model, like the normal model, has closed-form posterior expectations. Before displaying the exact formula for the error $\text{Err}_{1\text{st}}(I)$, we need to describe the weighted posterior in more detail. For each group g, we define three functions of w:

$$N_g(w) := \sum_{n:g^{(n)}=g} w_n, M_g(w) := \frac{\sum_{n:g^{(n)}=g} w_n x^{(n)}}{N_g(w)}, \Lambda_g(w) := \left(\frac{\sigma^2}{N_g(w)} + \tau^2\right)^{-1}.$$

While $N_g(w)$ sums up the weights of observations in group g, $M_g(w)$ is the weighted average of observations in this group, and $\Lambda_g(w)$ will be used to weigh $M_g(w)$ in forming the posterior mean of μ. Proof D.4 shows that $\mathbb{E}_w\mu$ is equal to

$$\frac{\sum_{g=1}^G \Lambda_g(w) M_g(w)}{\sum_{g=1}^G \Lambda_g(w)}.$$

To avoid writing $\sum_{g=1}^G \Lambda_g(w)$, we define $\Lambda(w) := \sum_{g=1}^G \Lambda_g(w)$. To lighten notation, for expectations under the original posterior, we write μ^* instead of $\mathbb{E}_{\mathbf{1}_N}\mu$ and N_g^* instead of $N_g(\mathbf{1}_N)$. The same shorthand applies to $N_g(\mathbf{1}_N)$, $M_g(\mathbf{1}_N)$, $\Lambda_g(\mathbf{1}_N)$ and $\Lambda(\mathbf{1}_N)$. In words, μ^* is the posterior mean of μ under the full-data posterior, N_g^* is the number of observations in group g of the original dataset, and so on. We also utilize the $\bar{x}_I$ and $\bar{x}$ notations defined the normal model section.

The error in the normal means model is given in the following lemma.

Lemma C.2. *In the normal means model, let the index set I be such that there exists $k \in \{1, 2, \ldots, G\}$ such that $g^{(n)} = k$ for all $n \in I$. Define*

$$F(I) := \frac{|I|^2}{N_k^*[N_k^* - |I|]}(M_k^* - \bar{x}_I) + \frac{|I|}{N_k^*}\frac{\sigma^2\Lambda_k^*}{N_k^*}(\mu^* - M_k^*),$$
$$E(I) := \frac{|I|}{N_k^*[N_k^* - |I|]}\sigma^2\Lambda_k(q^{-1}(I))\Lambda_k^*.$$

Then, $Err_{1st}(I)$ is equal to

$$\frac{\Lambda_k(q^{-1}(I))}{\Lambda^*}F(I) + \frac{\left(\sum_{g\neq k}\Lambda_g^*(M_g^* - M_k(q^{-1}(I)))\right)}{\Lambda^*\Lambda(q^{-1}(I))}E(I).$$

We prove Lemma C.2 in Proof D.4. The constraint where all observations in I belong to the same group k is made out of convenience: we can derive the error without this constraint, but the formula will be much more complicated.

A corrolary of Lemma C.2 is that the absolute value of the error behaves like $|I|^2/(G|N_k^*|^2)$.

Corollary C.1. *In the normal means model, for all groups g, assume that $N_g^* \geq \sigma^2/\tau^2$. Let the index set I be such that there exists $k \in \{1, 2, \ldots, G\}$ such that $g^{(n)} = k$ for all $n \in I$. For this k, assume that $N_k^* - |I| \geq \sigma^2/\tau^2$. Then,*

$$|Err_{1st}(I)| \leq C(\|x\|_\infty, \sigma, \tau)\frac{1}{G}\frac{|I|^2}{|N_k^*|^2}.$$

where $C(\|x\|_\infty, \sigma, \tau)$ is a constant that only depends on $\|x\|_\infty$, σ, and τ.

We prove Corollary C.1 in the Proof D.5. In addition to the assumptions Lemma C.2, the corrolary assumes that the number of observations in each group is not too small, and that after removing I, group k still has enough observations. This condition allows us to approximate Λ_k^* and $\Lambda_g(q^{-1}(I))$ with a constant. The factor $\|x\|_\infty$ in the bound comes from upper bounding $|M_g^* - M_k(q^{-1}(I))|$ by $2\max_{n=1}^N |x^{(n)}|$.

For two reasons, we conjecture that similiar qualitative differences also appear in the comparison between more complicated hierarchical and non-hierarchical models The fundamental task of estimating the population mean is embedded in many other statistical tasks, such as regression. In addition, the group structure imposed by the normal means model is also found in practically relevant hierarchical models.

C.2. Estimator properties

Recall from Section 3.3 that one concern regarding the quality of $\widehat{\Delta}$ is the $(\beta^{(1)}, \ldots, \beta^{(S)})$-induced sampling uncertainty. Theoretically analyzing this uncertainty is difficult, with one obstacle being that $\widehat{\Delta}$ is a non-smooth function of $(\beta^{(1)}, \ldots, \beta^{(S)})$. In this section, we settle for the easier goal of analyzing the sampling uncertainty of the influence estimates $\hat{\psi}_n$. We expect such theoretical characterizations to play a role in the eventual theoretical characterizations of $\widehat{\Delta}$, but we leave this step to future work.

In this analysis, we make more restrictive assumptions than those needed for Theorem 3.1 to hold. We assume that the sample $(\beta^{(1)}, \ldots, \beta^{(S)})$ comes from exact sampling: the independence across draws makes it easier to analyze sampling uncertainty. We focus on the quantity of interest equaling the posterior mean ($c_1 = 1, c_2 = 0$ in the sense of Assumption B.1): the scaling $c_1 = 1$ for the posterior mean is made out of convenience, and a smiliar analysis can be conducted when $c_2 \neq 0$, but we omit it for brevity. Finally, we need more stringent moment conditions than Assumption B.2.

Assumption C.1. *The functions $|g(\beta)^2 L(d^{(i)} \mid \beta)L(d^{(j)} \mid \beta)|$ (across i, j) have finite expectation under the full-data posterior.*

This moment condition guarantees that the sample covariance of $g(\beta)$ and $L(d^{(i)} \mid \beta)$ has finite variance under the full-data posterior: it plays the same as role as finite kurtosis in proofs sample variance consistency

With the assumptions in place, we begin by showing that the sampling uncertainty of $\hat{\psi}_n$ goes to zero in the limit of $S \to \infty$.

Lemma C.3. *Assume Assumption 2.1, Assumption B.1, Assumption B.2, Assumption C.1 holds. Let $\hat{\psi}$ be output of Algorithm 1 for $c_1 = 1$, $c_2 = 0$ and $(\beta^{(1)}, \ldots, \beta^{(S)})$ being an i.i.d. sample. Then, there exists a constant C such that for all n, for all S, Var$(\hat{\psi}_n) \leq C/S$.*

We prove Lemma C.3 in Proof D.7. That the variance of individual $\hat{\psi}_n$ goes to zero at the rate of $1/S$ is not surprising: $\hat{\psi}_n$ is a sample covariance, after all.

We use Lemma C.3 to show consistency of different estimators.

Theorem C.1. *Assume Assumption 2.1, Assumption B.1, Assumption B.2, and Assumption C.1 holds Let $\hat{\psi}$ be output of Algorithm 1 for $c_1 = 1$, $c_2 = 0$ and $(\beta^{(1)}, \ldots, \beta^{(S)})$ being an i.i.d. sample. Then $\max_{n=1}^N |\hat{\psi}_n - \psi_n|$ converges in probability to 0 in the limit $S \to \infty$, and $\widehat{\Delta}$ converges in probability to $\Delta(\alpha)$ in the limit $S \to \infty$.*

We prove Theorem C.1 in Proof D.8. Our theorem states that the vector $\hat{\psi}$ is a consistent estimator for the vector ψ and $\widehat{\Delta}$ is a consistent estimator for $\Delta(\alpha)$.

Not only is $\hat{\psi}$ consistent in estimating ψ, it is also asymptotically normal.

Theorem C.2. *Assume Assumption 2.1, Assumption B.1, Assumption B.2, and Assumption C.1 holds. Let $\hat{\psi}$ be output of Algorithm 1 for $c_1 = 1$, $c_2 = 0$ and $(\beta^{(1)}, \ldots, \beta^{(S)})$ being an i.i.d. sample. Then $\sqrt{S}(\hat{\psi} - \psi)$ converges in distribution to $N(\mathbf{0}_N, \Sigma)$ where Σ is the $N \times N$ matrix whose (i, j) entry, $\Sigma_{i,j}$, is the covariance between*

$(g(\beta) - \mathbb{E}_{\mathbf{1}_N} g(\beta)) \left(L(d^{(i)} \mid \beta) - \mathbb{E}_{\mathbf{1}_N} L(d^{(i)} \mid \beta)\right)$ *and* $(g(\beta) - \mathbb{E}_{\mathbf{1}_N} g(\beta)) \left(L(d^{(j)} \mid \beta) - \mathbb{E}_{\mathbf{1}_N} L(d^{(j)} \mid \beta)\right)$*, taken under the full-data posterior.*

We prove Theorem C.2 in Proof D.9. Heuristically, for each n, the distribution of $\hat{\psi}_n$ is the Gaussian centered at ψ_n, with standard deviation $\sqrt{\Sigma_{n,n}}/\sqrt{S}$.

C.2.1. Normal model with unknown precision.

While $\sqrt{\Sigma_{n,n}}/\sqrt{S}$ eventually goes to zero, for finite S, this standard deviation can be large, making $\hat{\psi}_n$ an imprecise estimate of ψ_n. To illustrate this phenomenon, we will derive $\Sigma_{n,n}$ in the context of a simple probabilistic model: a normal model with unknown precision.

We first introduce the model and the associated quantity of interest. The data is a set of N real values: $d^{(n)} = x^{(n)}$, where $x^{(n)} \in \mathbb{R}$. The parameters of interest are the mean μ and the precision τ of the population. The log-likelihood of an observation based on μ and τ is Gaussian: $\frac{1}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\tau[(x^{(n)})^2 - 2x^{(n)}\mu + \mu^2]$. The prior is chosen to be the following. μ is distributed from uniform over the real line, and τ is distributed from a gamma distribution. The quantity of interest is the posterior mean of μ.

For this probabilistic model, the assumptions of Theorem C.2 are satisfied. We show that the variance $\Sigma_{n,n}$ behaves like a quartic function of the observation $x^{(n)}$.

Lemma C.4. *In the normal-gamma model, there exists constants D_1, D_2, and D_3, where $D_1 > 0$, such that for all n, $\Sigma_{n,n}$ is equal to $D_1(x^{(n)} - \bar{x})^4 + D_2(x^{(n)} - \bar{x})^2 + D_3$.*

We prove Lemma C.4 in Proof D.10. D_1, D_2, D_3 are based on the posterior expectations: for instance, the proof shows that $D_1 = \frac{\mathbb{E}_{\mathbf{1}_N}[\tau^{-1}(\tau - \mathbb{E}_{\mathbf{1}_N}\tau)^2]}{4N}$. It is easy to show that for the normal-gamma model,

$$\mathrm{Cov}_{\mathbf{1}_N}(\mu, L(d^{(n)} \mid \mu, \tau)) = \frac{x^{(n)} - \bar{x}}{N}.$$

Hence, while the mean of $\hat{\psi}_n$ behaves like a linear function of $x^{(n)} - \bar{x}$, its standard deviation behaves like a quadratic function of $x^{(n)} - \bar{x}$. In other words, the more influence an observation has, the harder it is to accurately determine its influence!

D. Proofs

D.1. Taylor series proofs

Proof D.1 (Proof of Theorem 3.1)**.** *Theorem 3.1 is a special case of Theorem B.1. See Proof D.2 for the proof of the latter, which implies the proof of the former.*

Proof D.2 (Proof of Theorem B.1)**.** *At a high level, we rely on Fleming (1977, Chapter 5.12, Theorem 5.9) to interchange integration and differentiation.*

Although the theorem statement does not explicitly mention the normalizer, to show that the quantity of interest is continuously differentiable and compute partial derivatives, it is necessary to show that the normalizer is continuously differentiable and compute partial derivatives. To do so, we verify the following conditions on the integrand defining $Z(w)$:

1. *For any β, the mapping $w \mapsto p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)$ is continously differentiable.*
2. *There exists a Lebesgue integrable function f_1 such that for all $w \in \{w \in [0,1]^N : \max_n w_n \geq \delta\}$, $p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right) \leq f_1(\beta)$.*
3. *For each n, there exists a Lebesgue integrable function f_2 such that for all $w \in \{w \in [0,1]^N : \max_n w_n \geq \delta\}$, $\left|\frac{\partial}{\partial w_n} p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right| \leq f_2(\beta)$.*

The first condition is clearly satisfied. To construct f_1 that satisfies the second condition, we partition the parameter space $\mathbb{R}^V$ into a finite number of disjoint sets. To index these sets, we use a subset of $\{1, 2, \ldots, N\}$. If the indexing subset were

$I = \{n_1, n_2, \ldots, n_M\}$, the corresponding element of the partition is

$$B_I := \{\beta \in \mathbb{R}^V : \forall n \in I, L(d^{(n)} \mid \beta) \geq 0\}. \tag{5}$$

This partition allows us to upper bound the integrand with a function that is independent of w. Suppose $\beta \in B_I, I \neq \emptyset$. The maximum $\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)$ is attained by setting $w_n = 1$ for all $n \in I$ and $w_n = 0$ for all $n \notin I$. Suppose $\beta \in B_\emptyset$. As $L(d^{(n)} \mid \beta) < 0$ for all $1 \leq n \leq N$, and we are constrained by $\max_n w_n \geq \delta$, the maximum of $\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)$ is attained by setting $w_n = \delta$ for $\arg\max_n L(d^{(n)} \mid \beta)$ and $w_n = 0$ for all other n. In short, our envelope function is

$$f_1(\beta) := \begin{cases} p(\beta) \prod_{n \in I} \exp(L(d^{(n)} \mid \beta)) & \text{if } \beta \in B_I, I \neq \emptyset. \\ p(\beta) \left(\max_{n=1}^N \exp(\delta L(d^{(n)} \mid \beta))\right) & \text{if } \beta \in B_\emptyset. \end{cases}$$

The last step is to show f_1 is integrable. It suffices to show that the integral of f_1 on each B_I is finite. On $B_\emptyset$, integrating $p(\beta)\left(\exp(\delta L(d^{(n)} \mid \beta))\right)$ over $B_\emptyset$ is clearly finite: by Assumption 2.1, the integral of $p(\beta)\exp(\delta L(d^{(n)} \mid \beta))$ over $\mathbb{R}^V$ is finite, and $B_\emptyset$ is a subset of $\mathbb{R}^V$. As $f_1(\beta)$ is the maximum of a finite number of integrable functions, it is integrable. Similarly, the integral of f_1 over B_I where $I \neq \emptyset$ is atmost the integral of $p(\beta)\prod_{n\in I}\exp(L(d^{(n)} \mid \beta))$ over $\mathbb{R}^V$, which is finite by Assumption 2.1. To construct f_2 that satisfies the third condition, we use the same partition of $\mathbb{R}^V$, and the envelope function is $f_2(\beta) := L(d^{(n)} \mid \beta) f_1(\beta)$, since the partial derivative of the weighted log probability is clearly the product of the n-th log likelihood and the weighted log probability. The integrability of f_2 follows from Assumption B.2's guarantee that the expectation of $|L(d^{(n)} \mid \beta)|$ is finite under different weighted posteriors. In all, we can interchange integration with differentiation, and the partial derivatives are

$$\frac{\partial Z(w)}{\partial w_n} = Z(w) \times \mathbb{E}_w\left[L(d^{(n)} \mid \beta)\right].$$

We move on to prove that $\mathbb{E}_w g(\beta)$ is continuously differentiable and find its partial derivatives. The conditions on $g(\beta)\frac{1}{Z(w)}p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)$ that we wish to check are:

1. *For any β, the mapping $w \mapsto g(\beta)\frac{1}{Z(w)}p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)$ is continously differentiable.*
2. *There exists a Lebesgue integrable function f_1 such that for all $w \in \{w \in [0,1]^N : \max_n w_n \geq \delta\}$, $\left|g(\beta)\frac{1}{Z(w)}p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right| \leq f_3(\beta)$.*
3. *For each n, there exists a Lebesgue integrable function f_4 such that for all $w \in \{w \in [0,1]^N : \max_n w_n \geq \delta\}$, $\left|\frac{\partial}{\partial w_n} g(\beta)\frac{1}{Z(w)}p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right| \leq f_4(\beta)$.*

We have already proven that $Z(w)$ is continuously differentiable: hence, there is nothing to do for the first condition. It is straightforward to use Assumption B.2 and check that the second condition is satisfied by the function $f_3(\beta) := \frac{1}{Z(w)}g(\beta)f_1(\beta)$, and the third condition is satisfied by $f_4(\beta) := \frac{1}{Z(w)}g(\beta)L(d^{(n)} \mid \beta)f_1(\beta)$. Hence, we can interchange integration with differentiation. The partial derivatives of $\mathbb{E}_w g(\beta)$ is equal to tthe sume of two integrals. The first part is

$$\begin{aligned}
&\int \left(\frac{\partial Z(w)^{-1}}{\partial w_n} g(\beta)p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right) d\beta \\
&= -\left(\mathbb{E}_w\left[L(d^{(n)} \mid \beta)\right]\right)\int \left(\frac{1}{Z(w)} g(\beta)p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right) d\beta \\
&= -\mathbb{E}_w\left[L(d^{(n)} \mid \beta)\right] \times \mathbb{E}_w\left[g(\beta)\right].
\end{aligned}$$

The second part is

$$\int \left(\frac{1}{Z(w)} g(\beta)L(d^{(n)} \mid \beta)p(\beta)\exp\left(\sum_{n=1}^N w_n L(d^{(n)} \mid \beta)\right)\right) d\beta = \mathbb{E}_w\left[g(\beta)L(d^{(n)} \mid \beta)\right].$$

Putting two parts together, the partial derivative is equal to a covariance

$$\frac{\partial \mathbb{E}_w g(\beta)}{\partial w_n} = Cov_w\left[g(\beta), L(d^{(n)} \mid \beta)\right].$$

The proof that $\mathbb{E}_w g(\beta)^2$ is continuously differentiable is similar to that for $\mathbb{E}_w g(\beta)$. The partial derivative is

$$\frac{\partial [\mathbb{E}_w g(\beta)^2]}{\partial w_n} = Cov_w\left[g(\beta)^2, L(d^{(n)} \mid \beta)\right].$$

Since the posterior standard deviation is a differentiably continuous function of the mean and second moment, it is also differentiably continuous. The partial derivative of the posterior standard deviation is a simple application of the chain rule, and we omit the proof for brevity.

D.2. First-order accuracy proofs

Proof D.3 (Proof of Lemma C.1)**.** *Our proof finds exact formulas for the posterior mean and the partial derivatives of the posterior mean with respect to w_n. Then, we take the difference between the posterior mean and its Taylor series.*

In the normal model, the total log probability at w is equal to

$$\begin{aligned}
&\sum_{n=1}^{N} w_n \left[\frac{1}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}[(x^{(n)})^2 - 2x^{(n)}\mu + \mu^2]\right] \\
&= -\left(\frac{\sum_{n=1}^{N} w_n}{2\sigma^2}\right)\left(\mu - \frac{\sum_{n=1}^{N} w_n x^{(n)}}{\sum_{n=1}^{N} w_n}\right)^2 + C,
\end{aligned}$$

where C is a constant that does not depend on μ. Hence, the distribution of μ under w is normal with mean $(\sum_{n=1}^{N} w_n x^{(n)})/(\sum_{n=1}^{N} w_n)$ and precision $(\sum_{n=1}^{N} w_n)/(\sigma^2)$. The partial derivative of the posterior mean with respect to w_n is

$$\frac{x^{(n)}(\sum_{n=1}^{N} w_n) - (\sum_{n=1}^{N} w_n x^{(n)})}{(\sum_{n=1}^{N} w_n)^2}.$$

Plugging in $w = \mathbf{1}_N$, we have that ψ_n is equal to $(x^{(n)} - \bar{x})/N$.

After removing the index set I, the actual posterior mean is

$$\frac{N\bar{x} - |I|\bar{x}_I}{N - |I|},$$

while the Taylor series approximation is

$$\bar{x} - \sum_{n \in I} \frac{x^{(n)} - \bar{x}}{N} = \frac{N\bar{x} + |I|(\bar{x} - \bar{x}_I)}{N}.$$

The difference between the actual posterior mean and its approximation is as in the statement of the lemma.

Proof D.4 (Proof of Lemma C.2)**.** *Similar to the proof of Lemma C.1, we first find exact formulas for the posterior mean and its Taylor series.*

In the normal means model, the total log probability at w is

$$\begin{aligned}
&\sum_{g=1}^{G}\left[\log\left(\frac{1}{2\pi\tau^2}\right) - \frac{1}{2\tau^2}(\theta_g - \mu)^2\right] \\
&+ \sum_{n=1}^{N} w_n \left\{\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\left[(x^{(n)})^2 - 2x^{(n)}\theta_{g^{(n)}} + \theta_{g^{(n)}}^2\right]\right\}.
\end{aligned}$$

By completing the squares, we know that

- *The distribution of μ is normal:*

$$N\left(\frac{\sum_{g=1}^{G}\Lambda_g(w)M_g(w)}{\sum_{g=1}^{G}\Lambda_g(w)}, \frac{1}{\sum_{g=1}^{G}\Lambda_g(w)}\right)$$

- *Condition on μ, the group means are independent normals:*

$$\theta_g \mid \mu \sim N\left(\frac{\mu/\tau^2 + [N_g(w)M_g(w)]/\sigma^2}{1/\tau^2 + N_g(w)/\sigma^2}, \frac{1}{1/\tau^2 + N_g(w)/\sigma^2}\right).$$

To express the partial derivative of the posterior mean of μ with respect to w_n, it is helpful to define the following "intermediate" value between $\mathbb{E}_w\mu$ and $\mathbb{E}_w\theta_g$:

$$\tilde{\mu}_g(w) := \frac{M_g(w)N_g(w)/\sigma^2 + \mathbb{E}_w\mu/\tau^2}{N_g(w)/\sigma^2 + 1/\tau^2}.$$

In addition, we need the partial derivatives of the functions N_g, Λ_g, and M_g.

$$\frac{\partial N_g}{\partial w_n} = \begin{cases} 0 & \text{if } g \neq g^{(n)} \\ 1 & \text{if } g = g^{(n)} \end{cases},$$

$$\frac{\partial M_g}{\partial w_n} = \begin{cases} 0 & \text{if } g \neq g^{(n)} \\ \frac{x^{(n)} - M_g(w)}{N_g(w)} & \text{if } g = g^{(n)} \end{cases},$$

$$\frac{\partial \Lambda_g}{\partial w_n} = \begin{cases} 0 & \text{if } g \neq g^{(n)} \\ \sigma^2 \frac{\Lambda_g(w)^2}{N_g(w)^2} & \text{if } g = g^{(n)} \end{cases}.$$

If n is in the k-th group, the partial derivative of the posterior mean with respect to w_n is

$$\frac{1}{\Lambda(w)}\frac{1}{\sigma^2 + \tau^2 N_k(w)}\left(x^{(n)} - \tilde{\mu}_k(w)\right).$$

After removing only observations from the k-th group, the actual posterior mean is

$$\frac{\Lambda_k(q^{-1}(I))M_k(q^{-1}(I)) + \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)M_g(\mathbf{1}_N)}{\Lambda_k(q^{-1}(I)) + \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)}.$$

Between $w = q^{-1}(I)$ and $w = \mathbf{1}_N$, the N_g, M_g, Λ_g functions do not change for $g \neq k$. The Taylor series approximation of the posterior mean is

$$\frac{\Lambda_k(\mathbf{1}_N)\left[M_k(\mathbf{1}_N) + \sum_{n\in I}\left(\tilde{\mu}_k(\mathbf{1}_N) - x^{(n)}\right)/N_k(\mathbf{1}_N)\right] + \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)M_g(\mathbf{1}_N)}{\Lambda_k(\mathbf{1}_N) + \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)}.$$

If we denote

$$\begin{aligned} &A_1 := \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)M_g(\mathbf{1}_N), A_2 := \sum_{g\neq k}\Lambda_g(\mathbf{1}_N) \\ &B_1 := \Lambda_k(q^{-1}(I))M_k(q^{-1}(I)), B_2 := \Lambda_k(q^{-1}(I)) \\ &C_1 := \Lambda_k(\mathbf{1}_N)\left[M_k(\mathbf{1}_N) + \sum_{n\in I}\left(\tilde{\mu}_k(\mathbf{1}_N) - x^{(n)}\right)/N_k(\mathbf{1}_N)\right], C_2 := \Lambda_k(\mathbf{1}_N) \end{aligned},$$

then $Err_{1st}(I)$ is equal to $(A_1 + B_1)/(A_2 + B_2) - (A_1 + C_1)/(A_2 + C_2)$. The last equation is equal to

$$\frac{A_2(B_1 - C_1) + A_1(C_2 - B_2) + (B_1C_2 - C_1B_2)}{(A_2 + B_2)(A_2 + C_2)}.$$

We analyze the differences $C_2 - B_2$, $B_1C_2 - C_1B_2$, and $B_1 - C_1$ separately.

$C_2 - B_2$. This difference is

$$\frac{1}{\sigma^2/N_k(\mathbf{1}_N) + \tau^2} - \frac{1}{\sigma^2/N_k(q^{-1}(I)) + \tau^2}.$$

Since we remove $|I|$ from group k, $N_k(q^{-1}(I)) = N_k(\mathbf{1}_N) - |I|$. Hence, the difference $C_2 - B_2$ is

$$\sigma^2\Lambda_k(\mathbf{1}_N)\Lambda_k(q^{-1}(I))\frac{|I|}{N_k(\mathbf{1}_N)(N_k(\mathbf{1}_N) - |I|)},$$

which is exactly the $E(I)$ mentiond in the lemma statement.

$B_1C_2 - C_1B_2$. The difference is

$$\Lambda_k(\mathbf{1}_N)\Lambda_k(q^{-1}(I))\left\{M_k(q^{-1}(I)) - M_k(\mathbf{1}_N) - \frac{\sum_{n\in I}[\tilde{\mu}_k(\mathbf{1}_N) - x^{(n)}]}{N_k(\mathbf{1}_N)}\right\}.$$

We analyze the term in the curly brackets. It is equal to

$$\left\{M_k(q^{-1}(I)) - M_k(\mathbf{1}_N) - \frac{\sum_{n\in I}[M_k(\mathbf{1}_N) - x^{(n)}]}{N_k(\mathbf{1}_N)}\right\} + \sum_{n\in I}\left(\frac{M_k(\mathbf{1}_N) - \tilde{\mu}_k(\mathbf{1}_N)}{N_k(\mathbf{1}_N)}\right)$$

The left term is equal to

$$\frac{|I|^2(M_k(\mathbf{1}_N) - \bar{x}_I)}{N_k(\mathbf{1}_N)[N_k(\mathbf{1}_N) - |I|]}.$$

The right term is equal to

$$\frac{|I|}{N_k(\mathbf{1}_N)}\frac{\sigma^2\Lambda_k(\mathbf{1}_N)}{N_k(\mathbf{1}_N)}(\mathbb{E}_{\mathbf{1}_N}\mu - M_k(\mathbf{1}_N)).$$

The sum of the two terms is exactly $F(I)$ mentioned in the lemma statement. Overall, the difference $B_1C_2 - C_1B_2$ is equal to $\Lambda_k(\mathbf{1}_N)\Lambda_k(q^{-1}(I))F(I)$.

$B_1 - C_1$. If we introduce $D := \Lambda_k(\mathbf{1}_N)M_k(q^{-1}(I))$, then the difference $B_1 - C_1$ is equal to $(B_1 - D) + (D - C_1)$. The former term is

$$M_k(q^{-1}(I))(B_2 - C_2) = -M_k(q^{-1}(I))E(I).$$

The later term is

$$\Lambda_k(\mathbf{1}_N)\left\{M_k(q^{-1}(I)) - M_k(\mathbf{1}_N) - \frac{\sum_{n\in I}[\tilde{\mu}_k(\mathbf{1}_N) - x^{(n)}]}{N_k(\mathbf{1}_N)}\right\}.$$

We already know that the term in the curly brackets is equal to $F(I)$. Hence $B_1 - C_1$ is equal to $\Lambda_k(\mathbf{1}_N)F(I) - M_k(q^{-1}(I))E(I)$.

With the differences $C_2 - B_2$, $B_1C_2 - C_1B_2$, and $B_1 - C_1$, we can now state the final form of $Err_{1st}(I)$. The final numerator is

$$\left[\Lambda_k(q^{-1}(I)) + \sum_{g\neq k}\Lambda_g(\mathbf{1}_N)\right]\Lambda_k(\mathbf{1}_N)F(I)$$
$$+\left[\sum_{g\neq k}\Lambda_g(\mathbf{1}_N)M_g(\mathbf{1}_N) - M_k(q^{-1}(I))\sum_{g\neq k}\Lambda_g(\mathbf{1}_N)\right]E(I)$$

Divide this by the denominator $\left[\sum_g \Lambda_g(\mathbf{1}_N)\right]\left[\sum_g \Lambda_g(q^{-1}(I))\right]$, we have proven the lemma.

Proof D.5 (Proof of Corollary C.1)**.** *Under the assumption that $N_g^* \geq \sigma^2/\tau^2$, we have that $\Lambda_g(\mathbf{1}_N) \in \left[\frac{1}{2\tau^2}, \frac{1}{\tau^2}\right]$. Since $M_k^* - |I| \geq \sigma^2/\tau^2$, it is also true that $\Lambda_k(q^{-1}(I)) \in \left[\frac{1}{2\tau^2}, \frac{1}{\tau^2}\right]$.*

Because of Lemma C.2, an upper bound on $Err_{1st}(I)$ is

$$\frac{\Lambda_k(q^{-1}(I))}{\Lambda^*}\,|F(I)| + \left|\frac{\left(\sum_{g\neq k}\Lambda_g^*(M_g^* - M_k(q^{-1}(I)))\right)}{\Lambda^*\Lambda(q^{-1}(I))}\right|\,|E(I)|\,.$$

The fraction $\Lambda_k(q^{-1}(I))/\Lambda^*$ is at most $(\frac{1}{\tau^2})/\left(G\frac{1}{2\tau^2}\right)$, which is equal to $2/G$. The absolute value $|F(I)|$ is at most

$$\frac{2|I|^2\|x\|_\infty}{(N_k^*)^2} + \frac{2|I|\|x\|_\infty(\sigma^2/\tau^2)}{(N_k^*)^2} \leq \frac{2|I|^2\|x\|_\infty(\sigma^2/\tau^2+1)}{(N_k^*)^2}$$

The absolute value

$$\left|\frac{\left(\sum_{g\neq k}\Lambda_g^*(M_g^* - M_k(q^{-1}(I)))\right)}{\Lambda^*\Lambda(q^{-1}(I))}\right|$$

is at most

$$\frac{G(1/\tau^2)2\|x\|_\infty}{G^2(1/2\tau^2)} \leq \frac{4\|x\|_\infty}{G}.$$

Finally, the absolute value $|E(I)|$ is at most

$$\frac{|I|(\sigma^2/(4\tau^4))}{(N_k^*)^2} \leq \frac{|I|^2(\sigma^2/(4\tau^4))}{(N_k^*)^2}.$$

In all, the constant $C(\|x\|_\infty,\sigma,\tau)$ in the corollary's statement is

$$\|x\|_\infty\left(4(\sigma^2/\tau^2+1)+\sigma^2/\tau^4\right).$$

D.3. Consistency and asymptotic normality proofs

The following lemma, on covariance between sample covariances under i.i.d. sampling, will be useful for later proofs.

Lemma D.1. *Suppose we have S i.i.d. draws $(A^{(s)},B^{(s)},C^{(s)})_{s=1}^S$. Let f_1 be the (biased) sample covariance between the A's and the B's. Let f_2 be the (biased) sample covariance between the A's and C's. In other words,*

$$f_1 := \left(\frac{1}{S}\sum_{s=1}^S A^{(s)}B^{(s)}\right) - \left(\frac{1}{S}\sum_{s=1}^S A^{(s)}\right)\left(\frac{1}{S}\sum_{s=1}^S B^{(s)}\right),$$
$$f_2 := \left(\frac{1}{S}\sum_{s=1}^S A^{(s)}C^{(s)}\right) - \left(\frac{1}{S}\sum_{s=1}^S A^{(s)}\right)\left(\frac{1}{S}\sum_{s=1}^S C^{(s)}\right).$$

Suppose that the following are finite: $\mathbb{E}[(A-\mathbb{E}[A])^2(B-\mathbb{E}[B])(C-\mathbb{E}[C])]$, $Cov(B,C)$, $Var(A)$, $Cov(A,B)$, $Cov(A,C)$. Then, the covariance of f_1 and f_2 is equal to

$$\frac{(S-1)^2}{S^3}\mathbb{E}[(A-\mathbb{E}[A])^2(B-\mathbb{E}[B])(C-\mathbb{E}[C])]$$
$$+\frac{S-1}{S^3}Cov(B,C)Var(A) - \frac{(S-1)(S-2)}{S^3}Cov(A,B)Cov(A,C).$$

Proof D.6 (Proof of Lemma D.1)**.** *It suffices to prove the lemma in the case where $\mathbb{E}[A]=\mathbb{E}[B]=\mathbb{E}[C]=0$. Otherwise, we can subtract the population mean from the random variable: the value of f_1 and f_2 would not change (since covariance is invariant to constant additive changes). In otherwords, we want to show that the covariance between f_1 and f_2 is equal to*

$$\frac{(S-1)^2}{S^3}\mathbb{E}[A^2BC] + \frac{S-1}{S^3}\mathbb{E}[BC]\mathbb{E}[A^2] - \frac{(S-1)(S-2)}{S^3}\mathbb{E}[AB]\mathbb{E}[AC]. \tag{6}$$

Since f_1 is the biased sample covariance, $\mathbb{E}f_1 = \frac{S-1}{S}\mathbb{E}[AB]$. Similarly, $\mathbb{E}f_2 = \frac{S-1}{S}\mathbb{E}[AC]$. To compute Cov$(f_1, f_2)$, we only need an expression for $\mathbb{E}[f_1 f_2]$. The product $f_1 f_2$ is equal to the sum of D_1, D_2, D_3, D_4 where:

$$
\begin{aligned}
D_1 &:= -\left(\frac{1}{S}\sum_s A^{(s)}B^{(s)}\right)\left(\frac{1}{S}\sum_s A^{(s)}\right)\left(\frac{1}{S}\sum_s C^{(s)}\right),\\
D_2 &:= \left(\frac{1}{S}\sum_s A^{(s)}\right)^2\left(\frac{1}{S}\sum_s B^{(s)}\right)\left(\frac{1}{S}\sum_s C^{(s)}\right),\\
D_3 &:= -\left(\frac{1}{S}\sum_s A^{(s)}C^{(s)}\right)\left(\frac{1}{S}\sum_s A^{(s)}\right)\left(\frac{1}{S}\sum_s B^{(s)}\right),\\
D_4 &:= \left(\frac{1}{S}\sum_s A^{(s)}B^{(s)}\right)\left(\frac{1}{S}\sum_s A^{(s)}C^{(s)}\right).
\end{aligned}
$$

We compute the expectation of each D_j.

D_1. By expanding D_1, we know that $\mathbb{E}D_1 = \frac{1}{S^3}\sum_{i,j,k}\mathbb{E}[A^{(k)}B^{(k)}A^{(i)}C^{(j)}]$. The value of $\mathbb{E}[A^{(k)}B^{(k)}A^{(i)}C^{(j)}]$ depends on the triplet (i,j,k) in the following way:

$$
\mathbb{E}[A^{(k)}B^{(k)}A^{(i)}C^{(j)}] = \begin{cases} 0 & \text{if } i=k, j\neq k\\ \mathbb{E}[A^2BC] & \text{if } i=k, j=k\\ 0 & \text{if } i\neq k, j=k\\ \mathbb{E}[AB]\mathbb{E}[AC] & \text{if } i\neq k, j\neq k, i=j\\ 0 & \text{if } i\neq k, j\neq k, i\neq j\end{cases}
$$

We have used independence of $(A^{(s)}, B^{(s)}, C^{(s)})_{s=1}^S$ to factorize the expectation $\mathbb{E}[A^{(k)}B^{(k)}A^{(i)}C^{(j)}]$. For certain triplets, the factorization reveals that the expectation is zero. By accounting for all triplets, the expectation of D_1 is

$$
\frac{1}{S^3}\left[S\mathbb{E}[A^2BC] + S(S-1)\mathbb{E}[AB]\mathbb{E}[AC]\right].
$$

D_2. By expanding D_2, we know that $\mathbb{E}D_2 = \frac{1}{S^4}\sum_{i,j,p,q}\mathbb{E}[A^{(i)}A^{(i)}B^{(p)}C^{(q)}]$. We can do a similar case-by-case analysis of how $\mathbb{E}[A^{(i)}A^{(i)}B^{(p)}C^{(q)}]$ depend on the quartet (i,j,p,q). In the end, the expectation of D_2 is

$$
\frac{1}{S^3}\left[\mathbb{E}[A^2BC] + (S-1)\mathbb{E}[A^2]\mathbb{E}[BC] + 2(S-1)\mathbb{E}[AB]\mathbb{E}[AC]\right].
$$

D_3. By symmetry betwene D_1 and D_3, the expectation of D_3 is also

$$
\frac{1}{S^3}\left[S\mathbb{E}[A^2BC] + S(S-1)\mathbb{E}[AB]\mathbb{E}[AC]\right].
$$

D_4. By expanding D_4, we know that $\mathbb{E}D_4 = \frac{1}{S^2}\sum_{i,j}\mathbb{E}[A^{(i)}B^{(i)}A^{(j)}C^{(j)}]$. The case-by-case analysis of $\mathbb{E}[A^{(i)}B^{(i)}A^{(j)}C^{(j)}]$ for each (i,j) is simple, and is omitted. The expectation of D_4 is

$$
\frac{1}{S}\mathbb{E}[A^2BC] + \frac{S-1}{S}\mathbb{E}[AB]\mathbb{E}[AC].
$$

Simple algebra reveals that $\sum_{i=1}^4\mathbb{E}[D_i] - \frac{S-1}{S}\mathbb{E}[AB]\frac{S-1}{S}\mathbb{E}[AC]$ is equal to Equation (6).

Proof D.7 (Proof of Lemma C.3)**.** *In this proof, we will only consider expectations under the full-data posterior. Hence, to alleviate notation, we shall write $\mathbb{E}$ instead of $\mathbb{E}_{\mathbf{1}_N}$: similarly, covariance and variance evaluations are understood to be at $w = \mathbf{1}_N$.*

Applying Lemma D.1, the covariance of $\hat{\psi}_n$ and $\hat{\psi}_n$ i.e. the variance of $\hat{\psi}_n$ is equal to

$$\frac{(S-1)^2}{S^3}\mathbb{E}\{(g(\beta)-\mathbb{E}[g(\beta)])^2(L(d^{(n)}\mid\beta)-\mathbb{E}[L(d^{(n)}\mid\beta)])^2\}$$
$$+\frac{S-1}{S^3}Var(L(d^{(n)}\mid\beta))Var(g(\beta))-\frac{(S-1)(S-2)}{S^3}Cov(g(\beta),L(d^{(n)}\mid\beta))^2.$$

Define the constant C to be the maximum over n of

$$Cov(g(\beta),L(d^{(n)}\mid\beta))^2+Var(g(\beta))Var(L(d^{(n)}\mid\beta))$$
$$+\mathbb{E}\{(g(\beta)-\mathbb{E}[g(\beta)])^2(L(d^{(n)}\mid\beta)-\mathbb{E}[L(d^{(n)}\mid\beta)])^2\}.$$

Simple algebra shows that $Var(\hat{\psi}_n)\leq\frac{C}{S}$.

Proof D.8 (Proof of Theorem C.1)**.** *Similar to the proof of Lemma C.3, expectations (and variances and covariances) are understood to be taken under the full-data posterior.*

Since $\hat{\psi}_n$ is the biased sample variance, we know that

$$\mathbb{E}\hat{\psi}_n=\frac{S-1}{S}\psi_n.$$

The bias of $\hat{\psi}_n$ goes to zero at rate $1/S$. Because of Lemma C.3, the variance also goes to zero at rate $1/S$. Then, the application of Chebyshev's inquality shows that $\hat{\psi}_n\xrightarrow{p}\psi_n$. Since N is a constant, the pointwise convergence $|\hat{\psi}_n-\psi_n|\xrightarrow{p}0$ implies the uniform convergence $\max_{n=1}^N|\hat{\psi}_n-\psi_n|\xrightarrow{p}0$.

We now prove that $|\widehat{\Delta}-\Delta(\alpha)|\xrightarrow{p}0$. We first recall some notation. The ranks $r_1,r_2,\ldots,r_N$ sort the influences $\psi_{r_1}\leq\psi_{r_2}\leq\ldots\leq\psi_{r_N}$, and $\Delta(\alpha)=-\sum_{m=1}^{\lfloor N\alpha\rfloor}\psi_{r_m}\mathbb{I}\{\psi_{r_m}<0\}$. Similarly, $v_1,v_2,\ldots,v_N$ sort the estimates $\hat{\psi}_{v_1}\leq\hat{\psi}_{v_2}\leq\ldots\leq\hat{\psi}_{v_N}$, and $\widehat{\Delta}=-\sum_{m=1}^{\lfloor N\alpha\rfloor}\hat{\psi}_{v_m}\mathbb{I}\{\hat{\psi}_{v_m}<0\}$. It suffices to prove the convergence when $\lfloor N\alpha\rfloor\geq1$: in the case $\lfloor N\alpha\rfloor=0$, both $\widehat{\Delta}$ and $\Delta(\alpha)$ are equal to zero, hence the distance between them is identically zero. Denote the T unique values among ψ_n by $u_1<u_2<\ldots<u_T$. If $T=1$ i.e. there is only one value, let $\omega:=1$. Otherwise, let ω be the smallest gap between subsequent values: $\omega:=\min_t(u_{t+1}-u_t)$.

Suppose that $\max_{n=1}^N|\hat{\psi}_n-\psi_n|\leq\omega/3$: let A be the indicator for this event. For any n, each $\hat{\psi}_n$ is in the interval $[\psi_n-\omega/3,\psi_n+\omega/3]$. In the case $T=1$, clearly all k such that $\hat{\psi}_k$ is in $[\psi_n-\omega/3,\psi_n+\omega/]$ satisfy $\psi_k=\psi_n$. In the case $T>1$, since unique values of ψ_n are at least ω apart, all k such that $\hat{\psi}_k$ is in $[\psi_n-\omega/3,\psi_n+\omega/]$ satisfy $\psi_k=\psi_n$. This means that the ranks $v_1,v_2,\ldots,v_N$, which sort the influence estimates, also sort the true influences in ascending order: $\psi_{v_1}\leq\psi_{v_2}\leq\ldots\leq\psi_{v_N}$. Since the ranks $r_1,r_2,\ldots,r_N$ also sort the true influences, it must be true that $\psi_{v_m}=\psi_{r_m}$ for all m. Therefore, we can write

$$|\widehat{\Delta}-\Delta(\alpha)|=\left|\sum_{m=1}^{\lfloor N\alpha\rfloor}\left(\psi_{v_m}\mathbb{I}\{\psi_{v_m}<0\}-\hat{\psi}_{v_m}\mathbb{I}\{\hat{\psi}_{v_m}<0\}\right)\right|$$
$$\leq\sum_{m=1}^{\lfloor N\alpha\rfloor}\left|\psi_{v_m}\mathbb{I}\{\psi_{v_m}<0\}-\hat{\psi}_{v_m}\mathbb{I}\{\hat{\psi}_{v_m}<0\}\right|.$$

We control the absolute values $\left|\psi_{v_m}\mathbb{I}\{\psi_{v_m}<0\}-\hat{\psi}_{v_m}\mathbb{I}\{\hat{\psi}_{v_m}<0\}\right|$. For any index n, by triangle inequality, $\left|\psi_n\mathbb{I}\{\psi_n<0\}-\hat{\psi}_n\mathbb{I}\{\hat{\psi}_n<0\}\right|$ is at most

$$\mathbb{I}\{\hat{\psi}_n<0\}|\psi_n-\hat{\psi}_n|+|\psi_n||\mathbb{I}\{\hat{\psi}_n<0\}-\mathbb{I}\{\psi_n<0\}|.$$

The first term is at most $|\psi_n-\hat{\psi}_n|$. The second term is at most $\mathbb{I}\{|\psi_n-\hat{\psi}_n|\geq|\psi_n|,\psi_n\neq0\}$. We next prove a bound on $\left|\psi_n\mathbb{I}\{\psi_n<0\}-\hat{\psi}_n\mathbb{I}\{\hat{\psi}_n<0\}\right|$ that holds across n. Our analysis proceeds differently based on whether the set $\{n:\psi_n\neq0\}$ is empty or not.

- *$\{n : \psi_n \neq 0\}$ is empty. This means $\psi_n = 0$ for all n. Hence, $\mathbb{I}\{|\psi_n - \hat{\psi}_n| \geq |\psi_n|, \psi_n \neq 0\}$ is identically zero.*
- *$\{n : \psi_n \neq 0\}$ is not empty. We then know that $\min_n |\psi_n| > 0$. Hence, $\mathbb{I}\{|\psi_n - \hat{\psi}_n| \geq |\psi_n|, \psi_n \neq 0\}$ is upper bounded by $\mathbb{I}\{|\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}$. Since $|\psi_n - \hat{\psi}_n| \leq \max_n |\psi_n - \hat{\psi}_n|$, this last indicator is at most $\mathbb{I}\{\max_n |\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}$.*

To summarize, we have proven the following upper bounds on $|\widehat{\Delta} - \Delta(\alpha)|$. When $\{n : \psi_n \neq 0\}$ is empty, on A, $|\widehat{\Delta} - \Delta(\alpha)|$ is upper bounded by

$$\lfloor N\alpha \rfloor \max_{n=1} |\psi_n - \hat{\psi}_n| \tag{7}$$

When $\{n : \psi_n \neq 0\}$ is not empty, on A, $|\widehat{\Delta} - \Delta(\alpha)|$ is upper bounded by

$$\lfloor N\alpha \rfloor \max_{n=1} |\psi_n - \hat{\psi}_n| + \lfloor N\alpha \rfloor \mathbb{I}\{\max_n |\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}. \tag{8}$$

We are ready to show that $\Pr(|\widehat{\Delta} - \Delta(\alpha)| > \epsilon)$ converges to zero. For any positive ϵ, we know that

$$\Pr(|\widehat{\Delta} - \Delta(\alpha)| > \epsilon) \leq \Pr(|\widehat{\Delta} - \Delta(\alpha)| > \epsilon, A) + \Pr(A^c).$$

The later probability goes to zero because $\max_{n=1}^N |\hat{\psi}_n - \psi_n| \xrightarrow{p} 0$.

Suppose that $\{n : \psi_n \neq 0\}$ is empty. Using the upper bound Equation (7), we know that event in the former probability implies that $\max_{n=1}^N |\hat{\psi}_n - \psi_n| \geq \epsilon / \lfloor N\alpha \rfloor$: The probability of this event also goes to zero because $\max_{n=1}^N |\hat{\psi}_n - \psi_n| \xrightarrow{p} 0$.

Suppose that $\{n : \psi_n \neq 0\}$ is not empty. Using the upper bound Equation (8), we know that event in the former probability implies that $(\max_{n=1}^N |\hat{\psi}_n - \psi_n| + \mathbb{I}\{\max_n |\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}) \geq \epsilon / \lfloor N\alpha \rfloor$. Since $\max_{n=1}^N |\hat{\psi}_n - \psi_n|$ converges to zero in probability, $\mathbb{I}\{\max_n |\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}$ also converges to zero in probability. Hence, the probability that $(\max_{n=1}^N |\hat{\psi}_n - \psi_n| + \mathbb{I}\{\max_n |\psi_n - \hat{\psi}_n| \geq \min_n |\psi_n|\}) \geq \epsilon / \lfloor N\alpha \rfloor$ converges to zero.

In all, $\Pr(|\widehat{\Delta} - \Delta(\alpha)| > \epsilon)$ goes to zero in both the case where $\{n : \psi_n \neq 0\}$ is empty and the complement case. As the choice of ϵ was arbitrary, we have shown $\widehat{\Delta} \xrightarrow{p} \Delta(\alpha)$.

Proof D.9 (Proof of Theorem C.2)**.** *Similar to the proof of Lemma D.1, we only consider expectations under the full-data posterior. Hence, we will write $\mathbb{E}$ instead of $\mathbb{E}_{\mathbf{1}_N}$ to simplify notation. Variance and covariance operations are also understood to be taken un der the full-data posteiror. To lighten the dependence of the notation on the parameter β, we will write $g(\beta)$ as g and $L(d^{(n)} \mid \beta)$ as L_n when talking about the expectation of $g(\beta)$ and $L(d^{(n)} \mid \beta)$.*

Define the the following multivariate function

$$f(\beta) := \left[g(\beta), L(d^{(1)} \mid \beta), g(\beta)L(d^{(1)} \mid \beta), \ldots, L(d^{(N)} \mid \beta), g(\beta)L(d^{(N)} \mid \beta)\right]^T.$$

As defined, $f(\cdot)$ is a mapping from V-dimensional space to $2N+1$-dimensional space. Since $(\beta^{(1)}, \ldots, \beta^{(S)})$ is an i.i.d. sample, $\left(f(\beta^{(1)}), f(\beta^{(2)}), \ldots, f(\beta^{(S)})\right)$ is also an i.i.d. sample. Because of the moment conditions we have assumed, each $f(\beta)$ has finite variance. We apply the Lindeberg-Feller multivariate central limit theorem (van der Vaart, 1998, Proposition 2.27), and conclude that

$$\sqrt{S}\left(\frac{1}{S}\sum_s f(\beta^{(s)}) - \mathbb{E}f(\beta)\right) \xrightarrow{D} N(\mathbf{0}, \Xi)$$

where the limit is $S \to \infty$, and Ξ is a symmetric $(2N+1) \times (2N+1)$ dimensional matrix, which we specify next. It suffices to write down the formula for (i, j) entry of Ξ where $i \leq j$:

$$\Xi_{i,j} = \begin{cases} Var(g) & \text{if } i = j = 1 \\ Cov(g, L_n) & \text{if } i = 1, j > 1 \\ Cov(L_n, L_m) & \text{if } i = 2n, j = 2m \\ Cov(L_n, gL_m) & \text{if } i = 2n, j = 2m+1 \\ Cov(gL_n, L_m) & \text{if } i = 2n+1, j = 2m \\ Cov(gL_n, gL_m) & \text{if } i = 2n+1, j = 2m+1 \end{cases}.$$

To relate the asymptotic distribution of $f(\beta)$ to that of the vector $\hat{\psi}$, we now use the delta method. Define the following function which acts on $2N+1$ dimensional vectors and returns N dimensional vectors:

$$h([x_1, x_2, \ldots, x_{2N+1}]^T) := \left[x_3 - x_1x_2, x_5 - x_1x_4, x_7 - x_1x_6, \ldots, x_{2N+1} - x_1x_{2N}\right]^T.$$

Written this way, clearly $h(\cdot)$ transform the sample mean $\frac{1}{S}\sum_s f(\beta^{(s)})$ into the estimated influences: $\hat{\psi} = h\left(\frac{1}{S}\sum_s f(\beta^{(s)})\right)$. Furthermore, $h(\cdot)$ applied to $\mathbb{E}f(\beta)$ yields the vector of true influences: $\psi = h\left(\mathbb{E}f(\beta)\right)$. $h(\cdot)$ is continuously differentiable everywhere: its Jacobian is the $N \times (2N+1)$ matrix

$$\mathbf{J}_h = \begin{bmatrix} -x_2 & -x_1 & 1 & 0 & 0 & \ldots & 0 \\ -x_4 & 0 & 0 & -x_1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & \ldots & 0 \\ -x_{2N} & 0 & 0 & \ldots & 0 & \ldots & 1 \end{bmatrix},$$

which is non-zero. Therefore, we apply the delta method (van der Vaart, 1998, Theorem 3.1) and conclude that

$$\sqrt{S}\left(\hat{\psi} - \psi\right) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{J}_h\big|_{x=\mathbb{E}f(\beta)} \Xi (\mathbf{J}_h\big|_{x=\mathbb{E}f(\beta)})^T\right).$$

The (i,j) entry of the asymptotic covariance matrix is the dot product between the i-th row of $\mathbf{J}_h\big|_{x=\mathbb{E}f(\beta)}$ and the j-th column of $\Xi(\mathbf{J}_h\big|_{x=\mathbb{E}f(\beta)})^T$. The former is

$$[-\mathbb{E}L_i, 0, 0, \ldots, \underbrace{-\mathbb{E}g}_{2i \text{ entry}}, \underbrace{1}_{(2i+1) \text{ entry}}, \ldots, 0].$$

The later is

$$\begin{bmatrix} (-\mathbb{E}L_j)Cov(g,g) - (\mathbb{E}g)Cov(g,L_j) + Cov(g,gL_j) \\ \vdots \\ (-\mathbb{E}L_j)Cov(gL_N,g) - (\mathbb{E}g)Cov(gL_N,L_j) + Cov(gL_N,gL_j) \end{bmatrix}.$$

Taking the dot product, we have that the (i,j) entry of the asymptotic covariance matrix is equal to

$$\begin{aligned} &Cov(gL_i, gL_j) - (\mathbb{E}g)\left[Cov(gL_i, L_j) + Cov(gL_j, L_i)\right] \\ &- \left[(\mathbb{E}L_j)Cov(g, gL_i) + (\mathbb{E}L_i)Cov(g, gL_j)\right] \\ &+ (\mathbb{E}L_j)(\mathbb{E}L_i)Var(g) \\ &+ (\mathbb{E}g)^2 Cov(L_i, L_j) \\ &+ (\mathbb{E}g)\left[(\mathbb{E}L_j)Cov(g, L_i) + (\mathbb{E}L_i)Cov(g, L_j)\right] \end{aligned}$$

It is simple to check that the last display is equal to the covariance between $(g - \mathbb{E}[g])(L_j - \mathbb{E}[L_j])$ and $(g - \mathbb{E}[g])(L_i - \mathbb{E}[L_i])$.

Proof D.10 (Proof of Lemma C.4)**.** *We use the (shape, rate) parametrization of the gamma distribution. Let the prior over τ be Gamma(α, β) where $\alpha, \beta > 0$. Conditioned on observations, the posterior distribution of (μ, τ) is normal-gamma:*

$$\begin{aligned} \tau &\sim Gamma\left(\alpha + \frac{N}{2}, \beta + \frac{N}{2}\left[\frac{1}{N}\sum_{n=1}^{N}(x^{(n)})^2 - \bar{x}^2\right]\right), \\ \epsilon &\sim N(0,1), \\ \mu \mid \tau, \epsilon &= \bar{x} + \frac{\epsilon}{\sqrt{N\tau}}. \end{aligned}$$

In this section, since we only take expectations under the original full-data posterior, we will lighten the notation's dependence on w, and write $\mathbb{E}$ instead of $\mathbb{E}_{\mathbf{1}_N}$. Similarly, covariance and variance operators are understood to be under the full-data posterior.

For completeness, we compute $Cov(\mu, L(d^{(n)} \mid \mu, \tau))$. We know that $\mu - \mathbb{E}\mu = \epsilon/\sqrt{N\tau}$. The log likelihood, as a function of τ and ϵ, is

$$\frac{1}{2}\log\left(\frac{\tau}{2\pi}\right) - \frac{1}{2}\tau(x^{(n)} - \bar{x})^2 - \frac{1}{2N}\epsilon^2 + \frac{x^{(n)} - \bar{x}}{\sqrt{N}}\epsilon\sqrt{\tau}.$$

The covariance of μ and $L(d^{(n)} \mid \mu, \tau)$ is equal to the covariance between $\epsilon/\sqrt{N\tau}$ and $L(d^{(n)} \mid \mu, \tau)$. Since $\epsilon/\sqrt{N\tau}$ is zero mean, the covariance is equal to the expectation of the product. Since ϵ is indedependent of τ, many of the terms that form the expectatin of the product is zero. After some algebra, the only term that remains is

$$\mathbb{E}\left[\frac{x^{(n)} - \bar{x}}{N}\epsilon^2\right] = \frac{x^{(n)} - \bar{x}}{N}.$$

To compute the asymptotic variance of $\hat{\psi}_n$, it suffices to compute the expectation of $\frac{\epsilon^2}{N\tau}\left(L(d^{(n)} \mid \mu, \tau) - \mathbb{E}L(d^{(n)} \mid \mu, \tau)\right)^2$. The calculations are simple, but tedious, and we omit them. We will only state the result. The expectation of $\frac{\epsilon^2}{N\tau}\left(L(d^{(n)} \mid \mu, \tau) - \mathbb{E}L(d^{(n)} \mid \mu, \tau)\right)^2$ is

$$\begin{aligned}
&\left[\frac{1}{4N}\mathbb{E}[\tau^{-1}(\tau - \mathbb{E}\tau)^2]\right](x^{(n)} - \bar{x})^4 \\
&+ \left[\frac{3 + \mathbb{E}[\tau^{-1}(\tau - \mathbb{E}\tau)]}{N^2} - \frac{\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)]}{2N}\right](x^{(n)} - \bar{x})^2 \\
&+ \frac{1}{2N^3}\mathbb{E}[\tau^{-1}] + \frac{1}{2N}\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)^2] - \frac{1}{N^2}\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)^2].
\end{aligned}$$

Since the asymptotic variance is equal to this expectation minus the square of the covariance between $L(d^{(n)} \mid \mu, \tau)$ and μ, our final expression for the asymptotic variance $\Sigma_{n,n}$ is

$$\begin{aligned}
&\left[\frac{1}{4N}\mathbb{E}[\tau^{-1}(\tau - \mathbb{E}\tau)^2]\right](x^{(n)} - \bar{x})^4 \\
&+ \left[\frac{2 + \mathbb{E}[\tau^{-1}(\tau - \mathbb{E}\tau)]}{N^2} - \frac{\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)]}{2N}\right](x^{(n)} - \bar{x})^2 \\
&+ \frac{1}{2N^3}\mathbb{E}[\tau^{-1}] + \frac{1}{2N}\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)^2] - \frac{1}{N^2}\mathbb{E}[\tau^{-1}(\log\tau - \mathbb{E}\log\tau)^2].
\end{aligned}$$

The constants D_1, D_2, and D_3 mentioned in the lemma statement can be read off this last display. It is possible to replace the posterior functionals of τ with quantities that only depends on the prior (α, β) and the observed data. Such formulas might be helpful in studying the behavior of $\Sigma_{n,n}$ in the limit where some $x^{(n)}$ becomes very large.

E. Experimental Setup

In this section, we only describe the checks: for the actual results, see Section 4 and Appendices F.1 and F.2.

A practitioner with a particular definition of "small data" can set α to reflect their concern. We consider a number of α values. We set the maximum value of α to be 0.01. This choice is motivated by Broderick et al. (2023). Many analyses are non-robust to removing 1% of the data, and we a priori think that $\alpha > 1\%$ is a large amount of data to remove. We vary $\log_{10}(\alpha)$ in an equidistant grid of length 10 from -3 to -2. The ten values are 0.10%, 0.13%, 0.17%, 0.22%, 0.28%, 0.36%, 0.46%, 0.60%, 0.77% and 1.00%.

For the range of dropout fraction specified above and across three common quantities of interest corresponding to sign, significance, and significant result of opposite sign changes, we walk through what a practitioner would do in practice (although they would choose only one α and one decision). Our method proposes an influential data subset and a change in the quantity of interest, represented by a confidence interval.

Ideally, we want to check if our interval includes the result of the worst-case data to leave out. We are unable to do so, since we do not know how to compute the worst-case result in a reasonable amount of time. We settle for the following checks.

In the first check, for a particular MCMC run, we plot how the change from re-running minus the proposed data compares to the confidence interval. We recommend the user run this check if re-running MCMC a second time is not too computationally expensive.

Unfortunately, such refitting does not paint a complete picture of approximation quality. For instance, the MCMC run might be unlucky since MCMC is random. To be more comprehensive, we run additional checks. We do not expect users to run

these tests, as their computational costs are high. The central question is how frequently (under MCMC randomness) the confidence interval includes the result after removing the worst-case data. Since we estimate the worst-case change with a linear approximation, a natural way to answer this question is with two separate checks: while Appendix E.1 checks how frequently the confidence interval includes the result of the linear approximation i.e. the AMIP, Appendix E.3, checks whether the linear approximation is good. To understand *why* we observe the coverage in Appendix E.1, in Appendix E.2 we isolate the impact of the sorting step in the construction of our confidence interval.

E.1. Estimate coverage of confidence interval for AMIP

We estimate how frequently $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$ covers the AMIP by using another level of Monte Carlo. Recall that $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$ is intended to be a confidence interval covering $\Delta(\alpha)$ a fraction η of the time. If the estimated coverage is far from η, we have evidence that $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$ does not achieve the desired nominal coverage.

We draw J Markov chains: we set $J = 960$. On each chain, we estimate the influences and construct the confidence interval $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$. From all the chains, for each n, we have J estimates of ψ_n. We take the sample mean across chains, and denote this by ψ_n^*: because of variance reduction through averaging, ψ_n^* is a much better estimate of ψ_n than individual $\hat{\psi}_n$. We denote the indices of the $\lfloor N\alpha \rfloor$ most negative ψ_n^* by $U^*(\alpha)$. We sort ψ_n^* across n and sum the $\lfloor N\alpha \rfloor$ most negative ψ_n^*. This sum is denoted by $\Delta^*(\alpha)$: we use it in place of the ground truth $\Delta(\alpha)$. We use the sample mean of the indicators $\mathbb{I}\{\Delta^*(\alpha) \in [\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]\}$ as the point estimate of the coverage. We also report a 95% confidence interval for the coverage. This interval is computed using binomial tests designed in Clopper and Pearson (1934) and implemented as R's binom.test() function.

E.2. Estimate coverage of confidence intervals for sum-of-influence

It is possible that the estimated coverage of $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$ is far from the nominal η. We suspect that such a discrepancy comes from the sorting of $\hat{\psi}_n$ to construct $\Delta(\alpha)$. To modularize out the sorting, we consider a target of inference that is simpler than $\Delta(\alpha)$. At a high level, we fix an index set I, and define the target to be the sum of influences in I: $\sum_{n\in I} \psi_n$. On each sample $(\beta^{(1)}, \ldots, \beta^{(S)})$, our point estimate is $\sum_{n\in I} \hat{\psi}_n$: this estimate does not involve any sorting, while $\widehat{\Delta}$ does. We construct the confidence interval, $[V^{lb}, V^{ub}]$, from the block bootstrap distribution of $\sum_{n\in I} \hat{\psi}_n$. The difference between $[V^{lb}, V^{ub}]$ and $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$, which is constructed from the block bootstrap distribution of $\widehat{\Delta}$, is that the former is not based on sorting the influence estimates. If the actual coverage of $[V^{lb}, V^{ub}]$ is close to the nominal value, we have evidence that the miscoverage of $[\Delta^{lb}(\alpha), \Delta^{ub}(\alpha)]$ is due to this sorting.

From Appendix E.1 we use ψ_n^* and the associated $\Delta^*(\alpha)$ and $U^*(\alpha)$ as replacement for ground truths. We set I to be $U^*(\alpha)$. We run another set of J Markov chains: for each chain, we construct the confidence interval $[V^{lb}, V^{ub}]$ by sampling from the block bootstrap distribution of the estimator $\sum_{n\in I} \hat{\psi}_n$. We report the sample mean of the indicators $\mathbb{I}\{\sum_{n\in I} \psi_n^* \in [V^{lb}, V^{ub}]\}$ as our point estimate of the coverage. We also report a 95% confidence interval for the coverage. This interval is computed using binomial tests designed in Clopper and Pearson (1934) and implemented as R's binom.test() function.

E.3. Re-running MCMC on interpolation path

Ideally, we want to know the difference between the Maximum Influence Perturbation and the AMIP. As we have established, we do not know how to compute the former efficiently. We settle for checking the linearity approximation made in Section 3.1 i.e. estimating $\phi(w) - \phi(\mathbf{1}_N)$ with $\sum_n (w_n - 1)\psi_n$. In particular, we expect the first-order Taylor series approximation to be arbitrarily good for w arbitrarily close to $\mathbf{1}_N$. By necessity, we are interested in some w^* that has a non-trivial distance from $\mathbf{1}_N$. Plotting the quantity of interest $\phi(w)$ on an interpolation path between $\mathbf{1}_N$ and w^*, we get a sense of how much we have diverged from linearity by that point.

From Appendix E.1, we have ψ_n^* as our replacement for the ground truth ψ_n. We focus on $\alpha = 0.05$: 5% is a large amount of data to remove, and a priori we expect the linear approximation to be poor. Recall that $U^*(0.05)$ is the set of $\lfloor 0.05N \rfloor$ observations that are most influential according to sorted ψ_n^*. Let w^* be the N-dimensional weight vector that is 1 for observations in $U^*(0.05)$ and 0 otherwise. For $\zeta \in [0, 1]$, the linear approximation of $\phi(\zeta w^* + (1 - \zeta)\mathbf{1}_N)$ is $\phi(\mathbf{1}_N) + \zeta\Delta^*(0.05)$. In the extreme $\zeta = 0$, we do not leave out any data. In the extreme $\zeta = 1$, we leave out the entirety of

$U^*(0.05)$ i.e. 5% of the data. An intermediate value ζ roughly corresponds[5] to removing $(\zeta 5)\%$ of the data. We discretize $[0, 1]$ with 15 values: 0, 0.0010, 0.0016, 0.0027, 0.0044, 0.0072, 0.0118, 0.0193, 0.0316, 0.0518, 0.0848, 0.1389, 0.2276, 0.3728, 0.6105, 1. For each value on this grid, we run MCMC to estimate $\phi(\zeta w^* + (1-\zeta)\mathbf{1}_N)$, and compare it to the linear approximation.

F. Other Quality Checks

In our experiments, we find that our approximation works well for a simple linear model. But we find that it can struggle in hierarchical models with more complex structure. While Section 4 has plotted our confidence interval and the result after removing proposed data, we have not yet discussed the checks in Appendices E.1 to E.3. Furthermore, Section 4 has not described the context to each analysis. This section a) fills in the missing context and b) shows the outcome of the additional quality checks.

F.1. Linear model

We consider a slight variation of a microcredit analysis from Meager (2019). In Meager (2019), conclusions regarding microcredit efficacy were based on ordinary least squares (OLS). We refer the reader to Broderick et al. (2023, Section 4.3.2) for investigations of such conclusions' non-robustness. Here, we instead consider an analogous Bayesian analysis using MCMC, and we examine the robustness of conclusions from this analysis. Even for this very simple Bayesian analysis, it is possible to change substantive conclusions by removing a small fraction of the data.

Our quality checks suggest that our approximation is accurate. Our confidence interval contains the refit after removing the proposed data. The actual coverage of the confidence interval for AMIP is close to the nominal coverage. The actual coverage of the confidence interval for sum-of-influence is also close to the nominal coverage. Even for dropping 5% of the data, the linear approximation is still adequate.

F.1.1. Background.

Meager (2019) studies the microcredit data from Angelucci et al. (2015), which was an RCT conducted in Mexico. There are $N = 16{,}560$ households in the RCT. Each observation is $d^{(n)} = (x^{(n)}, y^{(n)})$, where $x^{(n)}$ is the treatment status and $y^{(n)}$ is the profit measured. The log-likelihood for the n-th observation is $L(d^{(n)} \mid \mu, \theta, \sigma) = -\frac{1}{2\sigma^2}(y^{(n)} - \theta x^{(n)} - \mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$. Here, the model parameters are baseline profit μ, treatment effect θ, and noise scale σ. The most interesting parameter is θ: as $x^{(n)}$ is binary, θ compares the means in the treatment and control groups. Meager (2019) estimates the model parameters with OLS.

Our variation of the above analysis is as follows. We put *t* location-scale distribution priors on the model parameters, with the additional constraint that the noise scale σ is positive. Recall that the *t* location-scale distribution has three hyperparameters: ν, μ, σ. ν is the degrees of freedom, μ is the location, and σ is the scale. The density at y of this distribution is

$$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

We set the the prior over μ to be *t* location-scale with degrees of freedom 3, location 0, and scale 1000. We set the the prior over θ to be *t* location-scale with degrees of freedom 3, location 0, and scale 1000. We set the the prior over σ to be *t* location-scale with degrees of freedom 3, location 0, and scale 1000.

We use Hamiltonian Monte Carlo (HMC) as implemented in Stan (Carpenter et al., 2017) to approximate the full-data posterior. We draw $S = 4000$ samples.

F.1.2. Runtime.

The running of our approximation takes very little time compared to the running of the original analysis. Generating the draws in Figure 1 took 3 minutes on MIT Supercloud (Reuther et al., 2018). For one α and one quantity of interest, it took less than 5 seconds to make a confidence interval for what happens if we remove the most extreme data subset. A user might check approximation quality by dropping a proposed subset and re-running MCMC: each such check took us around 3

[5]This correspondence is not exact, since for $\zeta < 1$, all observations in $U^*(0.05)$ are included in the analysis, only with downplayed contributions.

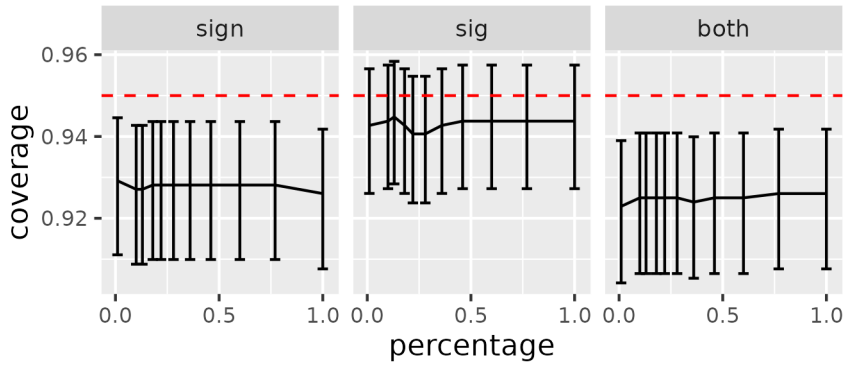

Figure 5. (Linear model) Monte Carlo estimate of AMIP confidence interval's coverage. Each panel corresponds to a target conclusion change. The dashed line is the nominal level $\eta = 0.95$. The solid line is the sample mean of the indicator variable for the event that ground truth is contained in the confidence interval. The error bars are confidence intervals for the population mean of these indicators.

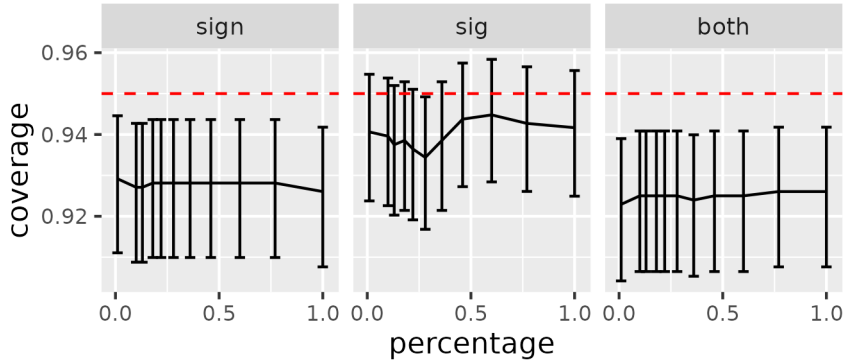

Figure 6. (Linear model) Monte Carlo estimate of sum-of-influence confidence interval's coverage. Each panel corresponds to a target conclusion change. The dashed line is the nominal level $\eta = 0.95$. The solid line is the sample mean of the indicator variable for the event that ground truth is contained in the confidence interval, and error bars are confidence intervals for the population mean of these indicators.

minutes, the runtime of the original analysis.

F.1.3. Additional quality checks.

Figure 5 shows that the actual coverage of the confidence interval for the AMIP is close to the nominal one, across α. As the half-width of each error bar is small (only 0.02), we believe that the difference between the true coverage and our point estimate of it is small. For either 'sign' or 'both' QoI, the error bars do not contain the nominal η. However, the difference between the point estimate and the nominal η is only 0.03 at worst, which is small. For the 'sig' QoI, the point estimate is within 0.005 of the nominal value, and the error bars contain the nominal η.

Figure 6 shows that the actual coverage of the confidence interval for the sum-of-influence is close to the nominal one across α. The absolute errors between our estimate of coverage and the nominal η are similar to those seen in Figure 5. This success suggests that the default block length, $L = 10$, is appropriate for this problem.

Figure 7 shows that the linear approximation works very well. It is somewhat remarkable that the linear approximation is this good even after dropping 5%, which we consider to be a large fraction of data. The horizontal axis ('scale') is the same as ζ in Appendix E.3. For all quantities of interest, the linear approximation and the refit lie mostly on top of each other: towards the right end of each panel, the approximation slightly underestimates the refit.

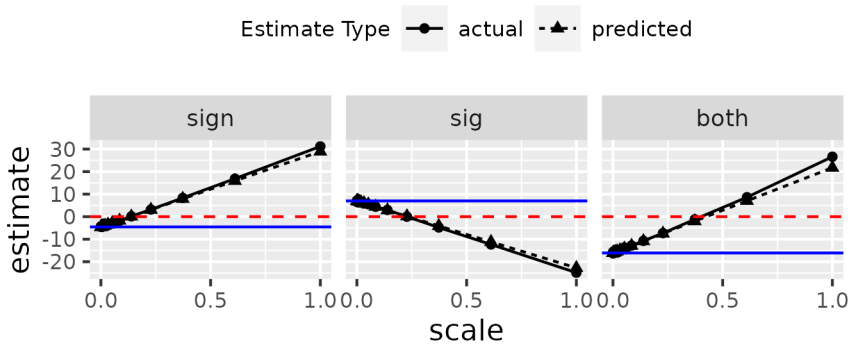

Figure 7. (Linear model) Quality of the linear approximation. Each panel corresponds to a target conclusion change. The solid blue line is the full-data fit. The horizontal axis is the distance from the weight that represents the full data. We plot both the refit from rerunning MCMC and the linear approximation of the refit.

F.2. Hierarchical model on tree mortality data

In the final experiment, we break from microcredit and look at ecological data. In particular, we consider a slight tweak of the analysis of European tree mortality from Senf et al. (2020). Senf et al. are acutely aware of generalization concerns. While previous work on tree death had been limited in both time and space, Senf et al. (2020) designs a large study that stretches across Europe and over 30 years, in hopes of making a broad-scale assessment. Our work shows that, even after an expansive study with generalization in mind, one might still worry about applying the findings at large, because of small-data sensitivity.

Our approximation struggles in this case. For the particular MCMC run used to estimate the full-data posterior, our confidence interval does not contain the refit after removing the proposed data. As each MCMC run is already highly time-consuming, we do not run quality checks on the whole dataset. We settle for running quality checks on a subsample of the data. On the subsampled data, the confidence interval for AMIP undercovers: the undercoverage is severe for one of the quantities of interest. However, the confidence interval for sum-of-influence is close to achieving the nominal coverage. For all three quantities of interest, the linear approximation is good up to removing roughly 1.1% of the data. For two of the three, it breaks down afterwards: for the remaining one, it continues to be good up to 3%, then falters.

As articulated in Appendix E, we think that dropping more than 1% of the data is already removing a large fraction. We are not worried about the Maximum Influence Perturbation for such α. So, that the linear approximation stops working after 1.1% is not a cause for concern.

F.2.1. Background.

Senf et al. (2020) studies the relationship between drought and tree death in Europe. To identify the association, they have compiled a dataset with $N = 87{,}390$ observations. Europe is divided into 2,913 regions, and the data spans 30 years: each observation is a set of measurements made in a particular region, which we denote as $l^{(n)}$, and at a particular year, which we denote as $t^{(n)}$. For our purposes, it suffices to know that the measurement of (the opposite of) drought is called climatic water balance, and we denote it as $x^{(n)}$: larger values of $x^{(n)}$ indicate that more water is available i.e. there is less drought. The response of interest, $y^{(n)}$, is excess death of tree canopy.

In our experiment, we mostly replicate (Senf et al., 2020)'s probabilistic model: we use the same likelihood, and make only an immaterial modification in the choice of priors. The likelihood for the n-th observation is exponentially modified Gaussian with standard deviation σ, scale λ and mean

$$\left(\mu^{(\text{time})}_{t^{(n)}} + \mu^{(\text{region})}_{l^{(n)}} + \mu\right) + \left(\theta^{(\text{time})}_{t^{(n)}} + \theta^{(\text{location})}_{l^{(n)}} + \theta\right) x^{(n)} + f(x^{(n)}),$$

with $f(x) := \sum_{i=1}^{10} B_i(x)\gamma_i$ where B_i's are fixed thin plate spline basis functions (Wood, 2003) and γ_i's are random: $\gamma_i \sim \text{Normal}(0, \sigma^2_{(\text{smooth})})$. In all, the parameters of interest are

- Fixed effects: μ and θ.

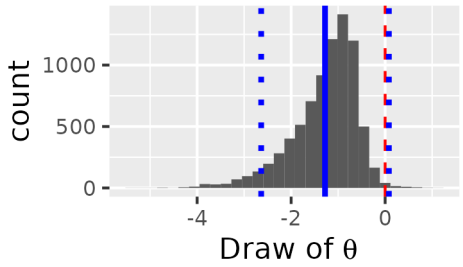

Figure 8. (Hierarchical model on subsampled tree mortality) Histogram of effect MCMC draws. See Figure 1 for the meaning of the distinguished lines.

- Random effects: time $(\mu^{(\text{time})}_{t^{(n)}}, \theta^{(\text{time})}_{t^{(n)}})$ and location $(\mu^{(\text{region})}_{l^{(n)}}, \theta^{(\text{location})}_{l^{(n)}})$.
- Degree of smoothing: $\sigma_{(\text{smooth})}$.

Since there are many regions (nearly 3,000) and periods of time (30), the number of random effects is large. Senf et al. (2020) uses brms()'s default priors for all parameters: in this default, the fixed effects are given improper uniform priors over the real line. To work with proper distributions, we set the priors for the random effects and degree of smoothing in the same way set by Senf et al. (2020). For fixed effects, we use t location-scale distributions with degrees of freedom 3, location 0, and scale 1000.

At a high level, both Senf et al. (2020)'s prior and our prior share strength across regions and times by modeling the random effects as coming from a some common *global* distributions. However, while Senf et al. (2020) uses an improper prior, we use a proper one. Numerically, there is no perceptible difference between the two. Theoretically, we prefer working with proper priors to avoid the integrability issue mentioned around Assumption 2.1.

Following Senf et al. (2020), we make conclusions based on posterior functionals of θ. Roughly speaking, θ is the average (across time and space) *association effect* that water balance has on excess tree death. We use $S = 8000$ HMC draws to approximate the posterior.

F.2.2. Runtime

The running of our approximation takes very little time compared to the running of the original analysis. Generating the draws in Figure 3 took 12 hours. For one α and one quantity of interest, it took less than 2 minutes to make a confidence interval for what happens if we remove the most extreme data subset. A user might check approximation quality by dropping a proposed subset and re-running MCMC: each such check took us around 12 hours, which is the runtime of the original analysis.

F.2.3. Results on subsampled data.

Running MCMC on the original dataset of size over 80,000 took 12 hours. In theory, we can spend time (on the order of thousands of hours) to run our quality checks, but we do not do so. Instead, we subsample 2,000 observations at random from the original dataset. Each MCMC on this subsample takes only 15 minutes, making it possible to run quality checks in a few hours instead of weeks. We hope that the subsampled data is representative enough of the original data that the quality checks on the subsampled data are indicative of the quality checks on the original data.

We use the same probabilistic model to analyze the subsampled data. Figure 8 plots the histogram of the association effect draws and sample summaries. Based on the draws, a forest ecologist might tentatively say that drought is positively associated with canopy mortality if they relied on the posterior mean, but refrain from conclusively deciding, since the uncertainty interval contains zero.

Figure 9 shows our confidence intervals and the actual refits. Similar to Figure 4, our confidence intervals predict a more extreme change than realized by the refit. The overestimation is most severe for 'both' QoI.

In Figure 10, the confidence interval for AMIP undercovers for all quantities of interest. The actual coverage decreases as

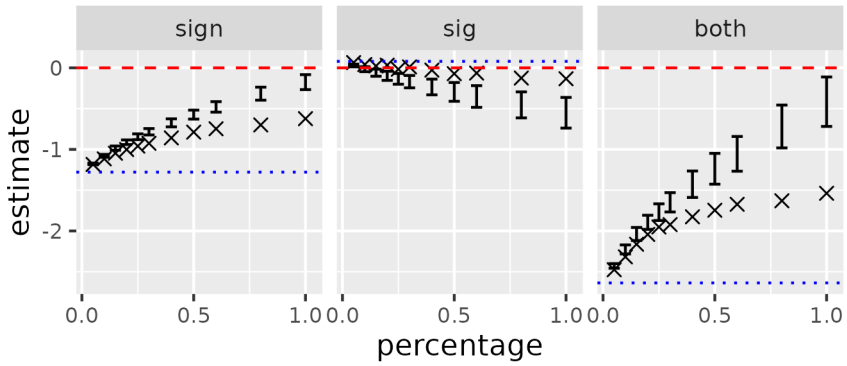

Figure 9. (Hierarchical model on subsampled tree mortality) Confidence interval and refit. See the caption of Figure 2 for the meaning of the panels and the distinguished lines.

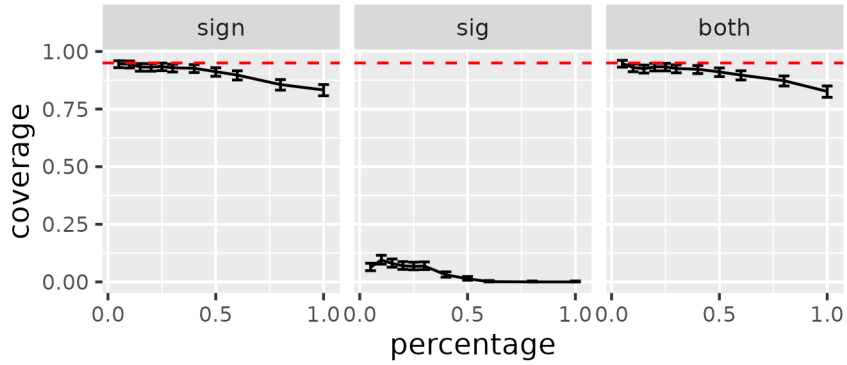

Figure 10. (Hierarchical model on subsampled tree mortality) Monte Carlo estimate of coverage of confidence interval for $\Delta(\alpha)$. See Figure 5 for the meaning of the panels and the distinguished lines.

α increases. The undercoverage is most severe for 'sig' QoI: while the nominal level is 0.95, the confidence interval for the true coverage only contains values less than 0.15. This translates to a relative error of over 84%. In other words, our confidence interval for significance change is too narrow, and rarely contains the AMIP. For 'both' QoI and 'sig' QoI, the worst-case relative error between the nominal and the estimated coverage, which occurs under the largest α, is 15.7%.

In Figure 11, the estimated coverage of the confidence interval for sum-of-influence is close to the nominal coverage. Note the stark contrast in the vertical scale of the 'sig' panel in Figure 10 with that in Figure 11. At worst, our point estimate of the true coverage is 0.04 less than the nominal level, which is only a 4.2% relative error. This success of the block bootstrap indicate that the undercoverage observed in Figure 10 can be attributed to the sorting step involved in the definition of $\hat{\Delta}$. We leave to future work to investigate why the interference cause by the sorting step is so much more severe for changing the significance than for changing sign or generating significant result of the opposite sign.

Figure 12 shows that the linear approximation is good for the posterior mean ('sign' QoI) and the left credible endpoint ('both' QoI) up to $\zeta = 0.2276$: in data percentages, this is roughly 1.1%. For larger ζ, the refit for 'both' QoI plateaus while the linear approximation continues to increase, and the linear approximation for posterior mean slightly underestimates it. For the left endpoint ('both' QoI), the linear approximation is close to the refit up to $\zeta = 0.6105$ (roughly 3% of data); afterwards, the left endpoint increases while the linear approximation continues to decrease.

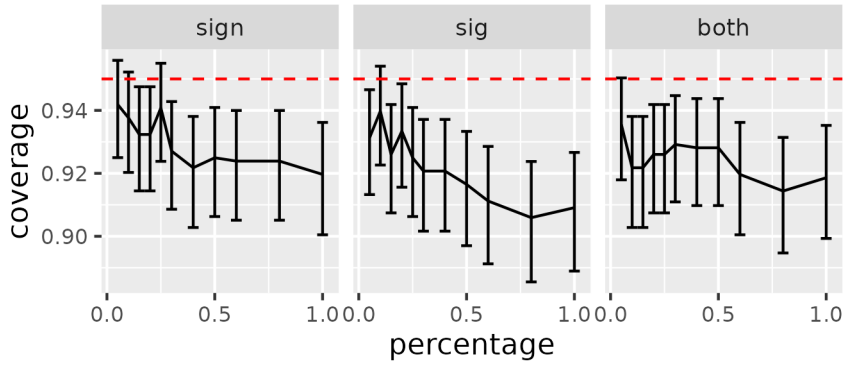

Figure 11. (Hierarchical model on subsampled tree mortality) Monte Carlo estimate of coverage of confidence interval for sum-of-influence. See Figure 6 for the meaning of the panels and the distinguished lines.

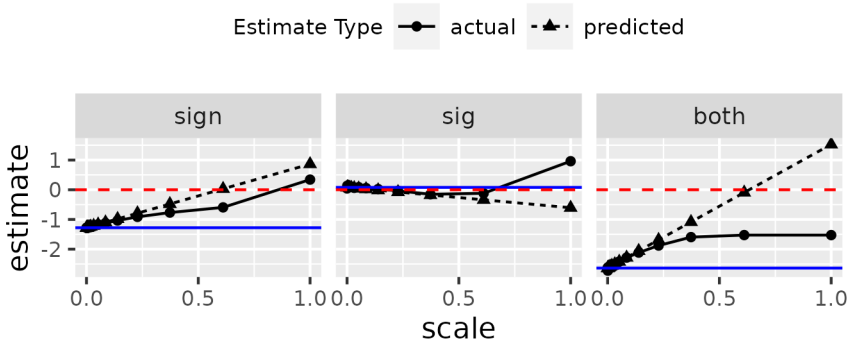

Figure 12. (Hierarchical model on subsampled tree mortality) Quality of linear approximation. See Figure 7 for the meaning of the panels and the distinguished lines.