

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Practically Feasible Solutions to a Set of Problems in Applied Statistics

Permalink

<https://escholarship.org/uc/item/1jz8f8rv>

Author

Ji, Feng

Publication Date

2022

Peer reviewed|Thesis/dissertation

Practically Feasible Solutions to a Set of Problems in Applied Statistics

by

Feng Ji

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sophia Rabe-Hesketh, Chair
Assistant Professor Jingshen Wang
Associate Professor Zachary Pardos

Spring 2022

Practically Feasible Solutions to a Set of Problems in Applied Statistics

Copyright 2022

by

Feng Ji

Abstract

Practically Feasible Solutions to a Set of Problems in Applied Statistics

by

Feng Ji

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Sophia Rabe-Hesketh, Chair

Many methodological advances in the statistics and machine learning literature are not accessible to substantive researchers and applied statisticians because of the highly technical and abstract writing and the lack of user-friendly tools for implementing the methods in practice. This dissertation provides practically feasible solutions for three challenges in applied statistics ranging from diagnosing and handling violations of the missing at random (MAR) assumption typically needed for valid statistical inference, evaluating multilevel models, and estimating standard errors for Bayesian quantile regression. All three solutions are generally applicable, even beyond the situations considered in the dissertation, all are presented in a relatively non-technical, intuitive way, and all are easy to implement with minimal modification of existing data analytic workflow and with available statistical software.

Chapter 1 proposes a generally-applicable and easy-to-implement test-based data-deletion estimator that first uses diagnostic tests to determine which MAR violations appear to be present and then proceeds with the corresponding data-deletion estimator. A data-deletion estimator starts by deleting observations for some variables for some individuals and then proceeds with standard (full information) maximum likelihood estimation based all available data. We start with a brief review of different missing data mechanisms with their corresponding graphical-model representations (i.e., directed acyclic graphs). We then examine the asymptotic and finite sample performance of data-deletion estimators for structural equation modeling (SEM), using a population study and a Monte Carlo study, under both MAR and different violations of MAR. Because the data-deletion estimators are tailored for specific MAR violations, we need a way to diagnose these violations and hence propose corresponding parametric and nonparametric diagnostic tests. We show that the tests have nominal Type I error rates and that their statistical power depends on the strength of MAR violation and proportion of missing data, two conditions that also determine the bias of naive

estimators. We therefore develop a test-based data-deletion estimator in which the outcome of the test for MAR violations determines the type of estimator that is used. We compare the test-based deletion estimator with some popular existing estimators through a series of simulation studies under various missing data mechanisms. Simulations show that the test-based data-deletion estimator outperforms other estimators considered when the missing data problem is severe and performs similarly otherwise, including under MAR.

Chapter 2 develops approximate resampling-based standard errors for Bayesian quantile regression. Quantile regression is useful in modelling treatment heterogeneity across different quantiles of the outcome distribution. One simple and popular Bayesian implementation uses the asymmetric Laplace distribution as working likelihood. This distribution is justified only because its maximum coincides with the minimum of the loss function used in classical estimation, and its scale parameter is arbitrary and not estimable from data. Unfortunately, the model-implied credible intervals are sensitive to the choice of the scale parameter and do not have the desirable frequentist nominal coverage. Resampling methods, on the other hand, have a long history in statistics for a variety of applications including variance estimation. Computationally intensive resampling (e.g., bootstrap) can be practically prohibitive for Bayesian data analysis given the computational cost of Bayesian estimation methods. We start by introducing computationally-efficient approximations for parameter estimates in resampled data, both in frequentist and Bayesian contexts. Then we develop corresponding approximate resampling-based standard errors for Bayesian asymmetric Laplace likelihood quantile regression, for both independent and clustered data (e.g., patients nested in clinics, pupils nested in schools, longitudinal/repeated measurements nested in individuals). We provide empirical evidence of the near nominal coverage of intervals based on the approximate standard errors under a wide range of conditions, as well as failure of the approach in some conditions.

Chapter 3 proposes a leave-one-out information criterion for multilevel models, also known as hierarchical or mixed-effects models. These models are used for clustered data and include random effects, such as random intercepts and random slopes, to represent between-cluster heterogeneity and induce within-cluster dependence. Model comparison and evaluation are important topics in the multilevel modelling literature. We start by introducing the target quantity of the Akaike information criterion (AIC) and linking it to leave-one-out cross validation, based on classic results. We then propose a leave-one-out information criterion (LOOIC) for independent data using the idea of approximate leave-one-out cross validation. We perform simulations to compare LOOIC with its exact counterpart and AIC, both in terms of model selection performance and in terms of how well they approximate the target quantity of the AIC. For clustered data and multilevel models, model comparison and evaluation are more complicated because predictions for left-out units can either be conditional on random effects (i.e., assume that the left-out unit belongs to the same cluster, in which

case the responses of other units in the cluster are informative), or marginal with respect to random effects (i.e., assume that the left-out unit belongs to a new cluster). Additionally, we can leave out a single unit or an entire cluster. We therefore discuss two variants of leave-one-out cross validation, leave-one-unit out (LOuO) cross validation and leave-one-cluster out (LOcO) cross validation, which correspond to conditional AIC (cAIC) and marginal AIC (mAIC), and propose the respective information criteria, LOuOIC and LOcOIC. We conduct a simulation study to compare LOcOIC with its exact counterpart and with mAIC. LOcOIC and mAIC have largely similar finite sample behavior. However, the idea of constructing approximate leave-one-out information criteria for multilevel models can be easily extended to other models and alternative ways of defining predictive errors.

To my parents
邓静 & 李舒民

Contents

Contents	ii
List of Figures	iv
List of Tables	v
1 Diagnosing and Handling Common Violations of Missing at Random	1
1.1 Introduction	1
1.2 Missingness mechanisms, recoverability and estimators	2
1.3 Performance of estimators	8
1.4 Diagnostic tests for violation of R-MAR	15
1.5 Test-based estimator	18
1.6 Concluding remarks	21
1.7 Appendix A: Derivation of IPW estimators	23
1.8 Appendix B: $\rho_{M_x^*Y X,Z}$, $\rho_{M_y^*X Z}$, P_x , and P_y in terms of model parameters . .	24
1.9 Appendix C: Performance of SEM, SEM-X, SEM-Y, and SEM-XY	26
1.10 Appendix D: Performance of test-based estimator	31
1.11 References	35
2 Approximate Resampling Standard Errors for Bayesian Quantile Regression	38
2.1 Introduction	38
2.2 Methodology	40
2.3 Bayesian ALD quantile regression	46
2.4 Bayesian ALD quantile regression with clustered data	56
2.5 Concluding Remarks	73
2.6 References	74
3 A Leave-one-out Information Criterion for Multilevel Models	79
3.1 Introduction	79
3.2 Connection between AIC and LOO	81

3.3	Approximation for eLOOIC	87
3.4	Study I: comparing LOOIC, eLOOIC and AIC with the target quantity . . .	90
3.5	Evaluation of multilevel models for clustered data	95
3.6	Study II: Comparing LOcOIC, eLOCOIC and mAIC for clustered data . . .	100
3.7	Concluding Remarks	105
3.8	References	106

List of Figures

1.1	Missingness graphs (DAGs) for common missingness mechanisms.	4
1.2	Estimated bias and mean squared error (MSE) under MNAR-X for $N = 500$. .	12
1.3	Estimated bias and mean squared error under MNAR-Y for $N = 500$	13
1.4	Estimated bias and mean squared error (MSE) under MNAR-XY for $N = 500$.	14
1.5	Power of test of R-MAR against $H_1^{\text{MNAR-Y}}$	18
2.1	Empirical coverage for different SE methods for $N = 200$	50
2.2	Empirical coverage for different SE methods for $N = 500$	51
2.3	Empirical coverage for different SE methods for $N = 1000$	52
2.4	Empirical coverage with frequentist point estimates for $N = 200$	53
2.5	Empirical coverage with frequentist point estimates for $N = 500$	54
2.6	Empirical coverage with frequentist point estimates for $N = 1000$	55
2.7	Empirical coverage with frequentist point estimates for $I = 10, J = 10$	61
2.8	Empirical coverage with frequentist point estimates for $I = 10, J = 50$	62
2.9	Empirical coverage with frequentist point estimates for $I = 10, J = 100$	63
2.10	Empirical coverage with frequentist point estimates for $I = 30, J = 10$	64
2.11	Empirical coverage with frequentist point estimates for $I = 30, J = 50$	65
2.12	Empirical coverage with frequentist point estimates for $I = 30, J = 100$	66
3.1	Comparison between exact-calculated estimates with its IJ approximation . . .	90
3.2	Finite sample bias estimates for AIC, eLOOIC, and LOOIC	92
3.3	Root mean squared error estimates for AIC, eLOOIC, and LOOIC	93
3.4	Finite sample bias estimates for mAIC, eLOcOIC, and LOcOIC	102
3.5	Root mean squared error estimates for mAIC, eLOcOIC, and LOcOIC	103

List of Tables

1.1	Estimates and $(100 \times \text{SE})$ for population study.	10
1.2	Decision rules for test-based estimator, where H_0 is that R-MAR holds.	19
1.3	Bias $(\times 1000)$ for SEM and test-based estimator across conditions.	20
1.4	MSE $(\times 1000)$ for SEM and test-based estimator across conditions.	20
1.5	Bias $(\times 1000)$ of SEM and SEM-X under MNAR-X.	26
1.6	MSE $(\times 1000)$ of SEM and SEM-X under MNAR-X.	27
1.7	Bias $(\times 1000)$ of SEM and SEM-Y under MNAR-Y.	27
1.8	MSE $(\times 1000)$ of SEM and SEM-Y under MNAR-Y.	28
1.9	Bias $(\times 1000)$ of SEM and SEM-XY under MNAR-XY.	29
1.10	MSE $(\times 1000)$ of SEM and SEM-XY under MNAR-XY.	30
1.11	Bias $(\times 1000)$ of SEM and test-based estimator (based on LRT) under MNAR-X.	31
1.12	MSE $(\times 1000)$ of SEM and test-based estimator (based on LRT) under MNAR-X.	32
1.13	Bias $(\times 1000)$ of SEM and test-based estimator (based on LRT) under MNAR-Y.	32
1.14	MSE $(\times 1000)$ of SEM and test-based estimator (based on LRT) under MNAR-Y.	33
1.15	Bias $(\times 1000)$ of SEM and test-based estimator (based on LRT) under R-MAR.	33
1.16	MSE $(\times 1000)$ of SEM and test-based estimator (based on LRT) under R-MAR.	34
2.1	EML and ECP for unclustered data when $\sigma = .1$	56
2.2	EML and ECP for unclustered data when $\sigma = .2$	57
2.3	EML and ECP for unclustered data when $\sigma = .5$	58
2.4	EML and ECP for unclustered data when $\sigma = 1$	59
2.5	EML and ECP for unclustered data when $\sigma = 2$	60
2.6	EML and ECP for unclustered data when $\sigma = 5$	67
2.7	EML and ECP for unclustered data when $\sigma = 10$	68
2.8	EML and ECP for clustered data when $\sigma = .1$	69
2.9	EML and ECP for clustered data when $\sigma = .5$	70
2.10	EML and ECP for clustered data when $\sigma = 1$	71
2.11	EML and ECP for clustered data when $\sigma = 5$	72
3.1	Comparing AIC, eLOOIC, and LOOIC	94
3.2	Comparing mAIC, eLOcOIC, and LOcOIC	104

Acknowledgments

I would love to express my deepest gratitude to my advisor, Sophia Rabe-Hesketh. I am enormously grateful to Sophia for her intellectual guidance throughout my doctoral studies, for her open-mindedness and academic curiosity in supporting me to explore many interests across topics, for her kindness and great sense of humor, for her infinite patience and the countless hours she put into my education, and for profoundly shaping my ways of thinking and solving problems. She truly fostered my growth and independence as a scholar and beyond. A simple paragraph is just not enough for all that Sophia has taught me. Working with Sophia has truly been a life-changing experience for me. It has been my great privilege and fortune to have her as my advisor. I will forever cherish what I have learned from her.

My heartfelt gratitude also goes to Jingshen Wang, for her unwavering support, continuous encouragement, and candid advice on many fronts. I wholeheartedly appreciate all the wonderful opportunities that Jingshen created for me. I would also like to sincerely thank Zachary Pardos, Alan Hubbard, and Anders Skrondal for their extensive knowledge and constructive advice as well as valuable contribution to my studies.

I am also incredibly grateful to my mentors outside of Berkeley and my collaborators, including but not limited to Joel Schneider, Mark Lai, Benjamin Deonovic, Gunter Maris, Karen Draney, Anja Ernst, Peng Yin, Heyuan Liu, Dexin Shi, Yuting Han, Jimmy de la Torre, Ran Liu, Bowen Wang, Scott Jordan, Irimi Moustaki, Roy Levy, Xinwei Ma, Gonzalo Vazquez-Bare, and Jinbo He. It has been my great pleasure working with them throughout my academic journey and I have learned so much from all of them.

I would also like to warmly thank many many of my friends whom I met on various occasions including Sophia's research group, Jingshen's research group, the biostatistics program, the SRM/QME program, BPH & GSE, the International House, UC Davis, as well as ACT & Google during my internships. They supported, accompanied, and inspired me in various forms during my doctoral journey. I have learned so much from them and am equally grateful for their presence in my life.

Finally, I heartily thank my friends back home and met prior to Cal, for their belief and support across time and space. Endless thanks to my family and especially my parents, Ji Shumin and Deng Jing, who do not understand my research but always try to support me in any way they possibly can. It is their tremendous love and sacrifice that allow them to send their only child to study 10138.87 km away from home. Special thanks to CPP, for being my best friend, always cheering me up, and the unconditional love. There is so much that I would like to say but so little that can be expressed – CPP, you are my consistency when I deal with all the randomness in my research and life.

Chapter 1

Diagnosing and Handling Common Violations of Missing at Random

1.1 Introduction

A common approach for handling missing data in regression analysis (e.g., linear, logistic, multilevel) is complete-case analysis or listwise deletion. Alternatively, regression models are sometimes embedded within multivariate models in order to analyze all available data by maximum likelihood (ML) estimation, Bayesian analysis, or multiple imputation. Following Little and Zhang (2011), we use the term “ignorable likelihood” (IL) for such approaches. Missing at random (MAR) and missing completely at random (MCAR) assumptions, originating from Rubin (1976), are often invoked as justifications for IL approaches, but these assumptions tend to be misunderstood as pointed out by Seaman et al. (2013) and Rabe-Hesketh and Skrondal (2022). When data are believed to be missing not at random (MNAR), it is commonly believed that it becomes necessary to model the missingness process explicitly. However, as demonstrated by Rabe-Hesketh and Skrondal (2015, 2022), valid inferences can often be obtained by making slight modifications to data, models, or estimators, such as data deletion followed by IL methods.

In groundbreaking work, Mohan et al. (2013) and Mohan and Pearl (2021) provide a framework for understanding missing data problems based on directed acyclic graphs (DAGs). Their DAGs, called m -graphs (m for “missing data”), represent the assumed relations among the variables of interest and include paths from these variables to missing data indicators to encode the (conditional) independence assumptions for the missing data mechanism. These representations make assumptions explicit and allow recoverability (or identifiability) of target quantities and testability of assumptions to be determined by applying graph theory. If target quantities can be estimated consistently based on the observed data, m -graphs also help to derive procedures for estimating these quantities.

Following this line of research, there have been many advances in handling missing data problems, primarily in computer science. This work is usually nonparametric and implicitly assumes that all variables are categorical so that the target quantities are joint and conditional probabilities that can be estimated by the corresponding sample proportions. In contrast, in this paper we use m -graphs to derive and justify parametric inference with missing data. For concreteness, we focus on linear regression and structural equation modeling (SEM) and on three types of missing not at random (MNAR) processes: (a) an explanatory variable directly affects missingness of a response variable (also when it is not observed); (b) a response variable directly affects missingness of an explanatory variable (also when it is not observed); and (c) both of these MAR violations are present.

For situations (a) and (b), Rabe-Hesketh and Skrondal (2015, 2022) show that a consistent estimator can be obtained by creating more missing data followed by IL methods. They demonstrate that their data-deletion approach can be viewed as making the process MAR. Rabe-Hesketh and Skrondal (2022) point out that this approach is closely related to the ordered factorization theorem of Mohan et al. (2013) but preferable for parametric models. Situation (c) can be addressed by an inverse probability weighted (IPW) estimator that we derive in Section 1.2 of this article.

In Section 1.2 we also provide a brief review of m -graphs and describe consistent estimators for the three types of MNAR considered. Thereafter, we use a population study in Section 1.3 to show that the estimators proposed for the above situations are actually consistent and a Monte Carlo study to investigate finite sample bias and mean squared error (MSE).

Because addressing MAR violations (a) and (b) requires different estimators, it is important to be able to diagnose which of the violations is present. We therefore propose both parametric and nonparametric diagnostic tests in Section 1.4. Simulations are used to assess the Type I error rates and power of the tests as a function of the strength of MAR violation, proportion of missing data, and sample size.

In Section 1.5 we develop a novel test-based estimator that first uses diagnostic tests for specific MAR violations and then proceeds with the estimator that is valid for that violation. The finite sample performance of the proposed test-based estimator is compared with the naive IL estimator under various missing data mechanisms. Finally, we close the article with some concluding remarks.

1.2 Missingness mechanisms, recoverability and estimators

We follow the terminology and notation of Rabe-Hesketh and Skrondal (2015, 2022) and Mohan and Pearl (2021) in this paper. An m -graph is a type of DAG that encodes relations

among multiple variables and their missingness (or selection) indicators, so that conditional independence can be inferred based on d -separation (e.g., Geiger et al., 1990).

Figure 1.1 shows m -graphs for different types of missingness mechanisms. Here Y is the response variable and X and Z are two explanatory variables. Whereas Z is always observed (filled circle), X and Y can be missing (hollow circles) with corresponding selection indicators, S^x and S^y , equal to 1 if the variable is observed and 0 if missing. In all cases, we assume that $(X_i, Z_i, Y_i, S_i^x, S_i^y)$ for unit i are independently and identically distributed and henceforth omit the i subscript for convenience. We let X^* denote a fully observed “proxy” variable, equal to “NA” when X is missing, so that $X^* = S^x X + (1 - S^x) \text{NA}$, and analogously for Y^* . For example, if the four realized values of X that would have been observed if there were no missing data are (3, 4, 6, 5), but the last value is missing, then the realized values of S^x are (1, 1, 1, 0) and the values of X^* are (3, 4, 6, NA).

In our setting, the definition of MAR in Mohan et al. (2013) and Mohan and Pearl (2021) is $(S^x, S^y) \perp\!\!\!\perp (X, Y) | Z$. More generally, missingness cannot depend on variables that can be missing (here X and Y), given the variables that cannot be missing (here Z). To emphasize how this MAR definition differs from Rubin’s (1976) original MAR notion, Rabe-Hesketh and Skrondal (2022) call it *realistic* missing at random (R-MAR), and we use this acronym henceforth. Rabe-Hesketh and Skrondal (2022) also relate R-MAR to Seaman et al.’s (2013) *everywhere* MAR and Pothoff et al.’s (2006) MAR+.

In Figure 1.1, Y could be income, X introversion and Z age. Income and introversion are based on survey responses, whereas age is available from an administrative database and therefore never missing. The m -graph in the top left panel includes paths $Z \rightarrow S^x$ and $Z \rightarrow S^y$, which encode that age (Z) causes the missingness of the covariate (X) and the outcome (Y). For instance, senior citizens may be more reluctant to disclose their income or complete a personality questionnaire. Because Z is always observed (no missing values), this situation corresponds to R-MAR.

We refer to the other situations in the figure as MNAR because R-MAR is violated. In the bottom left panel, R-MAR is violated because of one path, $X \rightarrow S^y$, where X is not always observed, and we call this situation MNAR-X. The path $X \rightarrow S^y$ could be due to more introvert individuals having a tendency not to report their income, given their age. MNAR-Y, shown in the bottom right panel of Figure 1.1, violates R-MAR because of the path $Y \rightarrow S^x$. For example, those who earn more may be less willing to provide their personal information, given age. Both MNAR-X and MNAR-Y can be handled by data deletion and standard IL based on all remaining data, as shown by Rabe-Hesketh and Skrondal (2022) and described in Section 1.2.1.

In the top right panel of Figure 1.1, the paths $X \rightarrow S^y$ and $Y \rightarrow S^x$ are both present, and we call this situation MNAR-XY. This situation was not considered by Rabe-Hesketh and Skrondal (2022).

We do not consider R-MAR violations in the form of the paths $X \rightarrow S^x$ and $Y \rightarrow S^y$. While $X \rightarrow S^x$ can be handled by simply conditioning on X (i.e., making inferences regarding

$P(Y, Z|X)$ for the subset of data where X is observed), more complex solutions work in specific situations when $Y \rightarrow S^y$ (e.g., Skrongdal and Rabe-Hesketh, 2014; Mohan, 2018).

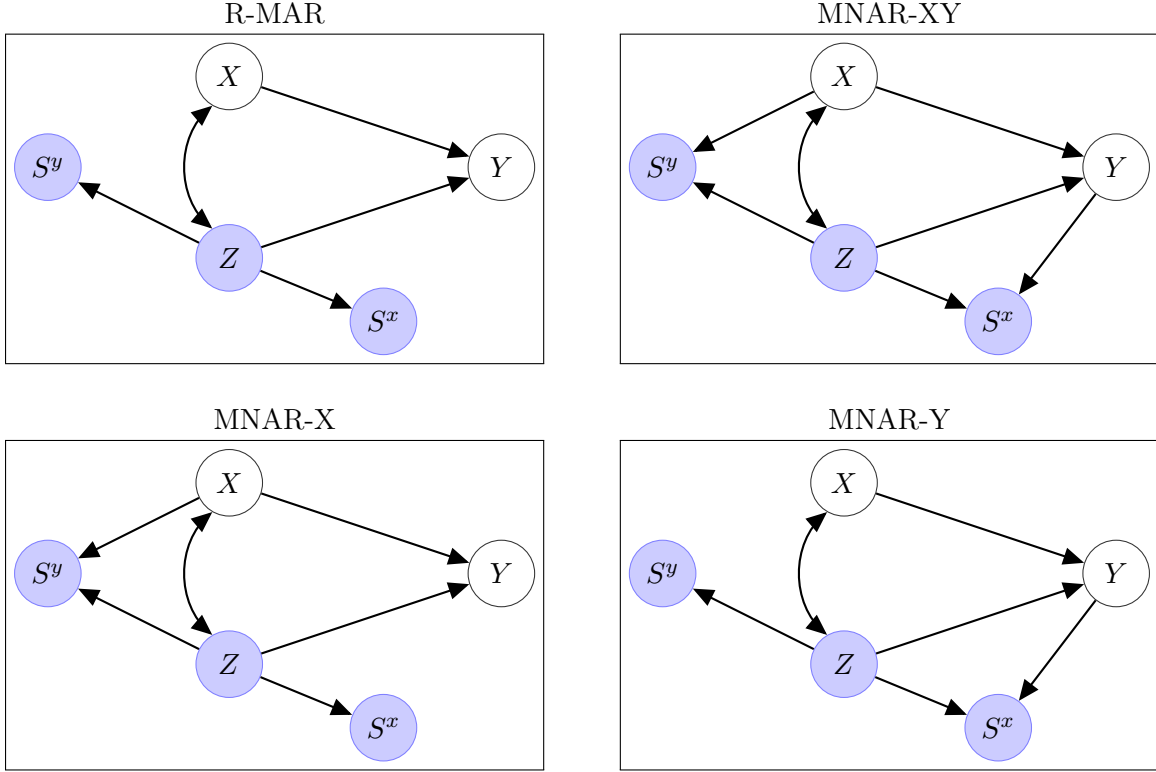


Figure 1.1: Missingness graphs (DAGs) for common missingness mechanisms.

Mohan and Pearl (2021) use graphs to derive whether joint and conditional distributions, such as $P(X, Z, Y)$ and $P(Y|X, Z)$, can be recovered (i.e., can be estimated consistently) when the missingness process is compatible with the graphs. They implicitly assume that all variables are categorical and that the distributions can be estimated from cross-tabulations. Here we assume that we have specified a correct parametric model for continuous variables, X , Z , and Y , namely a linear SEM compatible with the graphs in Figure 1.1. Our interest is in consistent estimation of the model parameters as in Mohan et al. (2018).

1.2.1 Estimators under MNAR-X

1.2.1.1 Joint distribution

We start by showing how to recover the joint distribution $P(X, Z, Y)$ under MNAR-X, originally discussed (without Z) in Example 1 by Mohan and Pearl (2021):

$$\begin{aligned} P(X, Z, Y) &= P(Z)P(X|Z)P(Y|X, Z) \\ &= \underbrace{P(Z)}_{a^A} \underbrace{P(X^*|S^x = 1, Z)}_{b^A} \underbrace{P(Y^*|S^x = 1, S^y = 1, X^*, Z)}_{c^A}. \end{aligned} \quad (1.1)$$

From the graph in Figure 1.1, we observe that S^x and X are independent conditional on Z (i.e., $S^x \perp\!\!\!\perp X|Z$), so that $P(X|Z) = P(X|S^x = 1, Z) = P(X^*|S^x = 1, Z)$. We also observe that $(S^x, S^y) \perp\!\!\!\perp Y|X, Z$, so that $P(Y|X, Z) = P(Y|S^x = 1, S^y = 1, X, Z) = P(Y^*|S^x = 1, S^y = 1, X^*, Z)$. To recover the joint distribution, we therefore proceed sequentially according to (1.1). In step a^A , we estimate $P(Z)$ using all units (Z is never missing); in step b^A , we include units for which X is observed, $S^x = 1$; and in step c^A , we include units for which both X and Y are observed, $S^x = S^y = 1$. This is an example of Mohan et al. (2013)'s ordered factorization.

Rabe-Hesketh and Skrondal (2015, 2022) suggest a data-deletion approach briefly described here. If S^x and S^y satisfy MNAR-X, then the selection process for Y is modified by deleting Y when $S^x = 0$, resulting in a modified selection indicator $\dot{S}^y$,

$$\dot{S}^y = \begin{cases} S^y & \text{if } S^x = 1 \\ 0 & \text{if } S^x = 0 \end{cases}.$$

Now selection of Y depends on X only when X is observed and does not depend on any missing data, as shown by conditioning on X^* (sometimes denoted X^{obs} in the missing data literature):

$$\text{M-MAR: } P(S^x, \dot{S}^y|X, Z, Y) = P(S^x|Z)P(\dot{S}^y|S^x, X^*, Z). \quad (1.2)$$

We have therefore *made* the missingness MAR and the idea is referred to as M-MAR. If the data deletion is performed in repeated samples, the modified missingness process is everywhere MAR (as defined in Seaman et al., 2013). IL methods for all remaining data hence have the desired frequentist properties (Seaman et al., 2013).

When X , Z , and Y follow a SEM, we call the data-deletion ML estimator SEM-X. Using the factorization in the first line of (1.1) to express the data-deletion log-likelihood in terms of the three probabilities, Rabe-Hesketh and Skrondal (2022) show that the subsets of units that contribute information on the parameters for each of these probabilities correspond to the subsets of units used in the sequential estimation procedure by Mohan et al. (2013). The advantage of data deletion is that we can proceed with our usual, nonsequential IL method after deleting some data and obtain standard errors as a byproduct.

1.2.1.2 Conditional distribution

It is evident from c^A in Equation (1.1) that the conditional distribution $P(Y|X, Z)$ can be recovered from complete cases for which $C = S^x = S^y = 1$. The m -graph satisfies *conditional* MAR (C-MAR) as defined by Rabe-Hesketh and Skrondal (2022) because $C \perp\!\!\!\perp Y|X, Z$. This condition for valid inference is also discussed by Little (1992) and corresponds to Example 2 in Mohan and Pearl (2021).

1.2.2 Estimators under MNAR-Y

1.2.2.1 Joint distribution

To recover the joint distribution under MNAR-Y in Figure 1.1, we can again use sequential estimation, now based on the ordered factorization

$$\begin{aligned} P(X, Z, Y) &= P(Z)P(Y|Z)P(X|Z, Y) \\ &= \underbrace{P(Z)}_{a^B} \underbrace{P(Y^*|S^y = 1, Z)}_{b^B} \underbrace{P(X^*|S^x = 1, S^y = 1, Z, Y^*)}_{c^B}. \end{aligned} \quad (1.3)$$

In step a^B , we estimate $P(Z)$ using all units (Z is never missing), in step b^B , we include units where $S^y = 1$; and in step c^B , we include units where $C = 1$.

After discarding values of X for units whose Y is missing, and defining the new selection indicator $\dot{S}^x$ as in Section 1.2.1.1 but now with X and Y interchanged, we obtain

$$\text{M-MAR: } P(\dot{S}^x, S^y|X, Z, Y) = P(S^y|Z)P(\dot{S}^x|S^y, Z, Y^*). \quad (1.4)$$

Again, the modified missingness process is everywhere MAR, and we can proceed with IL methods. When the assumed model for X , Z , and Y is a SEM, we refer to the data-deletion ML estimator as SEM-Y.

1.2.2.2 Conditional distribution

Unfortunately, the conditional distribution of interest $P(Y|X, Z)$ does not appear directly in the factorization in (1.3). Although the conditional distribution can be derived from the joint distribution by dividing by the marginal distribution $P(X, Y)$, the resulting expression (which may not have a closed form) will generally not be a straightforward function of the model parameters for $P(Y|X, Z)$. Therefore, SEM-Y is highly preferable to sequential estimation here. Alternatively, an IPW estimator can be derived as shown for a more complex case in Section 1.2.3 and Appendix A.

1.2.3 Estimators under MNAR-XY

1.2.3.1 Joint distribution

Mohan and Pearl (2021), Example 5, derive an expression for the joint distribution (without Z) that is a weighted version of the joint distribution in the complete-case sample. However, it is not clear how to apply this result unless all variables are categorical and the goal is to estimate the probabilities of the X by Z by Y table.

Instead, we use an estimating equation approach. Let $m(X, Z, Y)$ be the vector of score contributions for a unit based on the joint likelihood. For a correctly specified model with no missing data, $E_{X,Z,Y}[m(X, Z, Y)] = 0$ when evaluated at the correct parameter values. To obtain a consistent estimator of the parameters for $P(X, Z, Y)$ using complete cases (with $C = S^x S^y = 1$), we find inverse weights $\pi \equiv \pi(X, Z, Y)$ so that $E_{S^x, S^y, X, Z, Y} \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) \right] = 0$ when evaluated at the correct parameter values. We show in Appendix A that

$$E \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) \right] = E_{X,Z,Y} \left[\frac{1}{\pi} P(S^x = 1 | Z, Y) P(S^y = 1 | X, Z) m(X, Z, Y) \right],$$

where we suppressed the subscripts for the expectation when it is over the selection indicators and the variables X , Z , and Y that are not explicitly conditioned on. Therefore, $\pi = P(S^x = 1 | Z, Y) P(S^y = 1 | X, Z)$ and both probabilities, assumed to be positive, can be estimated in the complete-case sample by using, for instance, logistic regression. The inverse of the product of these estimates is then used as weights in pseudo maximum likelihood estimation. We call this IPW estimator SEM-XY. The estimator is also consistent under MNAR-X and MNAR-Y because these mechanisms are special cases of MNAR-XY.

1.2.3.2 Conditional distribution

We now let $m(X, Z, Y)$ be the vector of score contributions for a unit based on the likelihood of Y given X and Z . Without missing data, we assume that $E_{Y|X,Z}[m(X, Z, Y)] = 0$ at the correct parameter values. To estimate a model for $P(Y|X, Z)$ using complete cases when there are missing data, we want to find $\pi \equiv \pi(X, Z, Y)$ so that

$E_{S^x, S^y, Y|X,Z} \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) | X, Z \right] = 0$ at the correct parameter values. Appendix A shows that we obtain

$$E \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) | X, Z \right] = P(S^y = 1 | X, Z) E_{Y|X,Z} \left[\frac{1}{\pi} P(S^x = 1 | Z, Y) m(X, Z, Y) \right].$$

Assuming $P(S^y = 1 | X, Z) > 0$, the estimating equation becomes

$E_{Y|X,Z} \left[\frac{1}{\pi} P(S^x = 1 | Z, Y) m(X, Z, Y) \right] = 0$, so $\pi = P(S^x = 1 | Z, Y)$, also assumed to be positive. For linear regression, the pseudolikelihood estimator is weighted least squares (WLS), and we call the estimator WLS-X because the inverse weights are estimates of $P(S^x = 1 | Z, Y)$.

1.3 Performance of estimators

1.3.1 Population study of asymptotic performance

We evaluated the performance of six estimators for a simulated population-scale dataset where missingness mechanisms are compatible with the DAGs in Figure 1.1. The model for $P(X, Z, Y)$ is a SEM as described in Section 1.3.1.1. The six estimators are:

1. OLS for $P(Y|X, Z)$: OLS with complete cases
2. SEM for $P(X, Z, Y)$: SEM by MLE (with all available data)
3. SEM-X for $P(X, Z, Y)$: SEM by MLE after discarding Y when X is missing
4. SEM-Y for $P(X, Z, Y)$: SEM by MLE after discarding X when Y is missing
5. WLS-X for $P(Y|X, Z)$: WLS with weights for selection of X
6. SEM-XY for $P(X, Z, Y)$: SEM by pseudo MLE with weights for selection of X and Y .

1.3.1.1 Simulation design

For X and Z , we specify a bivariate normal distribution:

$$\begin{pmatrix} X \\ Z \end{pmatrix} \stackrel{i.i.d.}{\sim} N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{XX} & \psi_{XZ} \\ \psi_{ZX} & \psi_{ZZ} \end{pmatrix} \right], \quad (1.5)$$

where $\psi_{XX} = \psi_{ZZ} = 1$ and $\psi_{XZ} = 0.5$. Y is simulated from the linear model

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \epsilon, \quad \epsilon|X, Z \sim N(0, \sigma^2), \quad (1.6)$$

where $\beta_0 = \beta_X = \beta_Z = \sigma^2 = 1$.

The missingness indicators $M_x = 1 - S^x$ and $M_y = 1 - S^y$ are simulated from probit models:

$$M_y^* = \gamma_0 + \gamma_X X + \gamma_Z Z + u_0, \quad u_0|X, Z \sim N(0, 1) \quad (1.7)$$

$$M_x^* = \alpha_0 + \alpha_Z Z + \alpha_Y Y + u_1, \quad u_1|Z, Y \sim N(0, 1), \quad (1.8)$$

where

$$1 - S^x = M_x = \begin{cases} 1, & \text{if } M_x^* > 0 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad 1 - S^y = M_y = \begin{cases} 1, & \text{if } M_y^* > 0 \\ 0, & \text{otherwise} \end{cases}.$$

We set $\gamma_Z = \alpha_Z = 1$ and express the strength of dependence represented by the paths $X \rightarrow S^y$ and $Y \rightarrow S^x$ by the partial correlations, $\rho_{M_y^* X} \equiv (M_y^*, X|Z)$ and $\rho_{M_x^* Y} \equiv$

$(M_x^*, Y|X, Z)$, respectively. Appendix B gives expressions for these correlations that allow us to solve for γ_X and α_Y . We simulate four missingness processes defined by the four combinations of $\rho_{M_y^*X} \in \{0, 0.6\}$ and $\rho_{M_x^*Y} \in \{0, 0.6\}$. For MNAR-X, $\rho_{M_y^*X} = 0.6$ and $\rho_{M_x^*Y} = 0$; for MNAR-Y, $\rho_{M_y^*X} = 0$ and $\rho_{M_x^*Y} = 0.6$; for MNAR-XY, $\rho_{M_y^*X} = 0.6$ and $\rho_{M_x^*Y} = 0.6$ and for R-MAR, $\rho_{M_y^*X} = 0$ and $\rho_{M_x^*Y} = 0$. Appendix B also gives expressions for the marginal probabilities of observing X and Y , $P_x = P(S^x = 1)$ and $P_y = P(S^y = 1)$, which were set to 0.8 to solve for γ_0 and α_0 .

Datasets of size $N = 5,000,000$ were simulated with `lavaan` (Rosseel, 2012) in **R** (R Core Team, 2021). The IPW estimator was based on the correct probit models and we implemented WLS-X with `survey` (Lumley, 2019), and SEM-XY with `ipw` (van der Wal and Geskus, 2011) and `lavaan.survey` (Oberski, 2014). Sandwich estimators were used for the standard errors (SEs) of IPW estimators.

1.3.1.2 Results

Table 1.1 reports the point estimates of β_X and β_Z for all estimation methods and the estimates of ψ_{XZ} for the SEM estimators. The original estimated SEs have been multiplied by 100 and are presented in parentheses. The asymptotic SEs for any sample size m can be obtained by multiplying the reported estimates by $\sqrt{500/m}$. Point estimates expected to be consistent are indicated by asterisks next to the estimates. SEs of the most efficient estimators are shown in bold and the percentage increase in SE compared with the smallest SE is shown in square brackets for some alternative consistent estimators for each condition. Point estimates that differ from the generating value by more than 3% are shown in italics.

We see that SEM is inconsistent when R-MAR is violated although the inconsistency of $\hat{\beta}_X$ is small. In contrast, OLS is consistent for MNAR-X (and R-MAR) because $C \perp\!\!\!\perp Y|X, Z$, the sufficient condition for recovering $P(Y|X, Z)$ using complete cases. Notably, X in this case can cause its own missingness (a condition not included in the simulation). Under R-MAR, SEM is more efficient than OLS, which is why multiple imputation is often used to handle missing covariates. The efficiency gain is not large in the current case but could be improved by using multiple imputation with auxiliary variables.

OLS and SEM-X produce identical estimates of $E(Y|X, Z)$ because (i) without missing data, OLS estimates of regression coefficients are identical to ML estimates when the likelihood is based on $P(Y|X, Z)$, and (ii) SEM-X uses data from complete cases only for the parameters of $P(Y|X, Z)$, and these parameters are distinct from the parameters of $P(X, Z)$ (whose estimates are also based on incomplete cases).

Under MNAR-X and MNAR-Y, respectively, the data-deletion estimators SEM-X and SEM-Y are consistent and have smaller population SEs and are hence more efficient than the corresponding consistent IPW estimators, WLS-X and SEM-XY. In fact, SEM-XY is very inefficient, a general problem of IPW estimators discussed in Seaman and White (2013).

Table 1.1: Estimates and (100×SE) for population study.

		MNAR-X $X \rightarrow S^y$	MNAR-Y $Y \rightarrow S^x$	MNAR-XY	R-MAR
OLS	$\hat{\beta}_X$	1.000* (0.064)	0.955 (0.061)	0.963 (0.062)	1.000* (0.062) [4]
	$\hat{\beta}_Z$	1.000* (0.070)	0.902 (0.070)	0.905 (0.067)	1.000* (0.072) [7]
SEM	$\hat{\beta}_X$	0.993 (0.062)	0.999 (0.060)	1.000 (0.060)	1.000* (0.060)
	$\hat{\beta}_Z$	0.967 (0.065)	1.012 (0.068)	0.968 (0.065)	1.000* (0.067)
	$\hat{\psi}_{XZ}$	0.483 (0.056)	0.458 (0.057)	0.432 (0.056)	0.500* (0.057)
SEM-X: discard Y when X is missing	$\hat{\beta}_X$	1.000* (0.064)	0.955 (0.061)	0.963 (0.062)	1.000* (0.062)
	$\hat{\beta}_Z$	1.000* (0.070)	0.902 (0.070)	0.905 (0.067)	1.000* (0.072)
	$\hat{\psi}_{XZ}$	0.500* (0.058)	0.392 (0.056)	0.392 (0.056)	0.500* (0.058)
SEM-Y: discard X when Y is missing	$\hat{\beta}_X$	0.998 (0.062)	1.000* (0.060)	0.986 (0.279)	1.000* (0.060)
	$\hat{\beta}_Z$	0.988 (0.065)	1.000* (0.068)	0.956 (0.280)	1.000* (0.067)
	$\hat{\psi}_{XZ}$	0.384 (0.058)	0.500* (0.063)	0.161 (0.230)	0.500* (0.057)
WLS-X: with IPW for S^x	$\hat{\beta}_X$	1.000* (0.066) [3]	1.010* (1.055) [1661]	1.000* (0.120)	1.000* (0.064) [7]
	$\hat{\beta}_Z$	1.000* (0.076) [9]	1.001* (0.539) [688]	0.996* (0.202)	1.000* (0.079) [18]
SEM-XY: with IPW for S^x times IPW for S^y	$\hat{\beta}_X$	1.000* (0.286) [347]	1.020* (2.387) [3884]	0.989* (0.557) [362]	1.000* (0.115) [92]
	$\hat{\beta}_Z$	0.999* (0.660) [841]	0.997* (1.012) [1379]	0.992* (1.381) [582]	1.000* (2.792) [316]
	$\hat{\psi}_{XZ}$	0.484* (0.793) [1267]	0.490* (2.724) [4238]	0.464* (1.852)	0.505* (0.605) [969]

Consistent estimators are denoted by *.

Bold SE is smallest SE among estimators.

Square brackets show percentage increase in SE compared with smallest SE among estimators.

We note that WLS-X is consistent for $P(Y|X, Z)$ under all missingness mechanisms considered here, and the loss in efficiency relative to the most efficient consistent alternative (when one exists) is appreciable under MNAR-Y but much less so under MNAR-X. Under MNAR-XY, WLS-X is more efficient than SEM-XY and should therefore be used if we are interested in the parameters of $P(Y|X, Z)$ only. Generally, the inconsistency of $\hat{\beta}_Z$ is more severe than that of $\hat{\beta}_X$.

1.3.2 Simulation study of finite sample performance

In this section we assess the finite-sample frequentist properties of the estimators discussed above. We limit our investigation to a comparison of SEM with each consistent estimator, i.e., SEM-X, SEM-Y and SEM-XY for MNAR-X, MNAR-Y, MNAR-XY, respectively.

1.3.2.1 Simulation design

The simulation conditions are all combinations of the following factors:

1. Strength of dependence $\rho_{M_y^*X} \in \{0, 0.5, 0.9\}$.
2. Strength of dependence $\rho_{M_x^*Y} \in \{0, 0.5, 0.9\}$.
3. Marginal probability of observing X : $P_x \in \{0.3, 0.8\}$.
4. Marginal probability of observing Y : $P_y \in \{0.3, 0.8\}$.
5. Sample size: $N \in \{200, 500, 1000\}$.

For MNAR-X, $\rho_{M_y^*X} \in \{0.5, 0.9\}$ and $\rho_{M_x^*Y} = 0$; For MNAR-Y, $\rho_{M_y^*X} = 0$ and $\rho_{M_x^*Y} \in \{0.5, 0.9\}$; For MNAR-XY, $\rho_{M_y^*X} \in \{0.5, 0.9\}$ and $\rho_{M_x^*Y} \in \{0.5, 0.9\}$.

We replicate the simulation 100 times for each condition and estimate the parameters by SEM (MLE based on all available data) as well as the SEM-X, SEM-Y or SEM-XY estimator that is consistent for that condition.

1.3.2.2 Results

The estimated bias and mean squared error are reported for $\hat{\beta}_X$, $\hat{\psi}_{XZ}$, and $\hat{\beta}_Z$ in Figures 1.2 and 1.3 for $N = 500$. Tables with these results as well as the results for $N = 200$ and $N = 1,000$ are given in Web Appendix C (Tables 1.5 to 1.10).

The upper panels in Figures 1.2 and 1.3 show that the consistent estimators (SEM-X for MNAR-X and SEM-Y for MNAR-Y) are approximately unbiased whereas SEM is biased, especially when the probabilities of being observed, P_x and P_y , are both small and the strength of dependence is large. In these conditions, the data-deletion estimators also

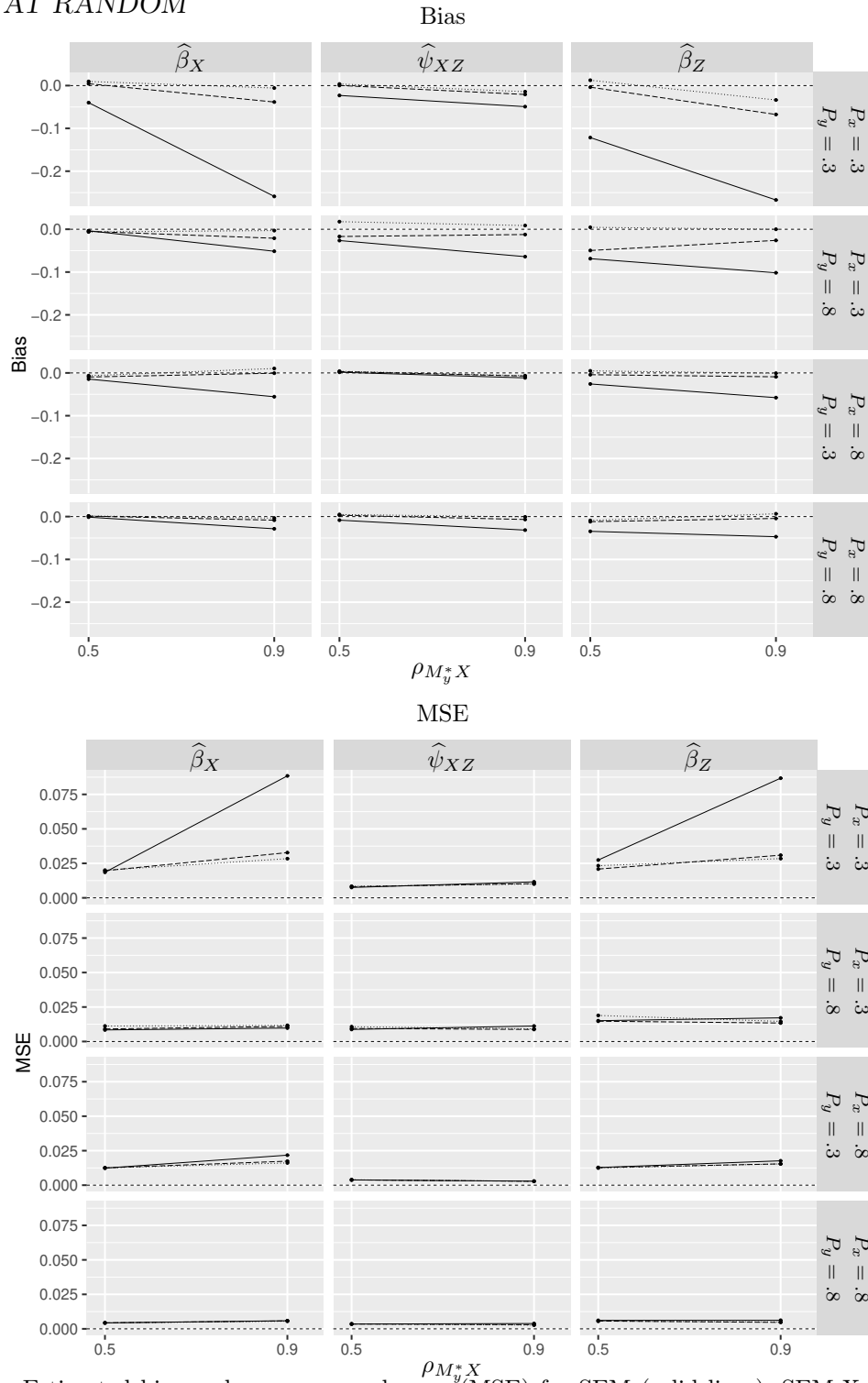


Figure 1.2: Estimated bias and mean squared error (MSE) for SEM (solid lines), SEM-X (dotted lines) and test-based (dashed lines) estimators under MNAR-X for $N = 500$.

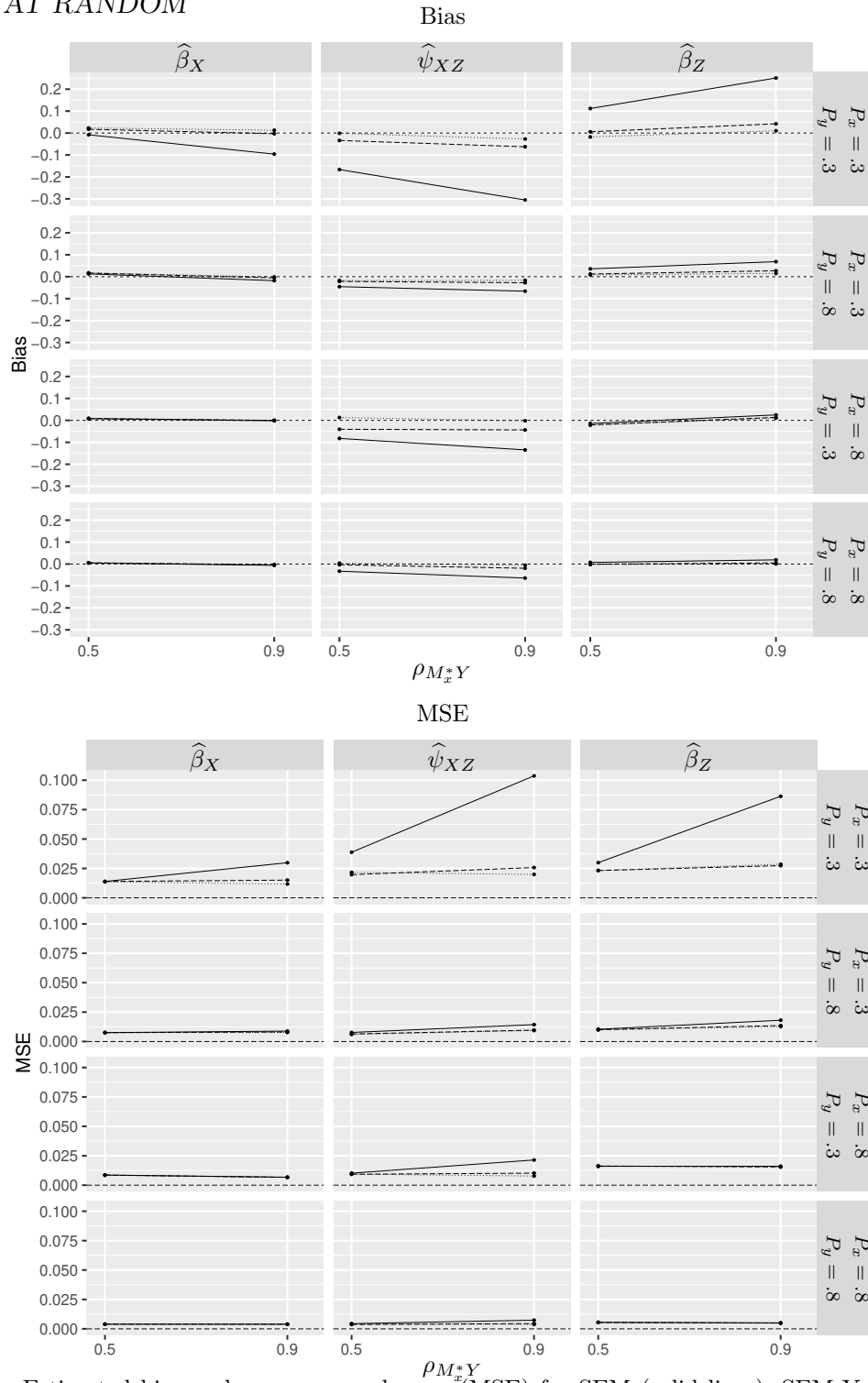


Figure 1.3: Estimated bias and mean squared error (MSE) for SEM (solid lines), SEM-Y (dotted lines) and test-based (dashed lines) estimators under MNAR-Y for $N = 500$.

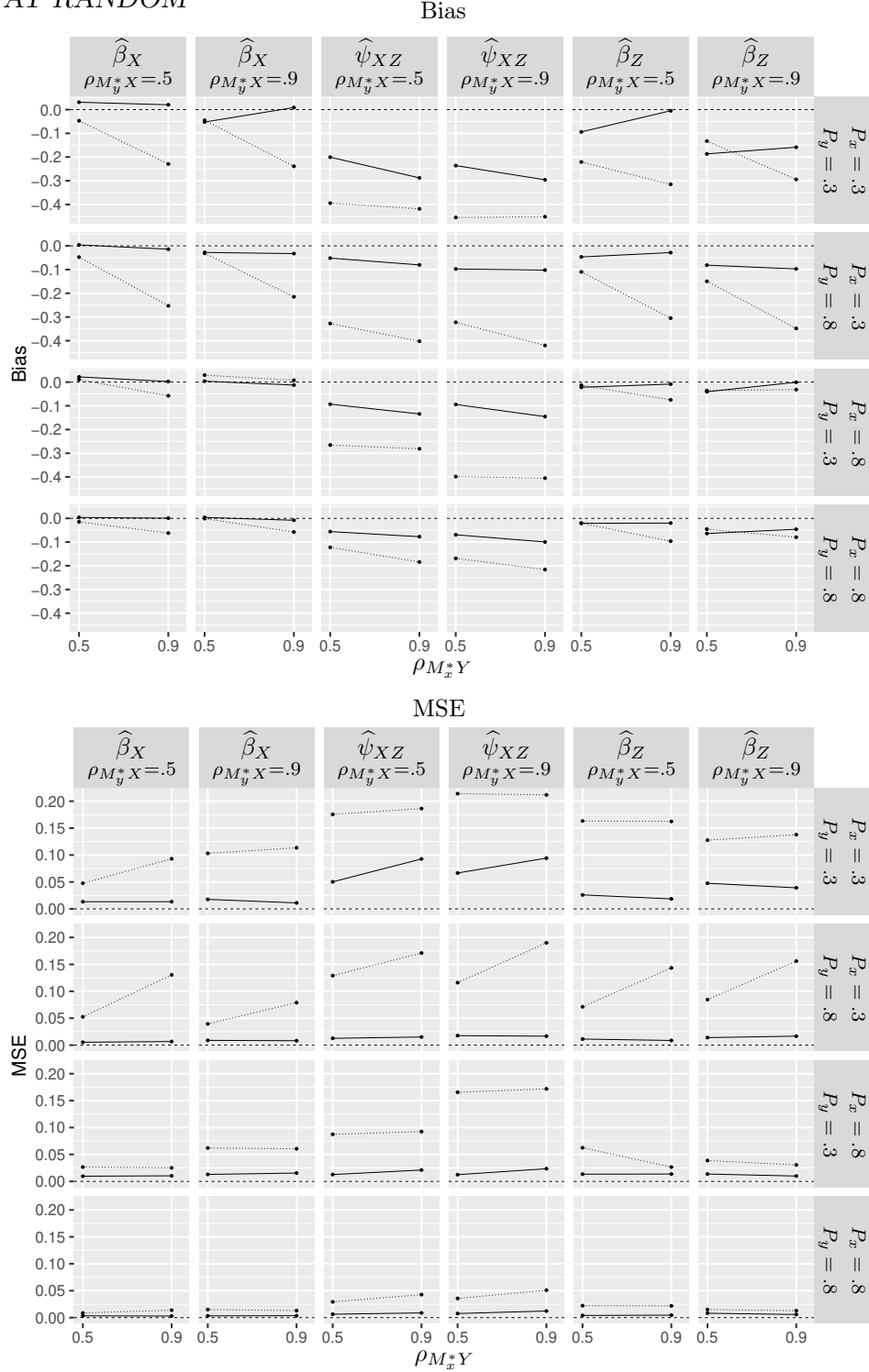


Figure 1.4: Estimated bias and mean squared error (MSE) for SEM (solid lines) and SEM-XY (dotted lines) estimators under MNAR-XY for $N = 500$.

have considerably smaller MSE than SEM in spite of the loss of data (see lower panels of Figures 1.2 and 1.3). The tables in Web Appendix C show that these results persist for $N = 1000$. However, when $N = 200$, the MSE of SEM is comparable to that of SEM-X and SEM-Y.

For MNAR-XY, Figure 1.4 shows that SEM-XY has greater bias and MSE than SEM for all simulation conditions although it is consistent unlike SEM.

1.4 Diagnostic tests for violation of R-MAR

When missingness is believed to be MNAR, the missingness or selection process is sometimes modeled jointly with the process of interest, for example by including the outcome of interest in the selection model (Hausman and Wise, 1979; Diggle and Kenward, 1994), allowing the error term of a probit selection model to be correlated with the error term of the linear regression model of interest (Heckman, 1979), or by including shared random effects in the selection and substantive models (Wu and Carroll, 1988). In such models, MAR can be tested by testing the null hypothesis that the parameters that represent the MAR violations are zero. However, such models and the associated tests rely on unverifiable assumptions (e.g., Molenberghs et al., 2008). Without such assumptions, MAR or everywhere MAR are not testable because they involve statements regarding the missing data.

Little (1988) developed tests for the much stronger missing completely at random (MCAR) assumption, but these tests are of limited use for approaches that require only the weaker MAR assumption. Fortunately, as pointed out by Pothoff et al. (2006), R-MAR (which they called MAR+) can be tested. While R-MAR is a stronger assumption than MAR (or everywhere MAR), so that violation of R-MAR does not imply violation of MAR, it is difficult to imagine realistic scenarios where MAR holds and R-MAR does not. We therefore consider R-MAR tests here.

1.4.1 Testability

Here we consider tests of the null hypothesis that the missingness process is R-MAR against alternative hypotheses expressed as specific violations of R-MAR. Mohan and Pearl (2021), Example 13, summarise the violations of R-MAR that are testable. Based on their work, we can test R-MAR against MNAR-X and MNAR-Y by expressing the alternative hypotheses as

$$\begin{aligned} H_1^{\text{MNAR-X}}: X^* &\not\perp S^y | Z, S^x = 1 \\ H_1^{\text{MNAR-Y}}: Y^* &\not\perp S^x | Z, S^y = 1. \end{aligned}$$

Conditioning on $S^x = 1$ and $S^y = 1$ above ensures that the two statements involve solely observed variables and that both tests are therefore feasible.

1.4.2 Likelihood ratio test (LRT)

We propose to use likelihood-ratio tests based on heteroscedastic regression where the groups defined by the selection indicators have different conditional expectations and variances versus a nested model that assumes equal conditional expectations and variances. Specifically, to test R-MAR against $H_1^{\text{MNAR-X}}$, we use the following model:

$$\begin{aligned} X^* &= \delta_0 + \delta_1 Z + \delta_2 S^y + \delta_3 S^y Z + \epsilon \\ \epsilon|Z, S^y &\sim N(0, (1 - S^y)\sigma_0^2 + S^y\sigma_1^2), \end{aligned}$$

and perform a LRT to test the null hypothesis that $\delta_2 = \delta_3 = 0$ and $\sigma_0^2 = \sigma_1^2$. This null hypothesis corresponds to:

$$\begin{aligned} E(X^*|Z, S^x = 1, S^y) &= E(X^*|Z, S^x = 1) \\ \text{Var}(X^*|Z, S^x = 1, S^y) &= \text{Var}(X^*|Z, S^x = 1). \end{aligned} \tag{1.9}$$

To test R-MAR against $H_1^{\text{MNAR-Y}}$, we use an analogous LRT based on a heteroscedastic regression of Y^* on Z , S^x and $S^x Z$, where the residual variance depends on S^x .

Bojinov et al. (2020) proposed a similar test but considered only conditional expectations as in (1.9). It seems more meaningful to test both the conditional mean and variance because this is a conditional independence test under normality. Moreover, both conditional mean and conditional variance independence are necessary (and sufficient) for estimating covariance matrices and structural equation models with missing data (Mohan et al., 2018). We anticipate that, under multivariate normality, testing the first two moments is more powerful than a generic conditional independence test such as the one described in the next subsection.

1.4.3 Kernel conditional independence (KCI) test

When $X, Y|Z$ is bivariate normal, conditional independence is equivalent to zero partial correlation,

$$X \perp\!\!\!\perp Y|Z \iff \text{Cov}(\epsilon_X, \epsilon_Y) = 0,$$

where $\epsilon_X \equiv X - E(X|Z)$ and $\epsilon_Y \equiv Y - E(Y|Z)$. Relaxing normality, Daudin (1980) provides a characterization of conditional independence as zero partial correlation of certain functions of the random variables. Defining $\tilde{\epsilon}_X \equiv f(X, Z) - E[f(X, Z) | Z]$ and $\tilde{\epsilon}_Y \equiv g(Y, Z) - E[g(Y, Z) | Z]$, conditional independence is characterized as

$$X \perp\!\!\!\perp Y|Z \iff \text{Cov}(\tilde{\epsilon}_X, \tilde{\epsilon}_Y) = 0$$

for all functions f and g that satisfy $E[f(X, Z)^2] < \infty$ and $E[g(Y, Z)^2] < \infty$, respectively. We can think of $\tilde{\epsilon}_X$ as a “residual” function of X, Z that is uncorrelated with any function

of Z and similarly for $\tilde{\epsilon}_Y$. Then conditional independence means that the residual functions $\tilde{\epsilon}_X$ and $\tilde{\epsilon}_Y$ have zero covariance.

Although conceptually simple, this characterization is hard to implement in practice. To address this problem, Zhang et al. (2012) considers a smaller class of functions from a reproducing kernel Hilbert space that makes the problem tractable and propose a kernel conditional independence (KCI) test. For a review of non-parametric conditional independence tests for continuous variables, see Li and Fan (2020).

1.4.4 Power of LRT and KCI tests

Zhang et al. (2012) showed that the KCI test has correct Type I error rates when variables are simulated from various distributions under different non-linear transformations. To assess Type I error rates and power of the two tests in the missing data setting, we perform simulations as described below. We confine our investigation to $H_1^{\text{MNAR-Y}}$ because the results for $H_1^{\text{MNAR-X}}$ should be identical due to symmetry.

1.4.4.1 Simulation design

We simulate data as in Section 1.3 but with X being the only partially observed variable, i.e., $S^y = 1$. The data-generating process is the same as in section 1.3.1.1. The simulation conditions are:

1. Strength of dependence $\rho_{M_x^*Y} \in (0, .1, .2, .3, .4, .5, .6, .7, .8, .9)$.
2. Marginal probability of observing X : $P_x \in (.2, .4, .6, .8)$.
3. Sample size $N \in (100, 200, 500, 1000)$.

For each condition, 100 datasets were simulated and both tests were performed at the 5% level. All analyses were carried out in **R** with the packages **lmvar** (Partners, 2019) for heteroscedastic regression and **CondIndTests** (Heinze-Deml et al., 2019) for the KCI test.

1.4.4.2 Results

Figure 1.5 shows that both tests have type I error rates at the nominal α level. KCI is comparable to LRT in terms of power in each of the simulation conditions, which is impressive given its nonparametric nature. Both tests are fairly powerful (power $\geq .75$) for a moderate effect size ($\rho_{M_x^*Y} \geq .4$) and sample size $N \geq 200$.

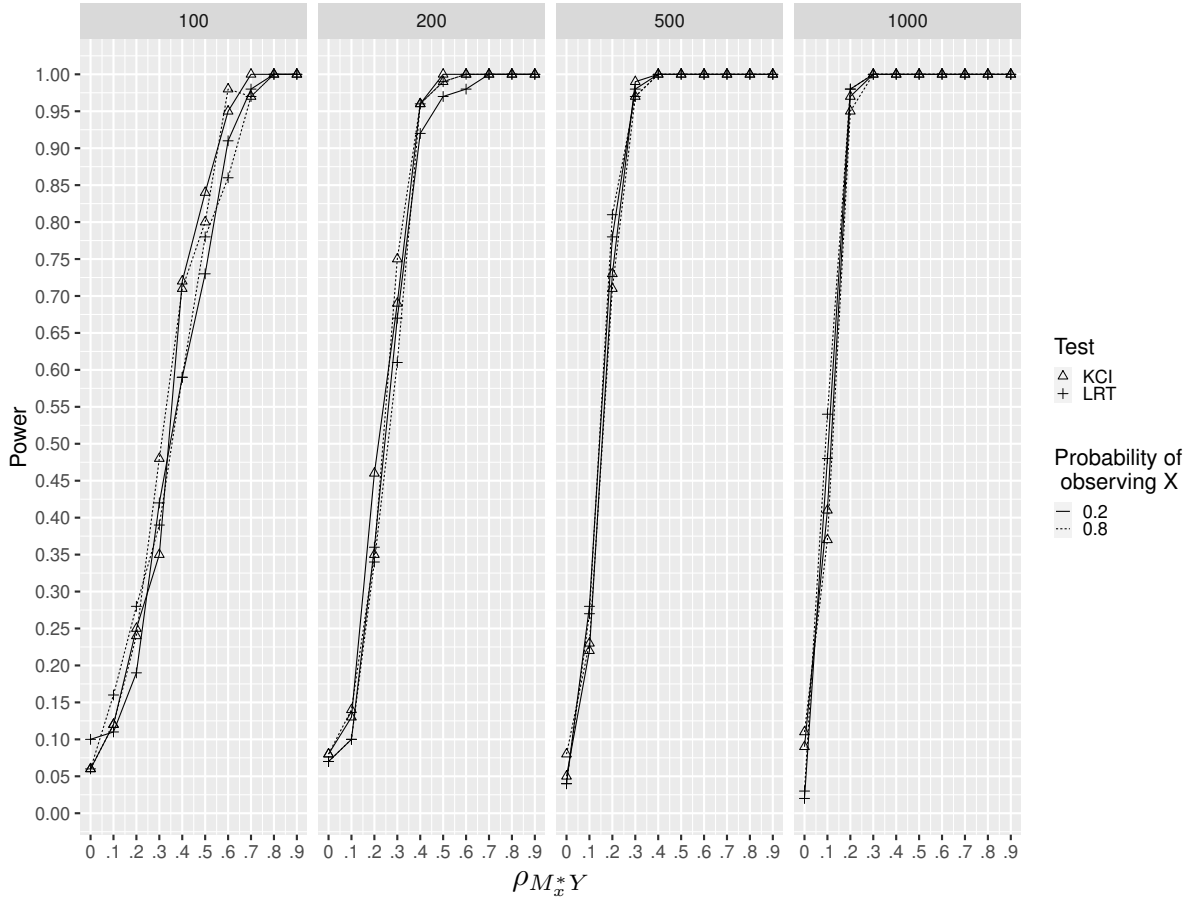


Figure 1.5: Power of test of R-MAR against $H_1^{\text{MNAR-Y}}$.

1.5 Test-based estimator

The simulations in Section 1.3 showed that SEM-X and SEM-Y have lower MSE than SEM if missingness is MNAR-X and MNAR-Y, respectively, especially when the strength of dependence is high and the probabilities of observing X and Y are low. It is fortunate that it is precisely in these conditions that the tests of R-MAR versus MNAR-X and MNAR-Y have high power. We therefore propose conducting these diagnostic tests and then proceeding with SEM-X if R-MAR is rejected in favor of MNAR-X and proceeding with SEM-Y if R-MAR is rejected in favor of MNAR-Y. If neither test is rejected, R-MAR is assumed to hold and SEM is used. If both tests are rejected, the process is assumed to be MNAR-XY. However, the MSE for SEM-XY is considerably greater than for SEM under MNAR-XY, so we will use SEM when both tests are rejected. These decision rules are shown in Table 1.2.

We use a significance level of 0.05 to avoid using the deletion estimators too readily given

Table 1.2: Decision rules for test-based estimator, where H_0 is that R-MAR holds.

		$H_1^{\text{MNAR-Y}}: Y^* \not\perp S^x Z, S^y = 1$	
		Reject H_0	Not Reject H_0
$H_1^{\text{MNAR-X}}: X^* \not\perp S^y Z, S^x = 1$	Reject H_0	MNAR-XY: SEM	MNAR-X: SEM-X
	Not Reject H_0	MNAR-Y: SEM-Y	R-MAR: SEM

the good performance of SEM for small violations of R-MAR. We call the sequence of testing followed by estimation a test-based estimator. To examine the utility of the test-based estimator, we conducted a simulation study to compare the bias and MSE of the test-based estimator with SEM.

1.5.1 Simulation study of finite sample performance

The data generating process is the same as in Section 1.3.1.1, with the following conditions:

1. Strength of dependence $\rho_{M_y^* X} \in \{0, 0.5, 0.9\}$.
2. Strength of dependence $\rho_{M_x^* Y} \in \{0, 0.5, 0.9\}$.
3. Marginal probability of observing X : $P_x \in \{0.3, 0.9\}$.
4. Marginal probability of observing Y : $P_y \in \{0.3, 0.9\}$.
5. Sample size $N \in \{200, 500, 1000\}$.

For each combination of simulation conditions, we replicated the simulation 100 times and applied two tests (LRT and KCI, at the 5% level) and two estimators (SEM and test-based), and obtained the estimated bias and MSE for $\hat{\beta}_X$, $\hat{\beta}_Z$ and $\hat{\psi}_{XZ}$.

Here we report the performance of the LRT only because the KCI test performed very similarly. Figures 1.2 and 1.3 show the estimated bias and MSE for the test-based estimator (dashed lines) compared with SEM (solid lines) and the consistent estimator SEM-X for MNAR-X and SEM-Y for MNAR-Y (dotted lines). In situations where the consistent estimator performs appreciably better than SEM, the test-based estimator is not much worse than the consistent estimator. Specifically, when the proportions observed X and Y are low (first row) and the R-MAR violation is strong (large $\rho_{M_y^* x}$ for MNAR-X or large $\rho_{M_x^* y}$ for MNAR-Y), SEM performs much worse than the consistent estimator, whereas the test-based estimator performs almost as well as the consistent estimator.

Tables 1.3 and 1.4 show detailed results aggregated across the nine combinations of the three values for $\rho_{M_y^* x}$ with the three values for $\rho_{M_x^* y}$, representing R-MAR, MNAR-X, MNAR-Y, and MNAR-XY. Across these missingness mechanisms, the test-based estimator

Table 1.3: Bias ($\times 1000$) for SEM and test-based estimator across conditions.

		P_x											
		0.3						0.8					
		P_y						P_y					
		0.3			0.8			0.3			0.8		
Method	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	200	-40	-51	-173	-20	-44	-56	-9	-19	-82	-5	-25	-47
	500	-43	-53	-175	-13	-37	-60	-3	-18	-76	-3	-23	-49
	1000	-55	-50	-177	-17	-38	-57	-15	-15	-79	-5	-22	-49
Test	200	-4	-63	-120	-6	-36	-44	-4	-16	-67	-2	-21	-36
	500	-0	-54	-126	-7	-34	-45	-4	-12	-60	-0	-18	-37
	1000	-12	-48	-128	-11	-33	-40	-7	-8	-62	-2	-16	-38

Table 1.4: MSE ($\times 1000$) for SEM and test-based estimator across conditions.

		P_x											
		0.3						0.8					
		P_y						P_y					
		0.3			0.8			0.3			0.8		
Method	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	200	52	73	69	19	30	25	33	31	19	10	13	11
	500	24	42	53	8	13	12	12	14	12	4	5	6
	1000	18	36	50	4	8	8	7	7	11	2	3	5
Test	200	47	64	61	20	33	25	33	32	20	10	13	11
	500	16	28	41	8	13	12	12	14	11	4	5	6
	1000	8	19	37	4	8	7	6	6	9	2	3	4

had the smallest bias and MSE for most simulation conditions and comparable performance otherwise. Specifically, the test-based estimator is almost uniformly better (i.e., smaller bias and MSE) than SEM when $N = 500$ and is uniformly better when $N = 1000$ for the conditions considered. The bias reduction for the test-based estimators is evident across the simulation conditions and can be as large as 99% (the highlighted cells in Table 1.3). The benefit of using the test-based estimator regarding MSE is large when the missingness of both X and Y is severe (i.e., $P_x = P_y = .3$, Table 1.4). Similar to the population results, $\hat{\beta}_Z$ and $\hat{\psi}_{XZ}$ have larger bias and MSE across all simulation conditions than $\hat{\beta}_X$.

The bias and MSE for each missingness mechanism separately are summarised in Web Appendix D (Tables 1.11 to 1.16). Again, the test-based estimator tends to outperform SEM when the missingness mechanisms are MNAR-X and MNAR-Y. The greater the proportion of missing data, either in X or Y , the greater the improvement in the bias and MSE, especially when the strength of dependence $\rho_{M_x^*Y}$ and $\rho_{M_y^*X}$ is high. Reassuringly, Tables 15 and 16 show that the test-based estimator performs nearly as well as SEM under R-MAR, with the larger discrepancies occurring for $N = 200$.

1.6 Concluding remarks

We have demonstrated that the data-deletion approaches SEM-X and SEM-Y are approximately unbiased and have smaller MSE than the conventional SEM estimator for MNAR-X and MNAR-Y, respectively. We proposed tests to diagnose whether the missingness process is MNAR-X, MNAR-Y or MNAR-XY. Surprisingly, the nonparametric test had similar power to the parametric counterpart and both tests performed well for moderate departures from R-MAR. We also proposed a test-based estimator in which the choice between SEM, SEM-X and SEM-Y is determined by the diagnostic tests. This estimator outperformed the SEM estimator under MNAR-X/MNAR-Y and performed nearly as well under R-MAR.

The model considered was a simple linear SEM without latent variables, but the estimators generalize to other multivariate models such as SEMs with latent variables and/or categorical response variables.

Our missing data assumptions for consistent estimation of target quantities were confined to properties of the missingness graphs. However, in some situations, the features of the data can also be utilized to address harder questions, such as a binary outcome causing its own missingness when the explanatory variable is binary (Mohan, 2018). See also Skrondal and Rabe-Hesketh (2014) on ways to protect target parameters against inconsistency due to MNAR in mixed models for binary responses.

Acknowledgements

This chapter comprises work done by the author of this dissertation, the chair of this dissertation committee, and Anders Skrondal. Permission was sought from all authors as well as from the Berkeley Graduate Division for the work to be included as a chapter in this dissertation.

This work was partially supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 26270.

1.7 Appendix A: Derivation of IPW estimators

For the joint model, we now show the derivation of the IPW estimator under MNAR-XY for Section 1.2.3:

$$\begin{aligned} E \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) \right] &= E_{X,Z,Y} \left[E \left(\frac{1}{\pi} S^x S^y m(X, Z, Y) \mid X, Z, Y \right) \right] \\ &= E_{X,Z,Y} \left[\frac{1}{\pi} E(S^x S^y \mid X, Z, Y) m(X, Z, Y) \right] \\ &= E_{X,Z,Y} \left[\frac{1}{\pi} E(S^x \mid X, Z, Y) E(S^y \mid X, Z, Y) m(X, Z, Y) \right] \quad (1.10) \end{aligned}$$

$$\begin{aligned} &= E_{X,Z,Y} \left[\frac{1}{\pi} E(S^x \mid Z, Y) E(S^y \mid X, Z) m(X, Z, Y) \right] \quad (1.11) \\ &= E_{X,Z,Y} \left[\frac{1}{\pi} P(S^x = 1 \mid Z, Y) P(S^y = 1 \mid X, Z) m(X, Z, Y) \right]. \end{aligned}$$

Here (1.10) follows because $S^x \perp\!\!\!\perp S^y \mid X, Z, Y$ and (1.11) follows because $S^x \perp\!\!\!\perp X \mid Y, Z$ and $S^y \perp\!\!\!\perp Y \mid X, Z$. We can estimate $P(S^x = 1 \mid Z, Y) = P(S^x = 1 \mid Z, Y, S^y = 1)$ in the complete-case sample. Similarly, because $S^x \perp\!\!\!\perp S^y \mid X, Z$, we can estimate $P(S^y = 1 \mid X, Z) = P(S^y = 1 \mid X, Z, S^x = 1)$.

For the conditional model, the derivation of the IPW estimator described in Section 1.2.3 is given below.

$$\begin{aligned} E \left[\frac{1}{\pi} S^x S^y m(X, Z, Y) \mid X, Z \right] &= E_{Y \mid X, Z} \left[E \left(\frac{1}{\pi} S^x S^y m(X, Z, Y) \mid X, Z, Y \right) \right] \\ &= E_{Y \mid X, Z} \left[\frac{1}{\pi} E(S^x \mid Z, Y) E(S^y \mid X, Z) m(X, Z, Y) \right] \quad (1.12) \\ &= E(S^y \mid X, Z) E_{Y \mid X, Z} \left[\frac{1}{\pi} E(S^x \mid Z, Y) m(X, Z, Y) \right] \\ &= P(S^y = 1 \mid X, Z) E_{Y \mid X, Z} \left[\frac{1}{\pi} P(S^x = 1 \mid Z, Y) m(X, Z, Y) \right]. \end{aligned}$$

Here (1.12) follows from steps analogous to those leading to (1.11).

1.8 Appendix B: $\rho_{M_x^*Y|X,Z}$, $\rho_{M_y^*X|Z}$, P_x , and P_y in terms of model parameters

For the data-generating model described in Section 1.3.1.1, we derive expressions for the strengths of dependence, $\rho_{M_x^*Y|X,Z}$, $\rho_{M_y^*X|Z}$, and probabilities, P_x and P_y , as functions of the model parameters.

$$\begin{aligned} \text{Var}(M_x^*|X, Z) &= \alpha_Y^2 \sigma^2 + 1 \\ \text{Var}(Y|X, Z) &= \sigma^2 \\ \text{Cov}(M_x^*, Y|X, Z) &= \alpha_Y \sigma^2 \\ \rho_{M_x^*Y|X,Z} &= \frac{\alpha_Y \sigma^2}{\sqrt{\alpha_Y^2 \sigma^2 + 1} \sqrt{\sigma^2}} \end{aligned} \tag{1.13}$$

$$\begin{aligned} \text{Var}(M_y^*|Z) &= \gamma_X^2 \psi_{XX} + 1 \\ \text{Var}(X|Z) &= (1 - \frac{\psi_{XZ}^2}{\psi_{XX}\psi_{ZZ}}) \psi_{XX} \\ \text{Cov}(M_y^*, X|Z) &= \gamma_X \psi_{XX} \\ \rho_{M_y^*X|Z} &= \frac{\gamma_X \psi_{XX}}{\sqrt{\gamma_X^2 \psi_{XX} + 1} \sqrt{(1 - \frac{\psi_{XZ}^2}{\psi_{XX}\psi_{ZZ}}) \psi_{XX}}} \end{aligned} \tag{1.14}$$

$$\begin{aligned} \text{E}(M_x^*) &= \alpha_0 + \alpha_Y \beta_0 \\ \text{Var}(M_x^*) &= 1 + \alpha_Z^2 \psi_{ZZ} + \alpha_Y^2 (\beta_X^2 \psi_{XX} + \beta_Z^2 \psi_{ZZ} + 2\beta_X \beta_Z \psi_{XZ} + \sigma^2) + 2\alpha_Z \alpha_Y (\beta_X \psi_{XZ} + \beta_Z \psi_{ZZ}) \\ P_x &= P(S^x = 1) = P(M_x^* \leq 0) \\ &= \Phi\left(\frac{-(\alpha_0 + \alpha_Y \beta_0)}{\sqrt{1 + \alpha_Z^2 \psi_{ZZ} + \alpha_Y^2 (\beta_X^2 \psi_{XX} + \beta_Z^2 \psi_{ZZ} + 2\beta_X \beta_Z \psi_{XZ} + \sigma^2) + 2\alpha_Z \alpha_Y (\beta_X \psi_{XZ} + \beta_Z \psi_{ZZ})}}\right) \end{aligned} \tag{1.15}$$

$$\begin{aligned} \text{E}(M_y^*) &= \gamma_0 + \gamma_Y \beta_0 \\ \text{Var}(M_y^*) &= \gamma_X^2 \psi_{XX} + \gamma_Z^2 \psi_{ZZ} + 2\gamma_Z \gamma_X \psi_{XZ} + 1 \\ P_y &= P(S^y = 1) = P(M_y^* \leq 0) = \Phi\left(\frac{-(\gamma_0 + \gamma_Y \beta_0)}{\sqrt{\gamma_X^2 \psi_{XX} + \gamma_Z^2 \psi_{ZZ} + 2\gamma_Z \gamma_X \psi_{XZ} + 1}}\right) \end{aligned} \tag{1.16}$$

All parameters, except α_0 , γ_0 , α_Y , and γ_X , are fixed. To achieve desired strengths of dependence, Equations (1.13) and (1.14) are solved for α_Y and γ_X , respectively. To achieve

desired marginal probabilities of observing X and Y , Equations (1.15), (1.16) are then solved for α_0 and γ_0 , respectively.

1.9 Appendix C: Performance of SEM, SEM-X, SEM-Y, and SEM-XY

Table 1.5: Bias ($\times 1000$) of SEM and SEM-X under MNAR-X.

			P_x											
			0.3						0.8					
			P_y			P_y			P_y			P_y		
			0.3			0.8			0.3			0.8		
Methods	$\rho_{M_y^*X}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	-30	-188	-27	-13	-31	-68	-15	-8	-5	-19	-18	-10
		500	-39	-121	-23	-3	-68	-26	-14	-25	-1	-1	-34	-8
		1000	-37	-144	-33	-8	-55	-43	-28	-24	0	-13	-22	-14
	0.9	200	-268	-239	-26	-39	-93	-66	-78	-48	-4	-22	-69	-15
		500	-258	-267	-49	-51	-101	-64	-55	-57	-11	-28	-47	-31
		1000	-282	-246	-32	-47	-96	-78	-74	-76	-6	-20	-49	-24
SEM-X	0.5	200	-18	-53	-1	-21	-24	-39	-9	-35	-7	-15	-15	-7
		500	-9	-12	-3	-6	-4	-17	-6	-4	-3	-1	-9	-4
		1000	-12	-4	-6	-0	-9	-0	-20	-4	-1	-8	-5	-0
	0.9	200	-6	-34	-6	-4	-18	-10	-19	-5	0	-3	-23	-13
		500	-5	-33	-14	-3	0	-8	-10	-0	-6	-3	-6	0
		1000	-21	0	-4	-3	-13	-3	-8	-13	0	-2	-4	-5

Table 1.6: MSE ($\times 1000$) of SEM and SEM-X under MNAR-X.

			P_x											
			0.3						0.8					
			P_y						P_y					
			0.3			0.8			0.3			0.8		
Methods	$\rho_{M_y^*X}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	32	83	24	26	23	19	31	36	10	9	14	5
		500	18	27	7	8	15	8	12	12	3	4	6	3
		1000	6	29	5	3	6	5	6	6	1	2	2	1
	0.9	200	150	110	27	16	24	18	46	28	10	12	14	8
		500	88	87	11	9	17	11	21	17	2	5	6	3
		1000	92	70	4	6	12	8	15	10	1	2	4	2
SEM-X	0.5	200	48	91	25	31	36	18	32	40	10	9	15	5
		500	20	23	8	11	18	10	12	12	3	4	5	3
		1000	7	14	3	5	6	4	6	6	1	2	2	1
	0.9	200	103	99	25	21	31	23	38	29	10	13	11	8
		500	28	28	10	11	14	9	16	15	2	5	5	2
		1000	15	16	4	6	8	3	9	5	1	2	2	1

Table 1.7: Bias ($\times 1000$) of SEM and SEM-Y under MNAR-Y.

		P_x												
		0.3						0.8						
		P_y						P_y						
		0.3			0.8			0.3			0.8			
Methods	$\rho_{M_x^*Y}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	-19	-108	-176	-3	-27	-33	-3	-36	-99	-5	-6	-30
		500	-8	-111	-166	-12	-35	-45	-7	-14	-82	-4	-7	-32
		1000	-47	-140	-173	-3	-19	-25	-3	-15	-94	-2	-10	-31
	0.9	200	-92	-275	-292	-16	-44	-47	-27	-9	-137	-11	-18	-62
		500	-96	-251	-305	-18	-68	-65	-1	-24	-134	-5	-19	-63
		1000	-107	-256	-299	-7	-45	-46	-0	-17	-129	-4	-22	-65
SEM-Y	0.5	200	-49	-15	-24	-4	-2	-3	-5	-47	-4	-3	-3	-6
		500	-23	-17	-1	-17	-8	-17	-8	-22	-13	-5	-2	-3
		1000	-16	-11	-2	-2	-8	-2	-2	-7	-8	-3	0	-7
	0.9	200	-14	-12	-1	-4	0	-3	-33	-27	-3	-8	-2	-12
		500	-13	-10	-27	-1	-14	-16	-0	-10	-2	-2	-0	-5
		1000	-3	-6	-10	-6	-10	-7	-3	-0	-21	-1	-4	-8

Table 1.8: MSE ($\times 1000$) of SEM and SEM-Y under MNAR-Y.

		P_x												
		0.3						0.8						
		P_y						P_y						
		0.3			0.8			0.3			0.8			
Methods	$\rho_{M_x^*Y}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	32	60	61	14	30	24	31	33	18	10	13	9
		500	13	30	38	7	10	7	8	16	10	4	5	4
		1000	8	28	35	3	5	5	5	6	10	1	2	2
	0.9	200	64	127	112	19	36	32	29	35	27	8	13	10
		500	29	86	103	8	18	14	6	16	21	4	5	7
		1000	22	80	94	3	8	6	4	7	18	1	2	5
SEM-Y	0.5	200	35	53	55	13	30	22	31	35	27	10	13	10
		500	13	23	21	7	9	6	8	16	9	3	5	4
		1000	5	10	7	2	4	4	5	6	4	1	2	1
	0.9	200	42	60	51	18	35	32	28	34	31	8	13	9
		500	11	29	19	7	12	9	6	15	7	3	4	3
		1000	6	11	8	2	6	4	3	6	4	1	2	2

Table 1.9: Bias ($\times 1000$) of SEM and SEM-XY under MNAR-XY.

Methods	$\rho_{M_y^*X}$	$\rho_{M_x^*Y}$	N	P_x											
				0.3						0.8					
				P_y			P_y			P_y			P_y		
				0.3	0.3	0.3	0.8	0.8	0.8	0.3	0.3	0.3	0.8	0.8	0.8
				$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	0.5	200	-30	-49	-195	-16	-50	-70	-4	-39	-101	-9	-22	-57
			500	-31	-93	-201	-4	-47	-52	-21	-21	-93	-3	-21	-56
			1000	-15	-70	-207	-6	-44	-57	-9	-25	-93	-3	-34	-53
		0.9	200	-0	-10	-281	-13	-49	-62	-4	-15	-141	0	-15	-74
			500	-20	-4	-288	-14	-28	-80	-2	-8	-134	-0	-20	-77
			1000	-1	-6	-304	-0	-33	-65	-4	-15	-143	-3	-17	-73
		0.9	200	-42	-177	-266	-39	-107	-87	-23	-21	-92	-11	-63	-72
			500	-52	-186	-236	-27	-81	-97	-4	-40	-94	-3	-64	-69
			1000	-38	-201	-238	-39	-92	-102	-3	-30	-95	-7	-54	-81
	0.9	0.9	200	-19	-165	-291	-35	-114	-89	-3	-1	-152	-9	-65	-104
			500	-8	-158	-296	-32	-97	-102	-12	0	-145	-7	-46	-99
			1000	-13	-177	-306	-43	-90	-88	-28	-8	-144	-4	-56	-96
		0.5	200	-21	-140	-382	-75	-185	-344	-49	-56	-315	-14	-43	-133
			500	-46	-220	-394	-47	-110	-327	-10	-13	-265	-14	-20	-122
			1000	-81	-164	-400	-53	-128	-288	-1	-51	-253	-6	-12	-81
SEM-XY	0.5	0.9	200	-238	-352	-435	-240	-341	-403	-61	-106	-303	-69	-84	-198
			500	-229	-315	-418	-252	-305	-402	-57	-74	-281	-62	-96	-184
			1000	-236	-337	-424	-243	-339	-393	-41	-75	-270	-57	-90	-174
		0.5	200	-95	-96	-454	-62	-148	-335	-25	-24	-415	-23	-24	-182
			500	-45	-132	-455	-31	-149	-322	-29	-35	-398	-1	-46	-169
			1000	-51	-111	-438	-59	-122	-316	-20	-17	-394	-11	-29	-139
		0.9	200	-202	-311	-462	-224	-388	-405	-18	-26	-405	-53	-88	-232
			500	-238	-294	-452	-214	-348	-420	-7	-31	-405	-57	-80	-216
			1000	-273	-265	-456	-178	-338	-387	-63	-31	-378	-55	-87	-204

Table 1.10: MSE ($\times 1000$) of SEM and SEM-XY under MNAR-XY.

			P_x												
			0.3						0.8						
			P_y						P_y						
			0.3			0.8			0.3			0.8			
Methods	$\rho_{M_y^*X}$	$\rho_{M_x^*Y}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	0.5	200	38	38	63	23	25	23	21	31	18	11	12	10
			500	13	26	50	5	11	12	9	13	13	3	4	6
			1000	6	13	47	3	7	8	3	5	11	1	3	4
		0.9	200	46	55	99	14	32	29	30	26	27	9	10	14
			500	13	18	93	6	8	15	10	13	21	3	4	8
			1000	7	7	97	3	5	8	4	7	22	1	2	6
	0.9	0.5	200	42	72	95	18	33	28	36	37	18	12	14	12
			500	17	47	67	8	14	17	12	13	12	3	8	7
			1000	6	46	61	4	11	13	8	6	11	2	4	8
		0.9	200	26	61	107	19	42	31	44	23	31	8	13	19
			500	11	39	94	8	16	16	15	9	24	3	5	12
			1000	6	37	97	5	12	12	8	5	22	1	4	10
SEM-XY	0.5	0.5	200	79	179	159	65	133	135	46	64	119	20	31	35
			500	47	163	175	52	71	129	26	62	87	8	22	29
			1000	55	115	172	28	77	96	17	36	75	12	17	24
		0.9	200	129	208	205	103	172	174	42	77	109	20	22	47
			500	93	162	186	130	143	171	25	26	92	13	21	42
			1000	82	146	186	84	142	164	15	30	82	9	13	38
	0.9	0.5	200	161	109	222	62	120	128	105	72	188	22	27	47
			500	103	127	214	39	84	116	62	38	165	15	15	35
			1000	62	64	199	32	64	109	49	30	165	11	7	30
		0.9	200	108	169	224	89	213	173	96	53	174	17	22	64
			500	113	137	212	78	156	190	60	30	172	13	13	50
			1000	131	105	214	53	137	156	70	29	149	9	12	46

1.10 Appendix D: Performance of test-based estimator

Table 1.11: Bias ($\times 1000$) of SEM and test-based estimator (based on LRT) under MNAR-X.

		P_x												
		0.3						0.8						
		P_y						P_y						
		0.3			0.8			0.3			0.8			
Methods	$\rho_{M_y^*X}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	-30	-188	-27	-13	-31	-68	-14	-8	-5	-19	-18	-10
		500	-39	-121	-23	-3	-68	-26	-14	-25	-1	-1	-34	-8
		1000	-37	-144	-33	-8	-55	-43	-28	-24	0	-13	-21	-14
	0.9	200	-268	-239	-26	-39	-93	-65	-78	-47	-4	-21	-69	-14
		500	-258	-267	-49	-51	-101	-64	-55	-57	-11	-28	-47	-31
		1000	-282	-246	-32	-47	-96	-78	-74	-76	-6	-20	-49	-24
Test	0.5	200	-8	-126	-19	-14	-11	-59	-9	-25	-5	-17	-6	0
		500	-4	-3	-0	-4	-49	-17	-9	-4	-3	-1	-11	-3
		1000	-8	-23	-10	-1	-10	-13	-20	-1	-1	-8	-3	-1
	0.9	200	-34	-55	0	-3	-9	-7	-39	-13	-1	-1	-29	-9
		500	-38	-67	-20	-21	-26	-12	0	-9	-7	-8	-4	-6
		1000	-57	-36	-2	-15	-22	-27	-15	-22	-1	-1	-7	-1

Table 1.12: MSE ($\times 1000$) of SEM and test-based estimator (based on LRT) under MNAR-X.

			P_x											
			0.3						0.8					
			P_y						P_y					
			0.3			0.8			0.3			0.8		
Methods	$\rho_{M_y^*X}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	32	83	24	26	23	19	31	36	10	9	14	5
		500	18	27	7	8	15	8	12	12	3	4	6	3
		1000	6	29	5	3	6	5	6	6	1	2	2	1
	0.9	200	150	110	27	16	24	18	46	28	10	12	14	8
		500	88	87	11	9	17	11	21	17	2	5	6	3
		1000	92	70	4	6	12	8	15	10	1	2	4	2
Test	0.5	200	36	84	24	29	28	18	31	37	10	9	15	6
		500	19	21	8	9	14	9	12	12	3	4	5	3
		1000	7	15	4	4	6	3	6	6	1	2	2	1
	0.9	200	117	87	26	18	27	18	39	28	10	13	10	8
		500	32	31	10	11	13	8	17	15	2	5	4	3
		1000	26	22	4	5	7	5	10	6	1	2	2	1

Table 1.13: Bias ($\times 1000$) of SEM and test-based estimator (based on LRT) under MNAR-Y.

		P_x													
		0.3						0.8							
		P_y						P_y							
		0.3			0.8			0.3			0.8				
Methods	$\rho_{M_x^*Y}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	
SEM	0.5	200	-18	-108	-176	-3	-27	-33	-3	-36	-99	-4	-6	-30	
		500	-8	-111	-166	-12	-35	-45	-7	-14	-82	-4	-7	-32	
		1000	-47	-140	-174	-3	-19	-25	-3	-15	-95	-2	-10	-31	
	0.9	200	-91	-275	-293	-16	-44	-47	-27	-9	-138	-11	-18	-62	
		500	-96	-251	-305	-18	-68	-65	-1	-24	-135	-5	-19	-63	
		1000	-107	-256	-299	-7	-45	-46	-0	-17	-129	-4	-22	-65	
	Test	0.5	200	-38	-23	-70	-2	-8	-15	-4	-43	-49	-4	-1	-3
			500	-17	-5	-34	-16	-12	-21	-8	-20	-41	-5	0	-3
			1000	-18	-21	-16	-0	-1	-4	-2	-8	-29	-3	-2	-1
0.9		200	-6	-36	-28	-11	-22	-25	-31	-19	-57	-8	-4	-19	
		500	-3	-42	-63	-5	-27	-27	0	-13	-44	-3	-5	-19	
		1000	-24	-61	-75	-4	-2	0	-2	-8	-49	-2	-11	-30	

Table 1.14: MSE ($\times 1000$) of SEM and test-based estimator (based on LRT) under MNAR-Y.

		P_x												
		0.3						0.8						
		P_y						P_y						
		0.3			0.8			0.3			0.8			
Methods	$\rho_{M_x^*Y}$	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	0.5	200	32	60	61	14	30	24	31	33	18	10	13	9
		500	13	30	38	7	10	7	8	16	10	4	5	4
		1000	8	28	35	3	5	5	5	6	10	1	2	2
	0.9	200	64	127	112	19	36	32	29	35	27	8	13	10
		500	29	86	103	8	18	14	6	16	21	4	5	7
		1000	22	80	94	3	8	6	4	7	18	1	2	5
Test	0.5	200	32	57	52	13	30	24	32	33	23	10	13	10
		500	13	23	19	7	10	6	8	16	9	4	5	3
		1000	5	12	9	3	4	4	5	6	5	1	2	1
	0.9	200	44	64	52	18	36	32	27	34	22	8	13	9
		500	15	27	25	8	13	9	6	15	10	3	4	4
		1000	11	27	27	2	6	4	3	7	8	1	2	3

Table 1.15: Bias ($\times 1000$) of SEM and test-based estimator (based on LRT) under R-MAR.

		P_x											
		0.3						0.8					
		P_y						P_y					
		0.3			0.8			0.3			0.8		
Methods	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	200	-5	-12	-4	29	-22	-15	-12	-12	-14	-5	-0	-0
	500	-7	-10	-7	17	-8	-5	-18	-16	-7	-4	-2	-4
	1000	-9	-2	-4	0	-4	-3	-8	-9	-1	-2	-4	0
Test	200	-10	-13	-8	22	-10	-21	-13	-13	-7	-5	-0	-2
	500	-11	-16	-4	17	-7	-3	-18	-17	-10	-4	-2	-4
	1000	-9	-3	-5	0	-3	-3	-8	-9	-3	-2	-3	0

Table 1.16: MSE ($\times 1000$) of SEM and test-based estimator (based on LRT) under R-MAR.

		P_x											
		0.3						0.8					
		P_y						P_y					
		0.3			0.8			0.3			0.8		
Methods	N	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$	$\hat{\beta}_X$	$\hat{\beta}_Z$	$\hat{\psi}_{XZ}$
SEM	200	36	51	27	21	24	15	28	30	10	9	12	6
	500	13	19	7	6	8	6	12	12	3	3	3	3
	1000	7	11	4	3	4	3	3	6	1	2	2	1
Test	200	40	68	28	27	29	18	28	30	17	9	12	7
	500	15	23	8	6	9	6	12	12	4	2	3	3
	1000	7	11	4	3	4	3	3	6	1	2	2	1

1.11 References

- Bojinov, I., Pillai, N., and Rubin, D. (2020). Diagnosing missing always at random in multivariate data. *Biometrika*, 107(1):246–253.
- Daudin, J. (1980). Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590.
- Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society, Series C*, 43(1):49–73.
- Geiger, D., Verma, T., and Pearl, J. (1990). d-separation: From theorems to algorithms. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence*, volume 10 of *Machine Intelligence and Pattern Recognition*, pages 139–148. North-Holland.
- Hausman, J. A. and Wise, D. A. (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica*, 47(2):455–473.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.
- Heinze-Deml, C., Peters, J., and Munk, A. M. S. (2019). Condindtests: Nonlinear conditional independence tests. R package version 0.1.5.
- Li, C. and Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(3):1–11.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.
- Little, R. J. (1992). Regression with missing X’s: A review. *Journal of the American Statistical Association*, 87(420):1227–1237.
- Little, R. J. and Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society, Series C*, 60(4):591–605.
- Lumley, T. (2019). survey: Analysis of complex survey samples, R package version 4.1.1.
- Mohan, K. (2018). On handling self-masking and other hard missing data problems. In *AAAI symposium*.
- Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1–16.

- Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in neural information processing systems*, volume 26, pages 1277–1285. Lake Tahoe.
- Mohan, K., Thoemmes, F., and Pearl, J. (2018). Estimation with incomplete data: The linear case. In *Twenty-seventh International Joint Conference on Artificial Intelligence*, pages 5082–5088.
- Molenberghs, G., Beunckens, C., Sotito, C., and Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, 70(2):371–388.
- Oberski, D. (2014). lavaan. survey: An R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, 57(1):1–27.
- Partners, P. (2019). lmvar: Linear regression with non-constant variances, R package version 1.5.2.
- Pothoff, R. F., Tudor, G. E., Pieper, K. S., and Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, 15:213–234.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rabe-Hesketh, S. and Skrondal, A. (2015). Ignoring non-ignorable missingness. Presidential Address at the International Meeting of the Psychometric Society, Beijing, China.
- Rabe-Hesketh, S. and Skrondal, A. (2022). Ignoring non-ignorable missingness. *Psychometrika*, in press.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. version 0.5–12 (beta). *Journal of Statistical Software*, 48(2):1–36.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Seaman, S. R., Galati, J., Jackson, J., and Carlin, J. (2013). What is meant by “missing at random”? *Statistical Science*, 28(2):257–268.
- Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.
- Skrondal, A. and Rabe-Hesketh, S. (2014). Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika*, 101(1):175–188.

- van der Wal, W. M. and Geskus, R. B. (2011). IPW: an R package for inverse probability weighting. *Journal of Statistical Software*, 43(13):1–23.
- Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of change in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44(1):175–188.
- Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence (UAI), Barcelona, Spain*.

Chapter 2

Approximate Resampling Standard Errors for Bayesian Quantile Regression

2.1 Introduction

Resampling techniques (Efron, 1982) have been demonstrated useful in applied statistics on many fronts. Examples include general procedures for estimating standard errors without deriving analytic solutions based on asymptotic approximations, bias correction in parameter estimation, hypothesis testing and bootstrap confidence intervals, and quantifying out-of-sample prediction errors (e.g., cross-validation).

The general applicability and flexibility comes at a high computational cost because the parameters must be re-estimated for each resampled dataset. Some previous studies (e.g., Efron, 1990; Hanley and MacGibbon, 2006) have developed ways to reduce the number of bootstrap replications needed – see Efron and Tibshirani (1994, Chapter 23) for a comprehensive review of such methods. More recently, Kleiner et al. (2014) combine the idea of subsampling and parallel computing so that many small-scale bootstraps can be computed in a parallel manner to lower the overall computational overhead.

In the machine learning literature, there have been a series of recent studies on how to approximate leave-one-out cross validation in an efficient and scalable fashion so that it becomes feasible to evaluate modern large-scale complex machine learning models (Koh and Liang, 2017; Giordano et al., 2019; Alaa and Van Der Schaar, 2020). Specifically, in the context of identifying influential training data points, Koh and Liang (2017) introduced the notion of influence functions to the machine learning community. The concept originated from the statistical literature on influence statistics (Cook, 1977; Pregibon, 1981; Cook and Weisberg, 1982) and robust statistics (e.g., Huber, 2004; Hampel et al., 2011).

Similarly to the “one-step Newton-Raphson” method for obtaining approximate leave-one-out maximum likelihood estimates (MLEs) based on MLEs for the full sample (e.g.,

Pregibon, 1981; Verweij and Van Houwelingen, 1993), the influence function can be used to derive a linear approximation for resampled-data estimates for general loss functions (e.g., Efron, 1982; Cain and Lange, 1984; Giordano et al., 2019). This approximation, called the Swiss Army Infinitesimal Jackknife by Giordano et al. (2019), replaces the computationally intensive step of re-estimating the parameters from scratch for each resampled dataset. Giordano et al. (2019) provide both practical implications and general theoretical guarantees for the infinitesimal Jackknife (IJ for short), as well as utilizing modern automatic differentiation techniques (e.g., Baydin et al., 2018) to implement the IJ efficiently for cross-validation of large-scale machine learning models. Many applications have been proposed based on this theoretical results. In related work, Broderick et al. (2020) applied the IJ approximation to sensitivity analysis and developed a computationally efficient method for identifying a subsample of the data that causes instability.

In a Bayesian setting, Efron (2015), Giordano et al. (2018), and Giordano (2020) have derived a linear approximation for the posterior mean parameters in resampled data that resembles the IJ and shows that the required gradients can be estimated straightforwardly as the covariance among the Markov chain Monte Carlo (MCMC) draws of the parameters and corresponding log-likelihood contributions. Inspired by this line of work as well as Lee et al. (2017) and Giordano et al. (2019), we develop IJ standard errors for both unclustered and clustered Bayesian quantile regression. Bayesian quantile regression is often implemented by treating the asymmetric Laplace distribution (ALD) as the likelihood function because frequentist quantile regression maximizes this likelihood for an arbitrary scale parameter. However, because the scale parameter of the ALD is arbitrary, posterior standard deviations (that are greatly influenced by the scale parameter) cannot be used as standard errors (Yang et al., 2016; Lee, 2020). Lee (2020) shows that the correction to the standard errors suggested by Yang et al. (2016) also often performs poorly. We therefore compare the approximate resampling standard errors based on the Bayesian IJ with both the naive posterior standard deviations and ordinary bootstrap standard errors in a simulation study. We also discuss how to extend the standard errors for Bayesian ALD quantile regression when the data are clustered and compare them with cluster-robust standard errors proposed for frequentist quantile regression by Hagemann (2017).

The structure of this chapter is as follows. In Section 2.2 we provide an overview of the IJ in a frequentist setting (Verweij and Van Houwelingen, 1993; Giordano et al., 2019) and then link it to the Bayesian counterpart (Efron, 2015; Giordano et al., 2018; Giordano, 2020) with a discussion on alternative methods (Lee et al., 2017; Vehtari et al., 2017). We also show how these approximations can be used to estimate standard errors for Bayesian ALD quantile regression, both for unclustered and clustered data. For the unclustered case, we present a simulation study in Section 2.3 to evaluate the empirical coverage and mean length of confidence intervals based on these standard errors, compared with other currently available standard errors. In Section 2.4, we evaluate the standard errors for the clustered case using a series of simulation studies. We end with a general discussion in Section 2.5.

2.2 Methodology

Resampling approaches typically require parameters to be estimated in each resampled dataset, and this is often not feasible because estimation is computationally demanding. In this section, we provide an overview of the general methodology for first order approximations for both maximum likelihood and Bayesian posterior mean parameter estimates for resampled data.

2.2.1 Approximation of MLE estimates in resampled data

The general idea, from Giordano et al. (2019), is to use parameter estimates for the full sample and a Taylor-expansion around these estimates to obtain approximate parameter estimates in resampled data. Let w_i be the number of times unit i is represented in the resampled data, and let w be the corresponding weighting vector. Then, for any M-estimators (although we focus primarily on maximum-likelihood estimators), we can define the estimates $\hat{\theta}(w)$ with regard to the weighting vector $w \in \mathbb{R}^n$ as the solution of the weighted estimating equation, with $w = 1_n$ recovering the full sample estimates

$$\sum_{i=1}^n w_i \ell'_i(\theta) = 0,$$

where $\ell'_i(\theta) \in \mathbb{R}^p$ is the column vector of gradients of the log-likelihood contribution $\ell_i(\theta)$ for the i th unit with respect to θ .

Giordano et al. (2019)'s definition 2 (also a classic result from robust statistics literature, e.g., Cook and Weisberg (1982), as well as approximated leave-one-out literature, e.g., Verweij and Van Houwelingen (1993)) shows that the first-order Taylor expansion for $\hat{\theta}(w)$ around $w = 1_n$ (the full sample) can be written as

$$\hat{\theta}(w) \approx \hat{\theta}_{\text{IJ}}(w) := \hat{\theta}(1_n) + \left. \frac{d\hat{\theta}(w)}{dw^\top} \right|_{w=1_n} (w - 1_n) \quad (2.1)$$

$$= \hat{\theta}(1_n) - H(1_n)^{-1} S(1_n) (w - 1_n), \quad (2.2)$$

where $H(1_n) \in \mathbb{R}^{p \times p}$ is the Hessian of the log likelihood function for the full sample (with weights set to 1 for each unit) evaluated at $\hat{\theta}(1_n)$ and $S(1_n) \in \mathbb{R}^{p \times n}$ is a matrix whose i th column contains the derivatives of unit i 's contribution to the log likelihood function with respect to the parameters, evaluated at $\hat{\theta}(1_n)$. The second equation has a similar form as the updating equation in the Newton-Raphson algorithm where the current estimates are based on the full sample and the next iteration is for the estimates in the reweighted sample. However, unlike a one-step Newton-Raphson update, the Hessian is based on the

unweighted data. Giordano et al. (2019) calls this approximation (along with modern auto-differentiation) a Swiss-army “Infinite-Jackknife (IJ)” approximation and provides theoretical guarantees. We hence denote the parameter estimates $\hat{\theta}_{\text{IJ}}(w)$.

More formally, to show that $\left. \frac{d\hat{\theta}(w)}{dw^\top} \right|_{w=1_n} = -H(1_n)^{-1} S(1_n)$, we make use of the fact that, for $\hat{\theta}(w)$ to maximize the log likelihood function in the reweighted sample, the score function evaluated at the estimates for the reweighted sample must satisfy $S(\hat{\theta}(w), w) = 0 \in \mathbb{R}^p$, where $S(\theta, w) = \sum_{i=1}^n w_i \ell'_i(\theta)$. Taking derivative with respect to w on both sides,

$$\begin{aligned} 0 &= \left. \frac{d}{dw^\top} S(\hat{\theta}(w), w) \right|_w \\ &= \left(\frac{\partial S(\hat{\theta}(w), w)}{\partial \theta^\top} \frac{d\hat{\theta}(w)}{dw^\top} + \frac{\partial S(\hat{\theta}(w), w)}{\partial w^\top} \right) \Big|_w. \end{aligned}$$

Rearranging terms and setting $w = 1_n$, we have

$$\left. \frac{d\hat{\theta}(w)}{dw^\top} \right|_{w=1_n} = -H(1_n)^{-1} S(1_n).$$

2.2.2 Bayesian approximation

Following the same logic, which has been discussed by Efron (2015), Giordano et al. (2018), and Giordano (2020), we can obtain approximate Bayesian estimates for resampled data in a similar fashion. The Bayesian approximation to the parameter estimates based on a weighted log likelihood is

$$\hat{\theta}^{\text{Bayes}}(w) \approx \hat{\theta}_{\text{IJ}}^{\text{Bayes}}(w) := \hat{\theta}^{\text{Bayes}}(1_n) + \left. \frac{d\hat{\theta}^{\text{Bayes}}(w)}{dw^\top} \right|_{w=1_n} (w - 1_n) \quad (2.3)$$

$$= \hat{\theta}^{\text{Bayes}}(1_n) + \text{cov}_{\theta|x}(\theta, \ell(x | \theta)) (w - 1_n). \quad (2.4)$$

The required matrix of derivatives is equal to the posterior covariance matrix of the parameters (i.e., $\theta \in \mathbb{R}^p$) and the log-likelihood contributions from individual units (i.e., $\ell(x | \theta) \in \mathbb{R}^n$), and this covariance matrix can be estimated straightforwardly by the corresponding empirical covariance matrix of the MCMC samples.

The derivation is outlined below, following section 2 of Efron (2015). To start, the Bayesian (posterior mean) estimator for the reweighted sample is

$$\hat{\theta}^{\text{Bayes}}(w) \approx \mathbb{E}[\theta | x, w] = \frac{\int \theta \exp(L(x | \theta, w)) p(\theta) d\theta}{\int \exp(L(x | \theta, w)) p(\theta) d\theta},$$

where x is the dataset, $L(x|\theta, w) = \sum_i \ell(x_i|\theta)$ is the log-likelihood for the reweighted data and $\ell(x_i|\theta)$ are the log likelihood contributions from the units. Following our previous notation, we will suppress reference to the data x and use the notation $\ell_i(\theta) := \ell(x_i|\theta)$ and $L(\theta, w) := L(x|\theta, w)$, so that $L(\theta, w) = \sum_i w_i \ell_i(\theta)$.

We can derive the derivatives by using the fact that $(\frac{M}{N})' = (\frac{M'}{N}) - \frac{N'}{N}$ and setting

$$M(w) = \int \theta \exp(L(\theta, w)) p(\theta) d\theta \quad (2.5)$$

$$N(w) = \int \exp(L(\theta, w)) p(\theta) d\theta, \quad (2.6)$$

to the numerator and denominator of the Bayesian estimator, $\hat{\theta}^{\text{Bayes}} = \frac{M(w)}{N(w)}$. The partial derivatives of $M(w)$ and $N(w)$ with respect to w are

$$M'(w) = \int L'(\theta, w) \theta \exp(L(\theta, w)) p(\theta) d\theta = N(w) E[L'(\theta, w) \theta] \quad (2.7)$$

$$N'(w) = \int L'(\theta, w) \exp(L(\theta, w)) p(\theta) d\theta = N(w) E[L'(\theta, w)], \quad (2.8)$$

where the expectations are the posterior expectations (over the parameters given the data) and $L'(\theta, w)$ is equal to the vector of log-likelihood contributions from the units, $L'(\theta, w) = \ell(\theta) := (\ell_1(\theta), \dots, \ell_n(\theta))^T$ since

$$\left. \frac{d \sum_i w_i \ell_i(\theta)}{dw_i} \right|_{w=1} = \ell_i(\theta).$$

Substituting these derivatives into $(\frac{M}{N})' = (\frac{M'}{N}) - \frac{N'}{N}$, we obtain

$$\begin{aligned} \left. \frac{d \hat{\theta}^{\text{Bayes}}(w)}{dw^\top} \right|_{w=1_n} &= \frac{M(w)}{N(w)} \left\{ \frac{N(w) E[L'(\theta, w) \theta]}{M(w)} - \frac{N(w) E[L'(\theta, w)]}{N(w)} \right\} \\ &= E[\theta] \left\{ \frac{E[\ell(\theta) \theta]}{E[\theta]} - E[\ell(\theta)] \right\} \\ &= E[\ell(\theta) \theta] - E[\theta] E[\ell(\theta)] \\ &= \text{cov}[\theta, \ell(\theta)]. \end{aligned}$$

There are other computationally efficient approximations for Bayesian resampling primarily using importance sampling ideas, which can provide approximated Bayesian resampling estimates, similarly as stated above. For example, Lee et al. (2017) utilizes posterior draws of one round of MCMC with the full data (i.e., proposal distributions) and the likelihood ratio of bootstrap-resampled data and original data as importance weights to estimate the posterior means for all the resampled data with a single Bayesian run.

In the context of model evaluation and leave one out, Vehtari et al. (2017) proposes to use a version of the generalized Pareto distribution to smooth the importance weights so that a better bias-variance trade-off can be achieved, and the procedure is referred to as PSIS, Pareto-smoothed importance sampling. Using this approach to approximate the leave-one-out predictive density for Bayesian model assessment is viewed as a critical step for the modern Bayesian workflow (e.g., Gelman et al., 2020). Recent extensions of this work include use of moment-matching together with importance sampling so that it works for complicated distributions without closed form (Paananen et al., 2021).

However, we apply the approach based on Efron (2015) and Giordano (2020) in the current study because we think it is the simplest one to implement in practice without extra steps required after carrying out a Bayesian data analysis (with posterior draws available). This approach therefore has huge potential in general applied Bayesian analysis.

2.2.3 Obtaining resampling estimates and standard errors

Both of the approximations for parameter estimates in resampled data can be used for jackknife or bootstrap standard errors by specifying the weighting vector w in different ways.

As noted in Efron and Tibshirani (1994, Chapter 20), many resampling schemes in statistics can be unified from a reweighted data perspectives. For example, given a generic full data sample, we can define a corresponding weighting vector $w = (w_1, w_2, \dots, w_i, \dots, w_{n-1}, w_n)^\top$, where $w_i = k$ means that unit i is included k times in the resampled data and 0 means that the unit is not in the reweighted sample. For multivariate and regression settings, we can define the data for each unit to be multidimensional or a tuple of outcome and corresponding covariates. For clustered data, we often treat the set of units within one cluster as a “unit” when considering different resampling schemes (e.g., leave-one-cluster out or cluster bootstrap, introduced below) because units within each cluster are not independent. This point has been long acknowledged and discussed in the multilevel modelling literature (e.g., Goldstein, 2011).

Leave- k -out is probably the simplest resampling scheme. When $k = 1$, the r^{th} weight vector, denoted $w^{(r)}$, has $w_r = 0$ and all other elements equal to 1. For clustered data that will be introduced later, we can also consider leaving one cluster out, i.e., giving weight 0 to all the units that belong to one cluster and 1 to all the other units. Leave-one-cluster out (or leave-one-person/group out, depending on the grouping unit) has been discussed in the context of inference and model assessment for multilevel/hierarchical data in both frequentist and Bayesian settings (Fang, 2011; Merkle et al., 2019).

To implement the bootstrap, we can draw the weight vector from a multinomial distribution (i.e., $w \sim \text{multinomial}(n, (\frac{1}{n}, \dots, \frac{1}{n}) \in \mathbb{R}^n)$). Or for clustered data, we can instead draw J integers from $\text{multinomial}(J, (\frac{1}{J}, \dots, \frac{1}{J}) \in \mathbb{R}^J)$ and assign the j th drawn integer to the elements of w that correspond to the j th cluster.

By specifying the weighting scheme described above to correspond to leave-one-unit (or cluster)-out, we can obtain the corresponding parameter estimates and hence empirical influence functions which have direct application to estimating IJ standard errors for both frequentist and Bayesian estimators.

Following the notation above, we can denote the full sample parameters estimates $\widehat{\theta}(1_n)$ and the leave-one-unit-out estimates $\widehat{\theta}(1_{(i)})$, where $1_{(i)}$ is the 1_n vector except that the i^{th} element is 0. If we denote loosely the “true” influence function regarding the parameter vector θ is I_i , then the jackknife estimate of the influence function (or so called “jackknife influence function”, equation (2.3) in Efron, 1992) is

$$\widehat{I}_i^{\text{ack}} = (n-1) \left(\widehat{\theta}(1_{(\cdot)}) - \widehat{\theta}(1_{(i)}) \right), \quad (2.9)$$

where $\widehat{\theta}(1_{(\cdot)}) = \frac{1}{n} \sum_{i=1}^n \widehat{\theta}(1_{(i)})$. Another similar alternative has been proposed by Devlin et al. (1975) of Bell Laboratories. They define the sample influence function (their equation 2) as

$$\widehat{I}_i^{\text{Bell}} = (n-1) \left(\widehat{\theta}(1_n) - \widehat{\theta}(1_{(i)}) \right), \quad (2.10)$$

where $\widehat{\theta}(1_{(\cdot)})$ in (2.9) is approximately equal to the full-sample estimate $\widehat{\theta}(1_n)$ in (2.10), and the approximation is exact for a linear statistic. The corresponding unscaled version of this difference $\widehat{\theta}(1_n) - \widehat{\theta}(1_{(i)})$ is often referred to as “jackknife influence values” (e.g., equation 11.10, Efron and Tibshirani, 1994). Notice that both of the vectors can be interpreted as a scaled change in parameter estimates when dropping unit i .

Use of empirical influence functions for estimating variances has been extensively studied and discussed in the literature. For example, chapter 21 in Efron and Tibshirani (1994) gives a detailed theoretical account of how influence functions can be used to derive variance estimates (their equation 21.17), and this information can also be found in section 2.7 of Davison and Hinkley (1997). Specifically, the asymptotic covariance matrix of parameters can be expressed in terms of the (first-order) influence function vector I_i as

$$\Sigma(\widehat{\theta}) = \frac{1}{n} \text{var}(I_i) = \frac{1}{n} E(I_i I_i^T). \quad (2.11)$$

The second equality holds because $E(I_i) = 0 \in \mathbb{R}^p$ under appropriate regularity conditions.

Combining equation (2.10) and (2.11), we can obtain one version of jackknife variance estimator based on $\widehat{I}_i^{\text{Bell}}$,

$$\left(\frac{n-1}{n} \right)^2 \sum_{i=1}^n \left(\widehat{\theta}(1_n) - \widehat{\theta}(1_{(i)}) \right) \left(\widehat{\theta}(1_n) - \widehat{\theta}(1_{(i)}) \right)^T,$$

which corresponds to equation (21.18) in Efron (1982). Similarly, combining equation (2.9) and (2.11), we can obtain another version of jackknife variance estimator based on $\hat{I}_i^{\text{jack}}$,

$$\left(\frac{n-1}{n}\right)^2 \sum_{i=1}^n \left(\hat{\theta}(1_{(\cdot)}) - \hat{\theta}(1_{(i)})\right) \left(\hat{\theta}(1_{(\cdot)}) - \hat{\theta}(1_{(i)})\right)^\top,$$

The well-known jackknife variance formula (Quenouille, 1956; Tukey, 1958) differs in the scaling factor, which ensures unbiasedness for some estimators (e.g., standard error of sample mean), namely

$$\left(\frac{n-1}{n}\right) \sum_{i=1}^n \left(\hat{\theta}(1_{(\cdot)}) - \hat{\theta}(1_{(i)})\right) \left(\hat{\theta}(1_{(\cdot)}) - \hat{\theta}(1_{(i)})\right)^\top.$$

This variance estimator can also be interpreted as combining a different version of the empirical influence function (equation 10.12, Efron and Hastie, 2021),

$\hat{I}_i^* = \sqrt{n(n-1)} \left(\hat{\theta}(1_{(i)}) - \hat{\theta}(1_{(\cdot)})\right)$, with (2.11). It is easy to verify that the above expression reduces to the estimator of the sampling variance of a sample mean.

There are many ways to define influence functions for a given parameter of interest. For example, the positive jackknife (e.g., Hinkley, 1978; Beran, 1984; Andrews and Hampel, 2015) estimate of influence is defined as

$$\hat{I}_i^{\text{positive-jack}} = (n+1) \left(\hat{\theta}(1_n) - \hat{\theta}(1_{(i)})\right),$$

which yields a positive jackknife variance estimate (Efron and Tibshirani, 1994):

$$\left(\frac{n+1}{n}\right)^2 \sum_{i=1}^n \left(\hat{\theta}(1_n) - \hat{\theta}(1_{(i)})\right) \left(\hat{\theta}(1_n) - \hat{\theta}(1_{(i)})\right)^\top.$$

The differences in scaling factors essentially correspond to different choices of ϵ , the proportion of mass we consider dropping, in defining the influence functions (e.g., $\epsilon = \frac{1}{n+1}$ for the positive jackknife case). The impact of the scaling factor on variance estimates vanishes when the sample size becomes large. In our application, we follow the original work on IJ variance estimation by Jaeckel (1972) and Buzas (1997), and also recent discussions of Alaa and Van Der Schaar (2020), Broderick et al. (2020), and Giordano (2020) and use n in defining the IJ empirical influence function for frequentist and Bayesian estimators, since it corresponds to the leave-one-out scenario. For the frequentist case, using the result in (2.2), the empirical influence function based on the IJ approximation is

$$\begin{aligned} \hat{I}_i^{\text{IJ}} &= n \left(\hat{\theta}_{\text{IJ}}(1_n) - \hat{\theta}_{\text{IJ}}(1_{(i)})\right) \\ &= nH(1_n)^{-1} S(1_n) (1_{(i)} - 1_n). \end{aligned} \tag{2.12}$$

Analogously, using the result in (2.4), the Bayesian IJ empirical influence function can be defined as

$$\begin{aligned}\widehat{I}_i^{\text{BIJ}} &= n \left(\widehat{\theta}_{\text{IJ}}^{\text{Bayes}}(1_n) - \widehat{\theta}_{\text{IJ}}^{\text{Bayes}}(1_{(i)}) \right) \\ &= -n \widehat{\text{cov}}_{\theta|x} \left(\widehat{\theta}, \ell(x | \widehat{\theta}) \right) (1_{(i)} - 1_n) .\end{aligned}\tag{2.13}$$

We can substitute $\widehat{I}_i^{\text{IJ}}$ or $\widehat{I}_i^{\text{BIJ}}$ as sample counterparts for I_i in equation (2.11) to obtain the corresponding standard errors. Following the advice by Efron (1981), we also multiply the final IJ variance estimate by an additional correction factor $\frac{n}{n-1}$ to reduce possible downward bias. The frequentist IJ variance estimator has been shown to be equal to the variance estimator obtained by the delta method under certain conditions in result 21.3 in Efron (1982), and to the ordinary jackknife variance estimator asymptotically under a wide range of conditions (Jaekel, 1972).

For clustered data, leave-one-unit out is replaced by leave-one-cluster out by defining the weights accordingly, and the sample size becomes J , the number of clusters. Specifically, we can define $1_{(j)}$, the leave-one-cluster out weights as a vector with each element equal to one except for the elements that belong to cluster j which are 0. Then we use the corresponding correction factor $\frac{J}{J-1}$. We will empirically evaluate the coverage of intervals constructed by using the Bayesian IJ standard errors for Bayesian ALD quantile regression for both independent data and clustered data through a series of simulation studies in following two sections.

2.3 Bayesian ALD quantile regression

Quantile regression (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001) provides a flexible general framework for inference on conditional quantiles.

In the original frequentist setup, the point estimates are often obtained as numerical solutions to the minimization problem, shown below for quantile τ and a model with one scalar covariate x_i (for simplicity):

$$\widehat{\beta}(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \rho_{\tau}(y_i - \beta_0 - \beta_1 x_i),$$

where $\rho_{\tau}(\cdot)$, often called “check function” for its shape, is defined as

$$\rho_{\tau}(u) = u\{\tau - 1(u < 0)\} = \begin{cases} u\tau, & \text{if } u \geq 0 \\ -u(1 - \tau), & \text{if } u < 0 \end{cases} .$$

Bayesian inference requires a likelihood function, and there have been several attempts in the literature to use a working likelihood for Bayesian quantile regression. Yang and He

(2012) proposed using the empirical likelihood and Kottas and Krnjajić (2009) specified a Dirichlet process mixture model, but the simplest and most popular working likelihood has been the asymmetric Laplace distribution (ALD), sometimes also referred as skew double exponential distribution (Yu and Moyeed, 2001; Yu and Zhang, 2005; Geraci and Bottai, 2007; Benoit and Van den Poel, 2012; Geraci and Bottai, 2014). ALD has been used as a default choice for many popular Bayesian regression packages such as **brms** (Bürkner, 2017), **bayesQR** (Benoit and Van den Poel, 2017) because of its simplicity, while most other Bayesian quantile methods (e.g., Yang and He, 2012; Yang and Tokdar, 2017) are currently not directly accessible to applied researchers.

Specifically, we assume $y|\mu, \sigma, \tau \sim \text{AL}(\mu, \sigma, \tau)$,

$$f_{\text{ALD}}(y_i|\mu_i, \sigma, \tau) = \frac{\tau(1 - \tau)}{\sigma} \exp\left(-\rho_\tau\left(\frac{y_i - \mu_i}{\sigma}\right)\right),$$

where μ_i , equal to $\beta_0 + \beta_1 x_i$ in the current example, is the location parameter of interest. The scale parameter σ does not affect the point estimates and is therefore set to an arbitrary fixed number (e.g., 1). The only motivation for ALD is that maximizing this likelihood corresponds directly to minimizing the loss function used in classical quantile regression, $\sum_{i=1}^n \rho_\tau(y_i - \mu_i)$, where $\mu_i = \beta_0 + \beta_1 x_i$. There does not appear to be any motivation in terms of a data-generating process. Importantly, the scale parameter σ makes no difference in maximum-likelihood estimation, but it will affect posterior variances in Bayesian inference. Therefore, not surprisingly, it has been shown that the point estimates are valid but that credible intervals based on ALD do not have correct frequentist coverage (Sriram, 2015; Yang et al., 2016; Lee, 2020).

Yang et al. (2016) proposed a simple adjustment to the posterior covariance matrix, similar to the frequentist sandwich idea, to improve the coverage of credible intervals with the ALD likelihood. However, as shown in an extensive simulation study in Lee (2020), naively relying on the ALD (model-based) interval or using Yang et al. (2016)’s adjustment both lead to poor coverage, and are sensitive to the choice of the scale parameter σ in finite samples.

Next, we investigate the coverage of confidence intervals based on the Bayesian IJ estimator of the standard error using the leave-one-out weighting scheme introduced above. Note that these are strictly not Bayesian credible intervals because the standard errors no longer correspond to posterior standard deviations. Through a series of simulation studies, we show that our proposed intervals have close to nominal frequentist coverage and are not sensitive to the choice of scale parameter σ .

2.3.1 Empirical coverage for Bayesian quantile regression

To make our simulation results comparable to Lee (2020)’s results on demonstrating that Yang et al. (2016)’s adjustment is very sensitive to the choice of the scale parameter of

ALD, we follow the simulation setup in Lee (2020), and specify the data-generating models described below.

- Model 1: Location shifted normal model: $y_i = \alpha + \beta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. The conditional quantile function for the τ^{th} conditional quantile is $Q_\tau(y_i | x_i) = [\alpha + \Phi^{-1}(\tau)] + \beta x_i$. This model has a shifted location since the intercept α shifts by $\Phi^{-1}(\tau)$ for different τ and a homoscedastic error which translates into parallel lines for the quantiles, with constant slope β for different τ .
- Model 2: Location shifted and scaled normal model: $y_i = \alpha + \beta x_i + (1 + \gamma x_i) \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. The conditional quantile function for the τ^{th} quantile is $Q_\tau(y_i | x_i) = [\alpha + \Phi^{-1}(\tau)] + [\beta + \gamma \Phi^{-1}(\tau)] x_i$. This model has a shifted location since the intercept α shifts by $\Phi^{-1}(\tau)$ for different τ . The heteroscedastic error results in different slopes, $\beta + \gamma \Phi^{-1}(\tau)$, for different τ .

For the models above, we consider three choices of sample size (200, 500, and 1000) and fix $\alpha = \beta = 2$ for both models and $\gamma = .3$ for Model 2. We set the ALD scale parameter σ to $\{.1, .2, .5, 1, 2, 5, 10\}$ and obtain model-based credible intervals (denoted as “ALD”) and Bayesian IJ-based intervals (denoted “IJ”) for five different quantiles, $\tau \in \{.1, .3, .5, .7, .9\}$. The empirical 90% coverage of the intervals (using a normality-based approximation centered at the posterior mean) is estimated for 100 replications for each of the simulation conditions. x_i is simulated from a standard Gaussian distribution. For comparison, we also include 90% confidence intervals based on frequentist point estimates and two kinds of standard errors, the sandwich estimator (denoted “nid”) and bootstrap (denoted “boot”).

The simulation is carried out in **R** (R Core Team, 2021). “ALD” and “IJ” variance estimates are obtained using Stan (Carpenter et al., 2017) while “nid” and “boot” variance estimates are obtained by using the **quantreg** package (Koenker, 2009).

2.3.2 Results

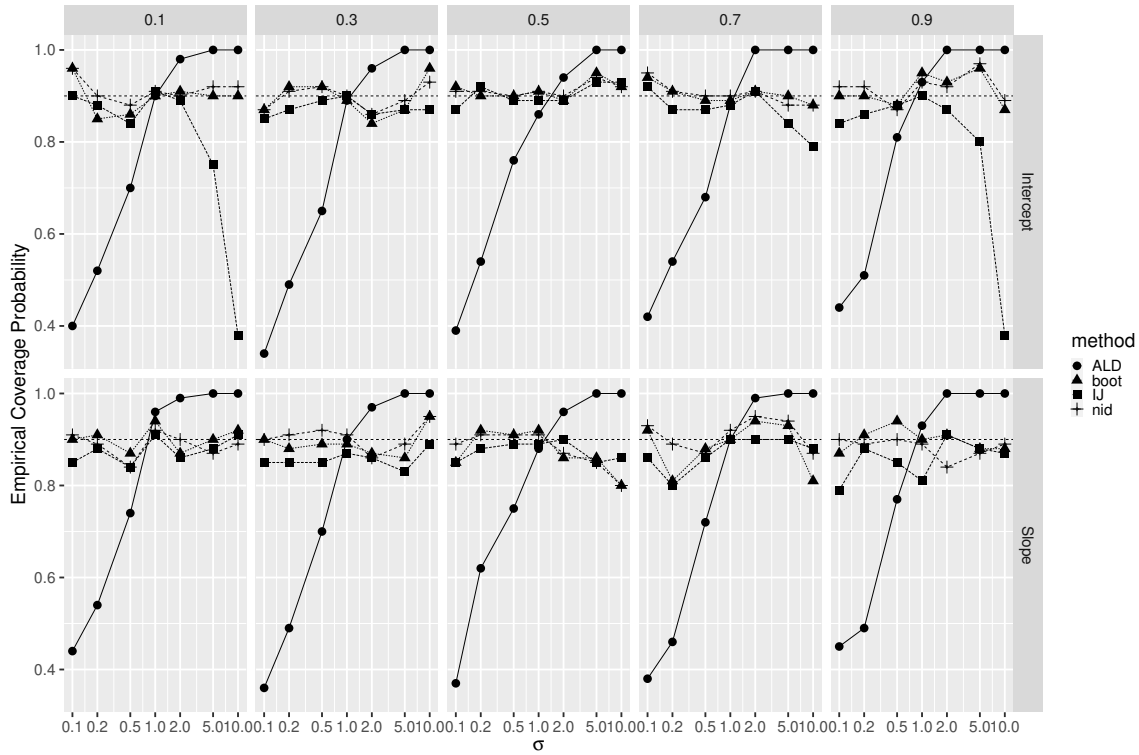
Figure 2.1 shows the empirical coverage of the intervals constructed by the four methods, represented by different line patterns, as a function of σ , for different quantiles τ (the columns of the array of panels). The rows are for different models and for the intercepts and slopes of the quantiles (Rows 1-2 for intercepts and slopes for Model 1 and rows 3-4 for intercepts and slopes for Model 2) for $N = 200$. Analogous graphs for $N = 500$ and $N = 1000$ are given in Figures 2.2 and 2.3, respectively.

For non-extreme quantiles (i.e., $\tau = .3, .5$ or $.7$), the coverage of the proposed IJ method performs similarly to the frequentist sandwich and bootstrap methods, all at around the nominal .9 level, across different choices of the scale parameter σ , sample size and data-generating model. This is expected because the Bayesian IJ method can be viewed as a

delta method that is closely related to the sandwich and bootstrap SE estimators (Chapter 21, Efron and Tibshirani, 1994).

However for the two extreme quantiles with large scale parameter (i.e., $\tau = .1$ or $.9$ and $\sigma = 5$ or 10), we see a deteriorated performance of the proposed IJ interval for both the intercept and the slope while the level of undercoverage is less so for the slope parameter. This is because the Bayesian point estimator for the intercept is severely biased for extreme quantiles with large scale parameters, as also observed in previous simulation studies (Lee, 2020, page 48-49). We can see that the under-coverage problem becomes less severe when we use the *frequentist* point estimates in Figures 2.4, 2.5, and 2.6 (numerical results are summarised in the “ECP” columns in Tables 2.1 to 2.7). We can also see clearly the problems with the ALD intervals, namely that their coverage is directly related to the choice of the scale parameter σ , which cannot be estimated from data. When σ is small, the interval tends to undercover and the opposite happens when σ is large. The results for the bootstrap intervals and IJ intervals are similar across conditions.

Table 2.1 to 2.7 also show the estimates for empirical mean length (EML) of different choices of intervals. We can see that the conclusion is consistent with what we observe for the empirical coverage. Specifically, the under-coverage observed before corresponds to a shorter interval length on average compared to the bootstrap interval. Also, the EML of the model-based ALD interval varies a lot depending on the scale parameter σ , corresponding to the sensitivity of coverage performance to the choice of the scale parameter in ALD. In general, the coverage for the intercept parameter is closer to the nominal level than that of the slope parameter.



$N = 200$, Model 2

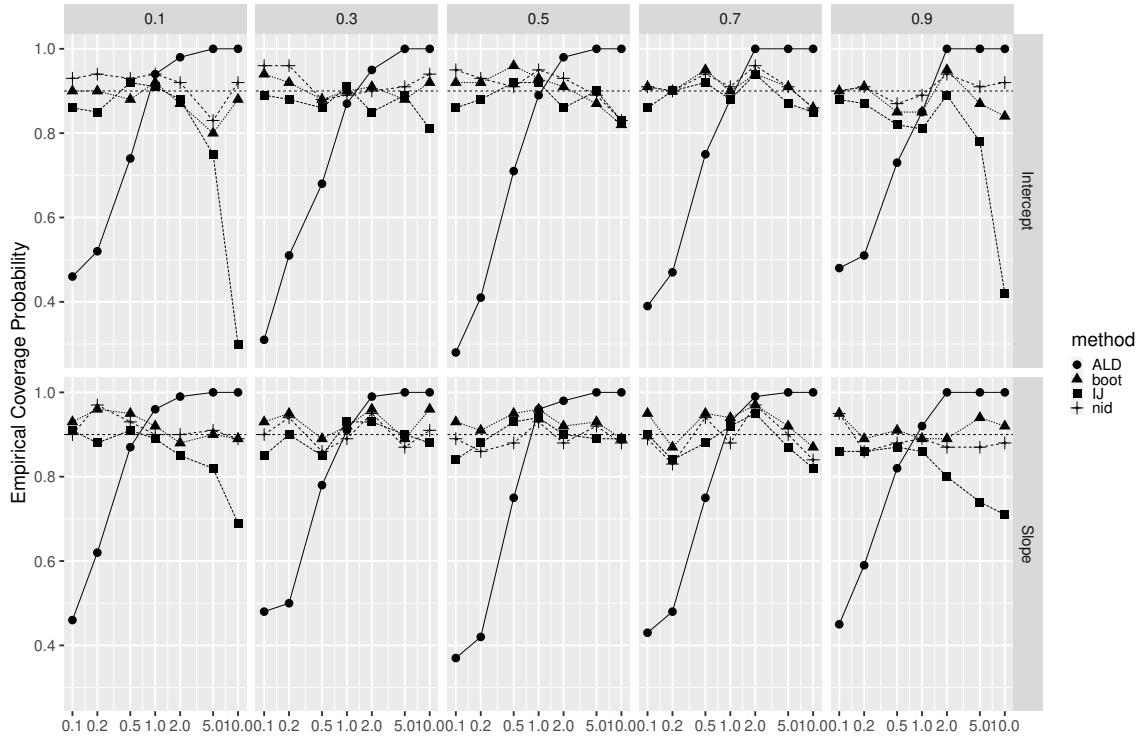
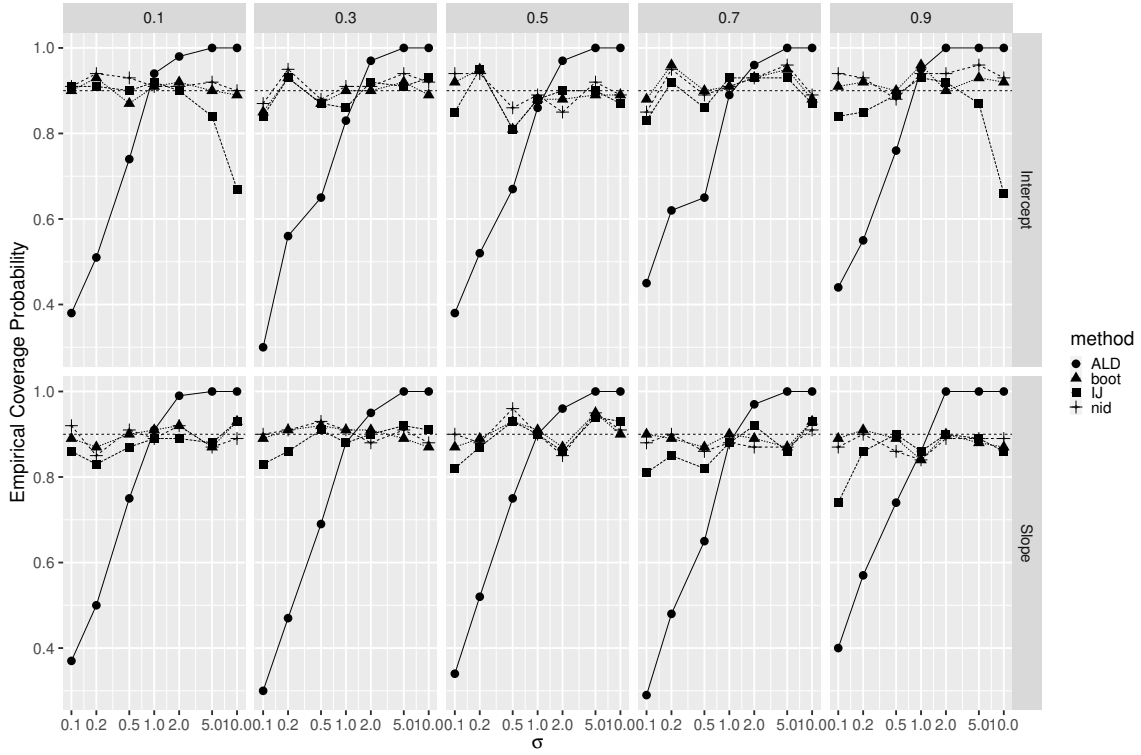


Figure 2.1: Empirical coverage for different SE methods as a function of the scale parameter σ for different quantiles (τ) for $N = 200$.



N = 500, Model 2

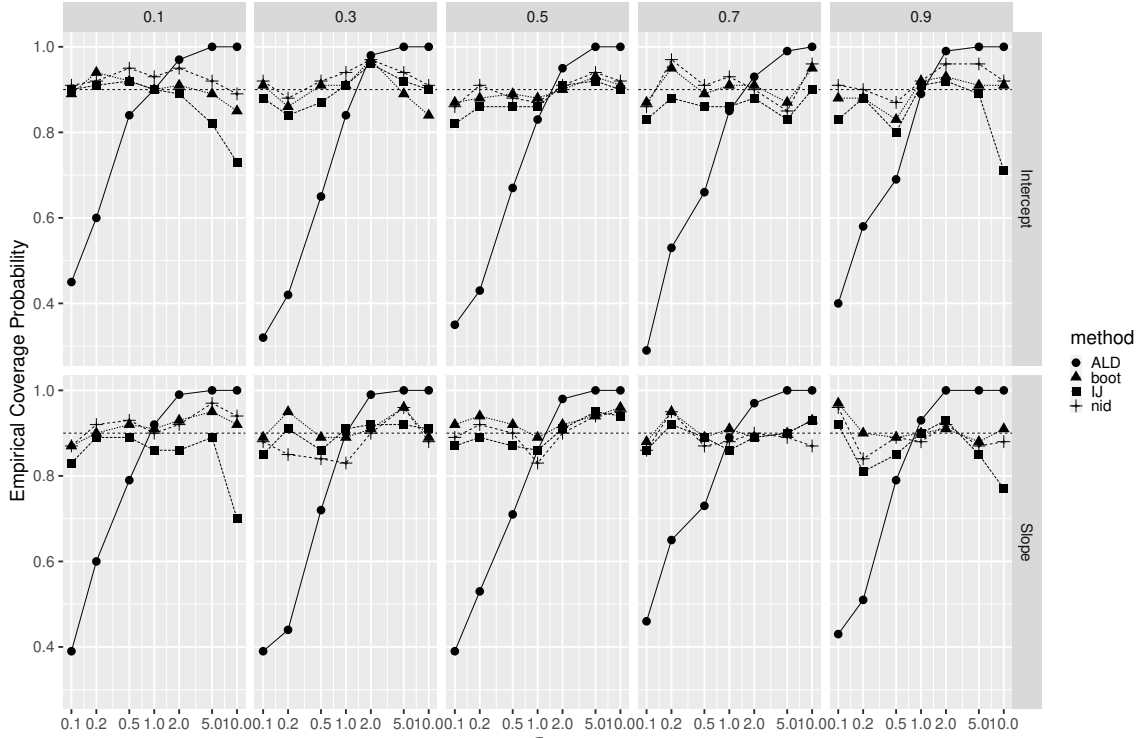
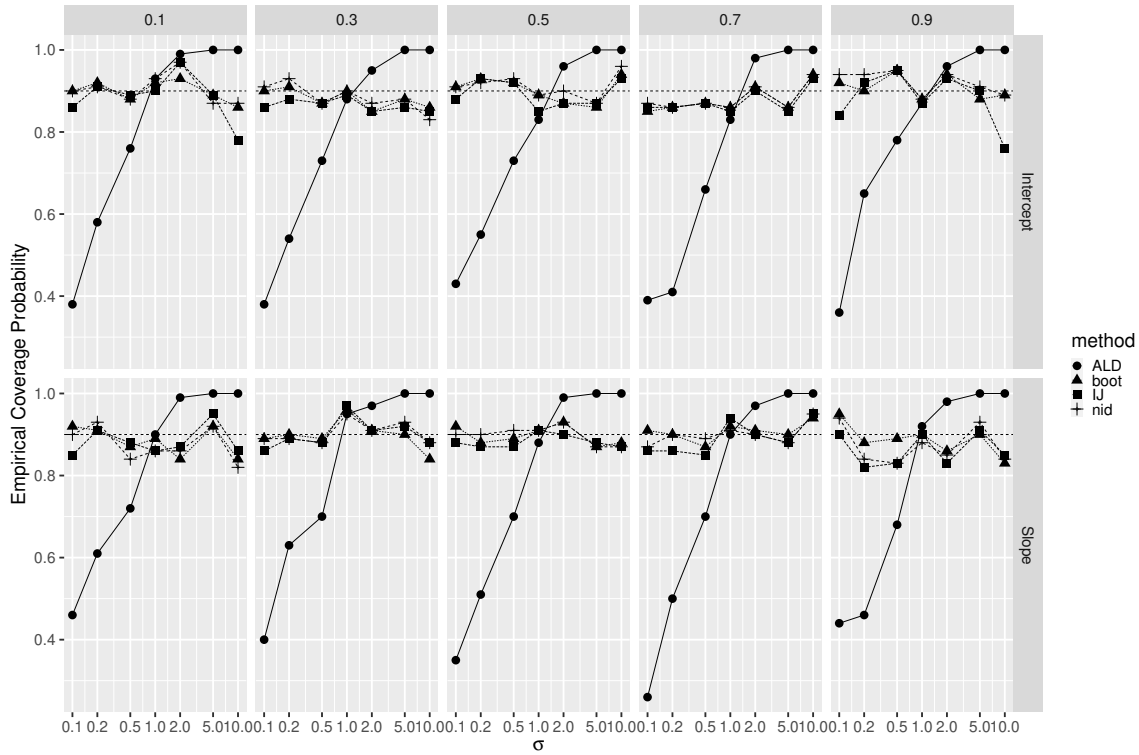


Figure 2.2: Empirical coverage for different SE methods as a function of the scale parameter σ for different quantiles (τ) for $N = 500$.



N = 1000, Model 2

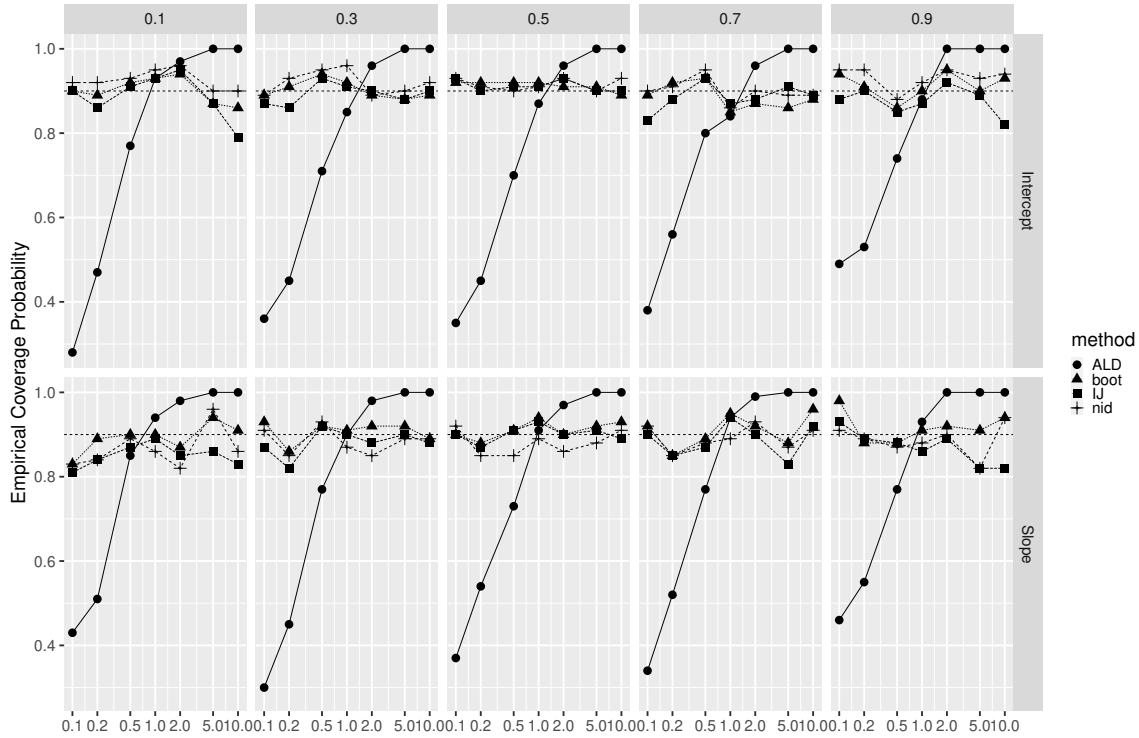
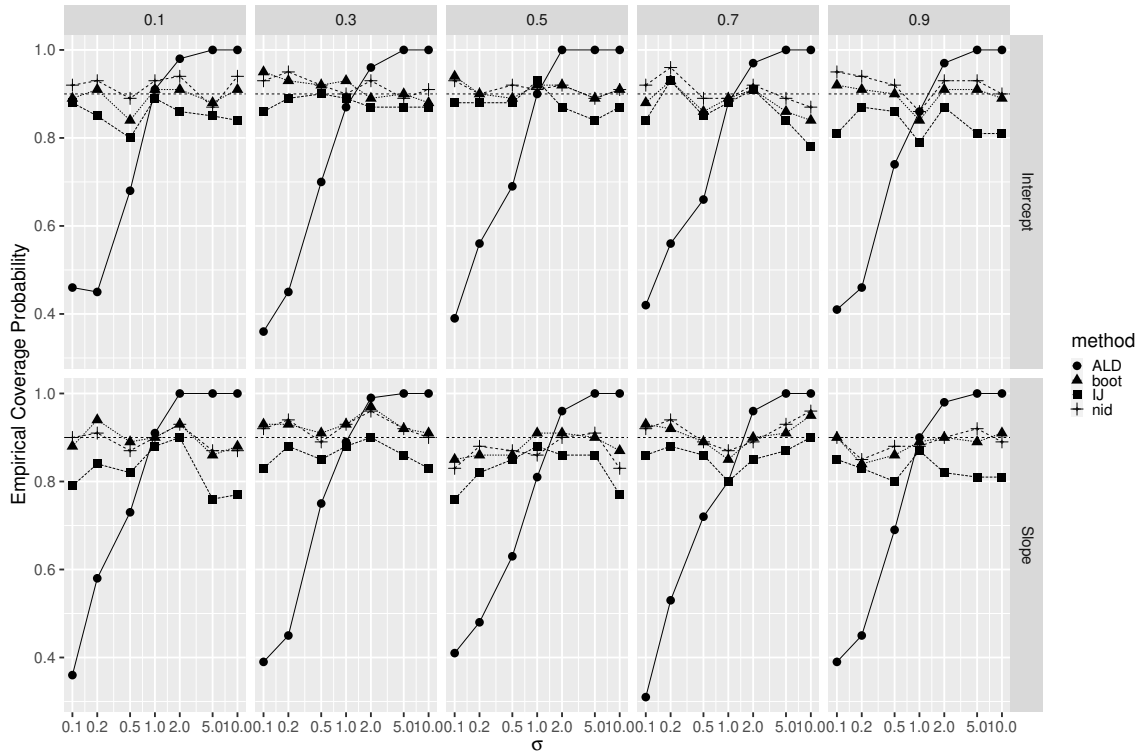


Figure 2.3: Empirical coverage for different SE methods as a function of the scale parameter σ for different quantiles (τ) for $N = 1000$.



$N = 200$, Model 2

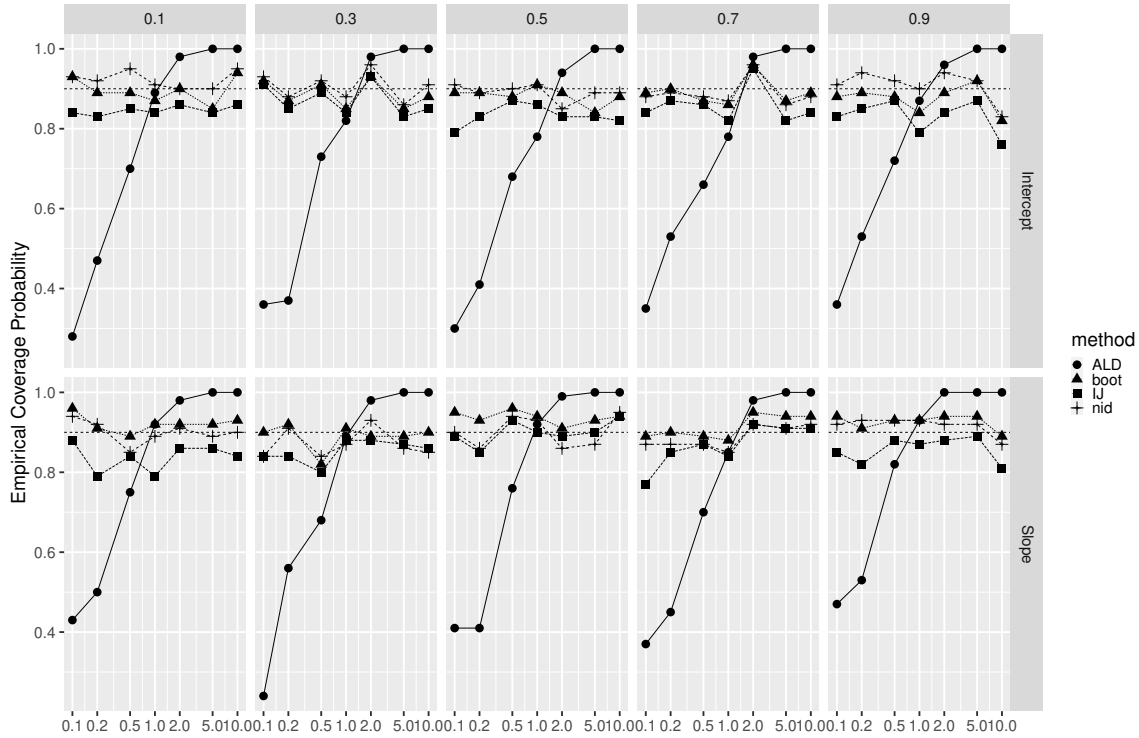
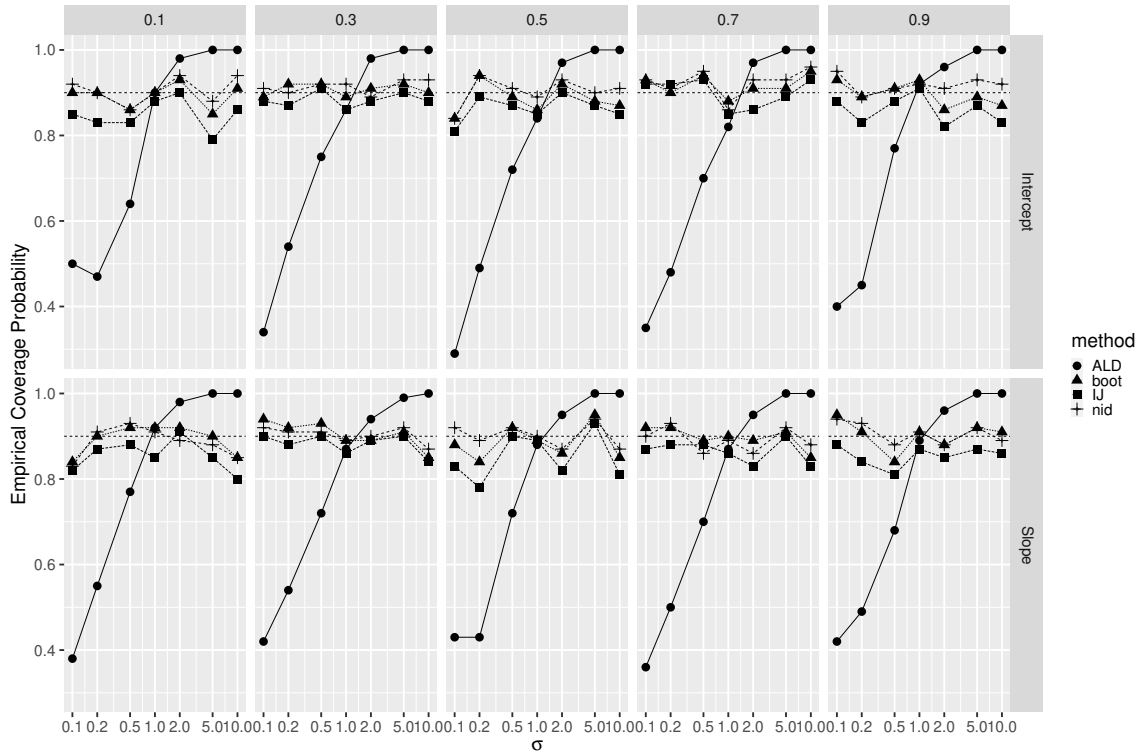


Figure 2.4: Empirical coverage for different $\hat{\sigma}$ SE methods as a function of the scale parameter σ with frequentist point estimates for different quantiles (τ) for $N = 200$.



N = 500, Model 2

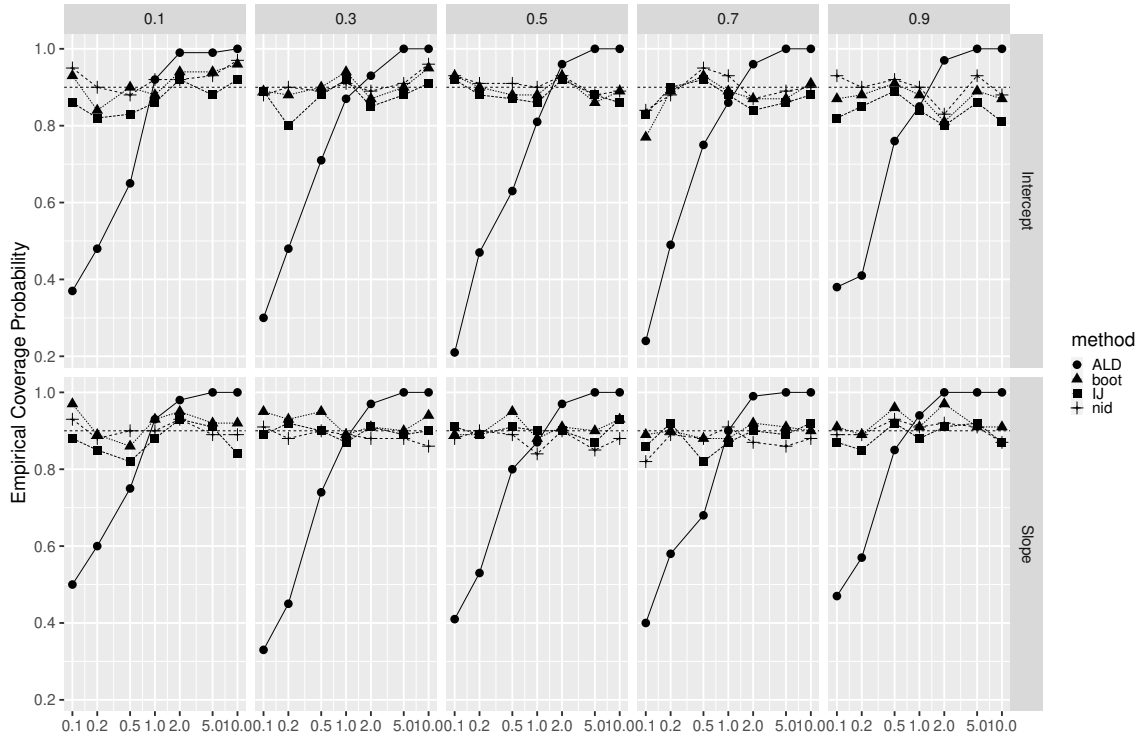
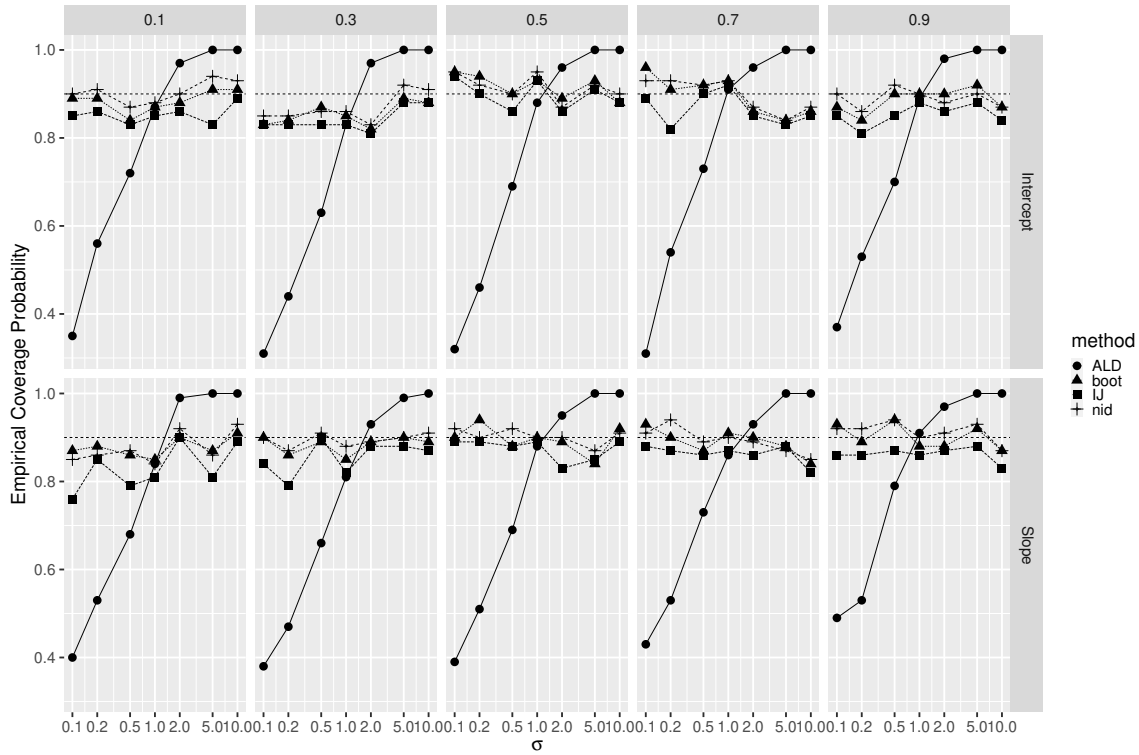


Figure 2.5: Empirical coverage for different $\hat{\sigma}$ SE methods as a function of the scale parameter σ with frequentist point estimates for different quantiles (τ) for $N = 500$.



N = 1000, Model 2

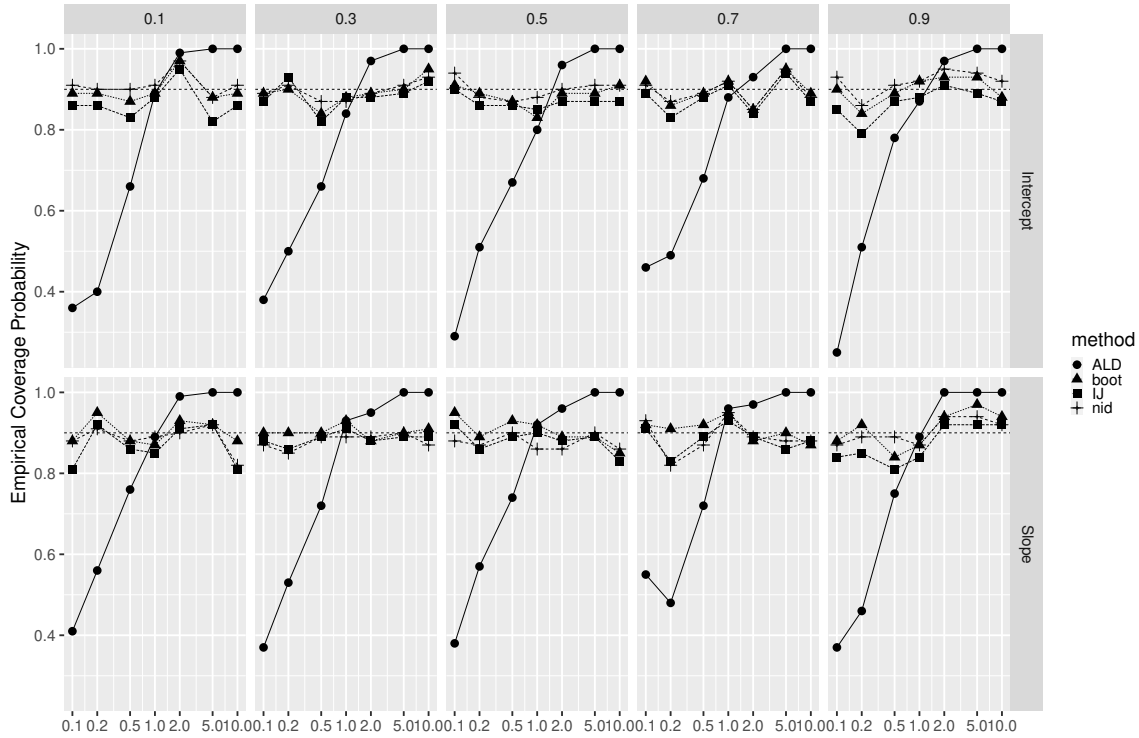


Figure 2.6: Empirical coverage for different $\hat{\sigma}$ SE methods as a function of the scale parameter σ with frequentist point estimates for different quantiles (τ) for $N = 1000$.

Table 2.1: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = .1$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	0.46	0.04	0.89	0.13	0.88	0.13	0.92	0.14	0.36	0.04	0.88	0.13	0.79	0.12	0.90	0.13
		500	0.50	0.02	0.90	0.08	0.85	0.08	0.92	0.08	0.38	0.02	0.84	0.08	0.82	0.08	0.83	0.08
		1000	0.35	0.02	0.89	0.06	0.85	0.05	0.90	0.06	0.40	0.02	0.87	0.06	0.76	0.05	0.85	0.06
	0.3	200	0.36	0.03	0.95	0.10	0.86	0.09	0.93	0.10	0.39	0.03	0.93	0.10	0.83	0.09	0.92	0.10
		500	0.34	0.02	0.89	0.06	0.88	0.06	0.91	0.06	0.42	0.02	0.94	0.06	0.90	0.06	0.92	0.06
		1000	0.31	0.01	0.83	0.04	0.83	0.04	0.85	0.04	0.38	0.01	0.90	0.04	0.84	0.04	0.90	0.04
	0.5	200	0.39	0.02	0.94	0.09	0.88	0.09	0.93	0.09	0.41	0.02	0.85	0.09	0.76	0.08	0.83	0.09
		500	0.29	0.02	0.84	0.06	0.81	0.06	0.84	0.06	0.43	0.02	0.88	0.06	0.83	0.06	0.92	0.06
		1000	0.32	0.01	0.95	0.04	0.94	0.04	0.95	0.04	0.39	0.01	0.90	0.04	0.89	0.04	0.92	0.04
	0.7	200	0.42	0.03	0.88	0.10	0.84	0.09	0.92	0.10	0.31	0.03	0.93	0.10	0.86	0.09	0.92	0.09
		500	0.35	0.02	0.93	0.06	0.92	0.06	0.93	0.06	0.36	0.02	0.92	0.06	0.87	0.06	0.90	0.06
		1000	0.31	0.01	0.96	0.04	0.89	0.04	0.93	0.04	0.43	0.01	0.93	0.04	0.88	0.04	0.91	0.04
	0.9	200	0.41	0.04	0.92	0.13	0.81	0.12	0.95	0.14	0.39	0.04	0.90	0.13	0.85	0.13	0.90	0.13
		500	0.40	0.02	0.93	0.08	0.88	0.08	0.95	0.08	0.42	0.02	0.95	0.08	0.88	0.08	0.94	0.08
		1000	0.37	0.02	0.87	0.06	0.85	0.05	0.90	0.06	0.49	0.02	0.93	0.06	0.86	0.06	0.92	0.06
2	0.1	200	0.28	0.04	0.93	0.14	0.84	0.13	0.93	0.14	0.43	0.03	0.96	0.12	0.88	0.10	0.94	0.11
		500	0.37	0.02	0.93	0.08	0.86	0.08	0.95	0.08	0.50	0.02	0.97	0.07	0.88	0.06	0.93	0.06
		1000	0.36	0.02	0.89	0.06	0.86	0.06	0.91	0.06	0.41	0.01	0.88	0.04	0.81	0.04	0.88	0.04
	0.3	200	0.36	0.03	0.92	0.10	0.91	0.10	0.93	0.10	0.24	0.02	0.90	0.09	0.84	0.08	0.84	0.08
		500	0.30	0.02	0.89	0.06	0.89	0.06	0.88	0.06	0.33	0.01	0.95	0.05	0.89	0.05	0.91	0.05
		1000	0.38	0.01	0.89	0.04	0.87	0.04	0.89	0.04	0.37	0.01	0.90	0.03	0.88	0.03	0.87	0.03
	0.5	200	0.30	0.02	0.89	0.10	0.79	0.09	0.91	0.10	0.41	0.02	0.95	0.09	0.89	0.08	0.90	0.08
		500	0.21	0.02	0.93	0.06	0.92	0.06	0.93	0.06	0.41	0.01	0.89	0.05	0.91	0.05	0.88	0.05
		1000	0.29	0.01	0.91	0.04	0.90	0.04	0.94	0.04	0.38	0.01	0.95	0.03	0.92	0.03	0.88	0.03
	0.7	200	0.35	0.03	0.89	0.10	0.84	0.10	0.88	0.10	0.37	0.02	0.89	0.09	0.77	0.09	0.87	0.08
		500	0.24	0.02	0.77	0.06	0.83	0.06	0.84	0.06	0.40	0.01	0.89	0.05	0.86	0.05	0.82	0.05
		1000	0.46	0.01	0.92	0.05	0.89	0.04	0.91	0.04	0.55	0.01	0.92	0.04	0.91	0.03	0.93	0.03
	0.9	200	0.36	0.04	0.88	0.13	0.83	0.12	0.91	0.14	0.47	0.03	0.94	0.12	0.85	0.11	0.92	0.12
		500	0.38	0.02	0.87	0.08	0.82	0.08	0.93	0.08	0.47	0.02	0.91	0.07	0.87	0.06	0.89	0.06
		1000	0.25	0.02	0.90	0.06	0.85	0.06	0.93	0.06	0.37	0.01	0.88	0.05	0.84	0.04	0.87	0.05

2.4 Bayesian ALD quantile regression with clustered data

Clustered data (or multilevel/hierarchical data) are very common in applied settings across different fields (e.g., patients nested in hospitals, residents nested in street blocks, students nested in classes, longitudinal/repeated-measure data). Such clustering violates the independence assumption relied on for estimating standard errors using leave-one-unit out methods. As discussed in Section 2.2.3, we will therefore use leave-one-cluster out methods instead.

For linear models, there has been an extensive literature on this problem, including explicit modeling of the between-cluster heterogeneity (e.g., Laird and Ware, 1982), cluster-robust standard errors (section 8.3, White, 1984), and bootstrapping based on resampling the clusters instead of the units. However, clustering has not received much attention in

Table 2.2: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = .2$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	0.45	0.05	0.91	0.13	0.85	0.12	0.93	0.13	0.58	0.05	0.94	0.13	0.84	0.12	0.91	0.12
		500	0.47	0.03	0.90	0.08	0.83	0.08	0.90	0.08	0.55	0.03	0.90	0.08	0.87	0.08	0.91	0.08
		1000	0.56	0.02	0.89	0.05	0.86	0.05	0.91	0.06	0.53	0.02	0.88	0.05	0.85	0.05	0.86	0.05
	0.3	200	0.45	0.04	0.93	0.10	0.89	0.09	0.95	0.10	0.45	0.04	0.93	0.10	0.88	0.10	0.94	0.10
		500	0.54	0.02	0.92	0.06	0.87	0.06	0.90	0.06	0.54	0.02	0.92	0.06	0.88	0.06	0.91	0.06
		1000	0.44	0.02	0.84	0.04	0.83	0.04	0.85	0.04	0.47	0.02	0.86	0.04	0.79	0.04	0.87	0.04
	0.5	200	0.56	0.03	0.90	0.09	0.88	0.09	0.90	0.09	0.48	0.03	0.86	0.09	0.82	0.09	0.88	0.09
		500	0.49	0.02	0.94	0.06	0.89	0.05	0.94	0.06	0.43	0.02	0.84	0.06	0.78	0.06	0.89	0.06
		1000	0.46	0.02	0.94	0.04	0.90	0.04	0.92	0.04	0.51	0.02	0.94	0.04	0.89	0.04	0.90	0.04
	0.7	200	0.56	0.04	0.93	0.10	0.93	0.10	0.96	0.10	0.53	0.04	0.92	0.11	0.88	0.10	0.94	0.10
		500	0.48	0.02	0.90	0.06	0.92	0.06	0.91	0.06	0.50	0.02	0.92	0.06	0.88	0.06	0.93	0.06
		1000	0.54	0.02	0.91	0.04	0.82	0.04	0.93	0.04	0.53	0.02	0.90	0.04	0.87	0.04	0.94	0.04
	0.9	200	0.46	0.05	0.91	0.13	0.87	0.12	0.94	0.13	0.45	0.05	0.84	0.13	0.83	0.12	0.85	0.12
		500	0.45	0.03	0.89	0.08	0.83	0.07	0.89	0.08	0.49	0.03	0.91	0.08	0.84	0.07	0.93	0.08
		1000	0.53	0.02	0.84	0.06	0.81	0.06	0.86	0.06	0.53	0.02	0.89	0.06	0.86	0.05	0.92	0.06
2	0.1	200	0.47	0.05	0.89	0.13	0.83	0.12	0.92	0.14	0.50	0.05	0.91	0.11	0.79	0.10	0.92	0.12
		500	0.48	0.03	0.84	0.08	0.82	0.08	0.90	0.08	0.60	0.03	0.89	0.07	0.85	0.07	0.88	0.07
		1000	0.40	0.02	0.89	0.06	0.86	0.06	0.90	0.06	0.56	0.02	0.95	0.05	0.92	0.04	0.91	0.04
	0.3	200	0.37	0.04	0.87	0.10	0.85	0.09	0.88	0.10	0.56	0.03	0.92	0.08	0.84	0.07	0.91	0.08
		500	0.48	0.02	0.88	0.06	0.80	0.06	0.90	0.06	0.45	0.02	0.93	0.05	0.92	0.05	0.88	0.05
		1000	0.50	0.02	0.90	0.04	0.93	0.04	0.91	0.04	0.53	0.01	0.90	0.03	0.86	0.03	0.85	0.03
	0.5	200	0.41	0.03	0.89	0.09	0.83	0.09	0.89	0.09	0.41	0.03	0.93	0.08	0.85	0.08	0.86	0.08
		500	0.47	0.02	0.90	0.06	0.88	0.06	0.91	0.06	0.53	0.02	0.89	0.05	0.89	0.05	0.90	0.04
		1000	0.51	0.02	0.89	0.04	0.86	0.04	0.88	0.04	0.57	0.01	0.89	0.03	0.86	0.03	0.87	0.03
	0.7	200	0.53	0.04	0.90	0.10	0.87	0.10	0.89	0.10	0.45	0.03	0.90	0.08	0.85	0.08	0.87	0.08
		500	0.49	0.02	0.89	0.06	0.90	0.06	0.88	0.06	0.58	0.02	0.90	0.05	0.92	0.05	0.89	0.05
		1000	0.49	0.02	0.86	0.04	0.83	0.04	0.87	0.04	0.48	0.01	0.91	0.03	0.83	0.03	0.82	0.03
	0.9	200	0.53	0.05	0.89	0.13	0.85	0.13	0.94	0.13	0.53	0.05	0.91	0.11	0.82	0.10	0.93	0.11
		500	0.41	0.03	0.88	0.08	0.85	0.08	0.90	0.08	0.57	0.03	0.89	0.07	0.85	0.06	0.89	0.07
		1000	0.51	0.02	0.84	0.06	0.79	0.05	0.86	0.06	0.46	0.02	0.92	0.05	0.85	0.04	0.89	0.04

the quantile regression literature, exceptions being recent work by Geraci and Bottai (2007, 2014) who develop multilevel quantile regression models and by Reich et al. (2010, 2011) who use Bayesian nonparametric and semiparametric models. Here we do not attempt to model the between-cluster heterogeneity. Instead, we ignore the clustering for point estimation and modify the Bayesian IJ standard errors to take the clustering into account, which is similar in purpose to recent work by Parente and Silva (2016) and Hagemann (2017) on cluster-robust standard errors for quantile regression.

Table 2.3: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = .5$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	0.68	0.09	0.84	0.13	0.80	0.12	0.89	0.14	0.73	0.08	0.89	0.13	0.82	0.11	0.87	0.13
		500	0.64	0.05	0.86	0.08	0.83	0.07	0.86	0.08	0.77	0.05	0.92	0.08	0.88	0.07	0.93	0.08
		1000	0.72	0.04	0.84	0.06	0.83	0.05	0.87	0.06	0.68	0.04	0.86	0.05	0.79	0.05	0.87	0.05
	0.3	200	0.70	0.06	0.92	0.10	0.90	0.09	0.92	0.10	0.75	0.06	0.91	0.10	0.85	0.09	0.89	0.09
		500	0.75	0.04	0.92	0.06	0.91	0.06	0.92	0.06	0.72	0.04	0.93	0.06	0.90	0.06	0.91	0.06
		1000	0.63	0.03	0.87	0.04	0.83	0.04	0.86	0.04	0.66	0.03	0.89	0.04	0.90	0.04	0.91	0.04
	0.5	200	0.69	0.06	0.89	0.09	0.88	0.09	0.92	0.09	0.63	0.06	0.86	0.09	0.85	0.09	0.87	0.09
		500	0.72	0.04	0.89	0.06	0.87	0.05	0.91	0.06	0.72	0.04	0.92	0.06	0.90	0.05	0.92	0.06
		1000	0.69	0.03	0.90	0.04	0.86	0.04	0.90	0.04	0.69	0.02	0.88	0.04	0.88	0.04	0.92	0.04
	0.7	200	0.66	0.06	0.86	0.10	0.85	0.09	0.89	0.10	0.72	0.06	0.89	0.10	0.86	0.10	0.89	0.09
		500	0.70	0.04	0.94	0.06	0.93	0.06	0.95	0.06	0.70	0.04	0.89	0.06	0.88	0.06	0.86	0.06
		1000	0.73	0.03	0.92	0.04	0.90	0.04	0.92	0.04	0.73	0.03	0.87	0.04	0.86	0.04	0.89	0.04
	0.9	200	0.74	0.09	0.90	0.13	0.86	0.12	0.92	0.13	0.69	0.08	0.86	0.13	0.80	0.11	0.88	0.12
		500	0.77	0.05	0.91	0.08	0.88	0.07	0.91	0.08	0.68	0.05	0.84	0.08	0.81	0.07	0.88	0.08
		1000	0.70	0.04	0.90	0.05	0.85	0.05	0.92	0.06	0.79	0.04	0.94	0.06	0.87	0.05	0.94	0.05
2	0.1	200	0.70	0.09	0.89	0.13	0.85	0.12	0.95	0.14	0.75	0.08	0.89	0.12	0.84	0.10	0.85	0.12
		500	0.65	0.05	0.90	0.08	0.83	0.08	0.88	0.08	0.75	0.05	0.86	0.07	0.82	0.06	0.90	0.07
		1000	0.66	0.04	0.87	0.06	0.83	0.06	0.90	0.06	0.76	0.03	0.88	0.05	0.86	0.04	0.88	0.05
	0.3	200	0.73	0.06	0.91	0.10	0.89	0.10	0.92	0.10	0.68	0.05	0.82	0.09	0.80	0.08	0.84	0.08
		500	0.71	0.04	0.90	0.06	0.88	0.06	0.89	0.06	0.74	0.03	0.95	0.05	0.90	0.05	0.90	0.05
		1000	0.66	0.03	0.84	0.04	0.82	0.04	0.87	0.04	0.72	0.02	0.90	0.03	0.89	0.03	0.89	0.03
	0.5	200	0.68	0.06	0.88	0.10	0.87	0.09	0.90	0.10	0.76	0.05	0.96	0.08	0.93	0.08	0.94	0.07
		500	0.63	0.04	0.88	0.06	0.87	0.06	0.91	0.06	0.80	0.03	0.95	0.05	0.91	0.05	0.89	0.05
		1000	0.67	0.03	0.87	0.04	0.86	0.04	0.87	0.04	0.74	0.02	0.93	0.03	0.89	0.03	0.90	0.03
	0.7	200	0.66	0.06	0.87	0.10	0.86	0.10	0.88	0.10	0.70	0.05	0.89	0.09	0.87	0.08	0.87	0.08
		500	0.75	0.04	0.93	0.06	0.92	0.06	0.95	0.06	0.68	0.03	0.88	0.05	0.82	0.05	0.88	0.05
		1000	0.68	0.03	0.89	0.04	0.88	0.04	0.89	0.04	0.72	0.02	0.92	0.04	0.89	0.03	0.87	0.03
	0.9	200	0.72	0.09	0.88	0.13	0.87	0.13	0.92	0.14	0.82	0.08	0.93	0.12	0.88	0.10	0.93	0.11
		500	0.76	0.05	0.91	0.08	0.89	0.07	0.92	0.08	0.85	0.05	0.96	0.07	0.92	0.06	0.93	0.06
		1000	0.78	0.04	0.89	0.06	0.87	0.05	0.91	0.06	0.75	0.03	0.84	0.04	0.81	0.04	0.89	0.04

2.4.1 Empirical coverage for Bayesian quantile regression for clustered data

To make our simulation results comparable to the results presented by Hagemann (2017) for cluster-robust inference for quantile regression, we follow the simulation setup in section 4 of Hagemann (2017). The data-generating model is

$$y_{ij} = \frac{1}{10}u_{ij} + x_{ij} + x_{ij}^2 u_{ij}.$$

Similar to the original paper, x_{ij} is the sum of $\sqrt{\rho z_j}$ and $\sqrt{1-\rho}\epsilon_{ij}$ with $\rho \in [0, 1)$, $z_j \sim N(0, 1)$ and, independently, $\epsilon_{ij} \sim N(0, 1)$, so that x_{ij} has an intraclass correlation equal to ρ and a variance of 1. The variable u_{ij} is simulated from a Gaussian distribution $N(0, \frac{1}{3})$.

Table 2.4: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 1$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	0.91	0.12	0.91	0.13	0.89	0.12	0.93	0.14	0.91	0.12	0.90	0.13	0.88	0.11	0.90	0.13
		500	0.90	0.07	0.90	0.08	0.88	0.07	0.90	0.08	0.92	0.08	0.92	0.08	0.85	0.07	0.91	0.08
		1000	0.87	0.05	0.87	0.06	0.85	0.05	0.88	0.06	0.84	0.05	0.85	0.06	0.81	0.05	0.84	0.06
	0.3	200	0.87	0.09	0.93	0.10	0.89	0.09	0.90	0.10	0.89	0.09	0.93	0.10	0.88	0.09	0.93	0.10
		500	0.86	0.05	0.89	0.06	0.86	0.06	0.92	0.06	0.87	0.05	0.89	0.06	0.86	0.06	0.89	0.06
		1000	0.83	0.04	0.85	0.04	0.83	0.04	0.86	0.04	0.81	0.04	0.85	0.04	0.82	0.04	0.88	0.04
	0.5	200	0.90	0.08	0.92	0.09	0.93	0.09	0.91	0.09	0.81	0.08	0.91	0.10	0.88	0.09	0.86	0.09
		500	0.84	0.05	0.86	0.06	0.85	0.06	0.89	0.06	0.88	0.05	0.89	0.06	0.89	0.05	0.90	0.06
		1000	0.88	0.04	0.93	0.04	0.93	0.04	0.95	0.04	0.88	0.04	0.90	0.04	0.89	0.04	0.90	0.04
	0.7	200	0.88	0.09	0.89	0.10	0.88	0.09	0.89	0.10	0.80	0.08	0.85	0.10	0.80	0.09	0.87	0.10
		500	0.82	0.05	0.88	0.06	0.85	0.06	0.86	0.06	0.87	0.05	0.90	0.06	0.86	0.06	0.89	0.06
		1000	0.91	0.04	0.93	0.04	0.92	0.04	0.93	0.04	0.86	0.04	0.91	0.04	0.87	0.04	0.90	0.04
	0.9	200	0.86	0.12	0.84	0.13	0.79	0.11	0.86	0.13	0.90	0.12	0.89	0.13	0.87	0.11	0.88	0.13
		500	0.92	0.08	0.93	0.08	0.91	0.07	0.92	0.08	0.89	0.08	0.91	0.08	0.87	0.07	0.91	0.08
		1000	0.89	0.05	0.90	0.05	0.88	0.05	0.90	0.06	0.91	0.05	0.88	0.06	0.86	0.05	0.90	0.05
2	0.1	200	0.89	0.12	0.87	0.13	0.84	0.12	0.91	0.14	0.92	0.11	0.92	0.11	0.79	0.10	0.89	0.11
		500	0.92	0.08	0.88	0.08	0.86	0.07	0.92	0.08	0.93	0.07	0.93	0.07	0.88	0.06	0.90	0.06
		1000	0.89	0.05	0.89	0.06	0.88	0.06	0.91	0.06	0.89	0.05	0.87	0.05	0.85	0.04	0.89	0.04
	0.3	200	0.82	0.08	0.85	0.10	0.84	0.09	0.88	0.10	0.89	0.08	0.91	0.08	0.88	0.07	0.87	0.08
		500	0.87	0.05	0.94	0.06	0.92	0.06	0.91	0.06	0.88	0.05	0.89	0.05	0.87	0.05	0.89	0.05
		1000	0.84	0.04	0.88	0.04	0.88	0.04	0.87	0.04	0.93	0.03	0.93	0.03	0.91	0.03	0.89	0.03
	0.5	200	0.78	0.08	0.91	0.09	0.86	0.09	0.91	0.10	0.92	0.07	0.94	0.08	0.90	0.07	0.93	0.08
		500	0.81	0.05	0.88	0.06	0.86	0.06	0.90	0.06	0.87	0.04	0.88	0.05	0.90	0.05	0.84	0.04
		1000	0.80	0.04	0.83	0.04	0.85	0.04	0.88	0.04	0.92	0.03	0.92	0.03	0.90	0.03	0.86	0.03
	0.7	200	0.78	0.09	0.86	0.10	0.82	0.09	0.87	0.10	0.85	0.08	0.88	0.09	0.84	0.08	0.85	0.08
		500	0.86	0.05	0.89	0.06	0.88	0.06	0.93	0.06	0.90	0.05	0.88	0.05	0.87	0.05	0.91	0.05
		1000	0.88	0.04	0.92	0.04	0.91	0.04	0.92	0.04	0.96	0.03	0.95	0.03	0.93	0.03	0.95	0.03
	0.9	200	0.87	0.12	0.84	0.12	0.79	0.11	0.90	0.13	0.93	0.11	0.93	0.12	0.87	0.10	0.93	0.10
		500	0.85	0.07	0.88	0.08	0.84	0.07	0.90	0.08	0.94	0.07	0.91	0.07	0.88	0.06	0.91	0.06
		1000	0.87	0.05	0.92	0.06	0.88	0.05	0.92	0.06	0.89	0.05	0.87	0.04	0.84	0.04	0.87	0.04

The outcome model above can be written as a quantile regression model with covariates x_{ij} and x_{ij}^2 ,

$$Q_{\tau}(y_{ij}|x_{ij}) = \alpha(\tau) + \beta_1(\tau)x_{ij} + \beta_2(\tau)x_{ij}^2$$

with model-implied parameters $\alpha(\tau) = \frac{\Phi^{-1}(\tau)}{\sqrt{300}}$, $\beta_1(\tau) = 1$, and $\beta_2(\tau) = \frac{\Phi^{-1}(\tau)}{\sqrt{3}}$, where Φ is the standard normal CDF.

We consider two cluster sizes, $I \in \{10, 30\}$, three choices of number of clusters, $J \in \{10, 50, 100\}$, and two choices of intraclass correlation, $\rho \in \{.3, .8\}$. We vary the scale parameter in ALD to be $\sigma \in \{.1, .5, 1, 5\}$ and obtain model-based credible intervals (denoted as "ALD") and clustered Bayesian IJ intervals (denoted as "IJ") for five different quantiles, $\tau \in \{.1, .3, .5, .7, .9\}$.

The empirical 90% coverage is estimated for 100 replications for each of the simulation conditions. For comparison, we also include the current state-of-the-art clustered gradient

Table 2.5: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 2$.

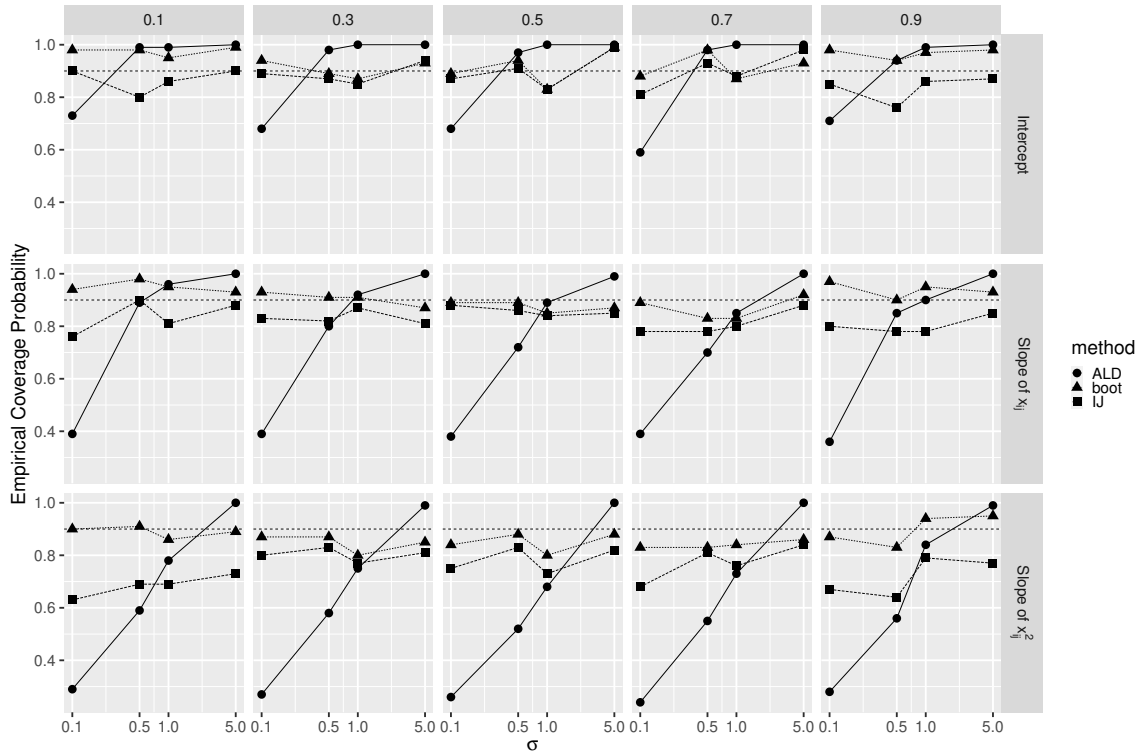
Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	0.98	0.17	0.91	0.12	0.86	0.11	0.94	0.14	1.00	0.17	0.93	0.13	0.90	0.10	0.93	0.13
		500	0.98	0.11	0.93	0.08	0.90	0.07	0.94	0.08	0.98	0.11	0.92	0.08	0.91	0.07	0.89	0.08
		1000	0.97	0.08	0.88	0.06	0.86	0.05	0.90	0.06	0.99	0.08	0.90	0.06	0.90	0.05	0.92	0.06
	0.3	200	0.96	0.12	0.89	0.10	0.87	0.09	0.93	0.10	0.99	0.12	0.97	0.10	0.90	0.09	0.96	0.10
		500	0.98	0.08	0.91	0.06	0.88	0.06	0.89	0.06	0.94	0.08	0.89	0.06	0.89	0.06	0.90	0.06
		1000	0.97	0.05	0.82	0.04	0.81	0.04	0.83	0.04	0.93	0.05	0.89	0.04	0.88	0.04	0.89	0.04
	0.5	200	1.00	0.11	0.92	0.09	0.87	0.09	0.92	0.09	0.96	0.11	0.91	0.09	0.86	0.08	0.90	0.09
		500	0.97	0.07	0.92	0.06	0.90	0.06	0.93	0.06	0.95	0.07	0.86	0.06	0.82	0.05	0.87	0.06
		1000	0.96	0.05	0.89	0.04	0.86	0.04	0.87	0.04	0.95	0.05	0.89	0.04	0.83	0.04	0.90	0.04
	0.7	200	0.97	0.12	0.91	0.10	0.91	0.09	0.92	0.10	0.96	0.12	0.90	0.10	0.85	0.09	0.89	0.10
		500	0.97	0.08	0.91	0.06	0.86	0.06	0.93	0.06	0.95	0.08	0.89	0.06	0.83	0.06	0.86	0.06
		1000	0.96	0.05	0.86	0.04	0.85	0.04	0.87	0.04	0.93	0.05	0.90	0.04	0.86	0.04	0.89	0.04
	0.9	200	0.97	0.17	0.91	0.13	0.87	0.11	0.93	0.14	0.98	0.17	0.90	0.13	0.82	0.10	0.90	0.12
		500	0.96	0.11	0.86	0.08	0.82	0.07	0.91	0.08	0.96	0.11	0.88	0.08	0.85	0.07	0.88	0.08
		1000	0.98	0.08	0.90	0.06	0.86	0.05	0.88	0.06	0.97	0.08	0.88	0.06	0.87	0.05	0.91	0.05
2	0.1	200	0.98	0.17	0.90	0.13	0.86	0.11	0.90	0.14	0.98	0.16	0.92	0.11	0.86	0.09	0.91	0.11
		500	0.99	0.11	0.94	0.08	0.92	0.08	0.92	0.08	0.98	0.10	0.95	0.07	0.93	0.06	0.93	0.06
		1000	0.99	0.08	0.97	0.06	0.95	0.05	0.97	0.06	0.99	0.07	0.93	0.05	0.91	0.04	0.90	0.04
	0.3	200	0.98	0.12	0.93	0.10	0.93	0.09	0.96	0.10	0.98	0.11	0.89	0.09	0.88	0.08	0.93	0.08
		500	0.93	0.08	0.87	0.06	0.85	0.06	0.89	0.06	0.97	0.07	0.91	0.05	0.91	0.05	0.88	0.05
		1000	0.97	0.05	0.89	0.04	0.88	0.04	0.89	0.04	0.95	0.05	0.88	0.03	0.88	0.03	0.89	0.03
	0.5	200	0.94	0.11	0.89	0.10	0.83	0.09	0.85	0.09	0.99	0.10	0.91	0.08	0.89	0.07	0.86	0.07
		500	0.96	0.07	0.93	0.06	0.92	0.06	0.93	0.06	0.97	0.07	0.91	0.05	0.90	0.05	0.90	0.05
		1000	0.96	0.05	0.89	0.04	0.87	0.04	0.90	0.04	0.96	0.04	0.89	0.03	0.88	0.03	0.86	0.03
	0.7	200	0.98	0.12	0.96	0.10	0.95	0.09	0.96	0.10	0.98	0.11	0.95	0.08	0.92	0.08	0.92	0.08
		500	0.96	0.08	0.87	0.06	0.84	0.06	0.87	0.06	0.99	0.07	0.92	0.05	0.90	0.05	0.87	0.05
		1000	0.93	0.05	0.85	0.04	0.84	0.04	0.85	0.04	0.97	0.05	0.88	0.03	0.89	0.03	0.89	0.03
	0.9	200	0.96	0.18	0.89	0.14	0.84	0.12	0.94	0.15	1.00	0.17	0.94	0.12	0.88	0.10	0.92	0.12
		500	0.97	0.11	0.81	0.08	0.80	0.08	0.83	0.08	1.00	0.10	0.97	0.07	0.91	0.06	0.92	0.07
		1000	0.97	0.08	0.93	0.06	0.91	0.05	0.95	0.06	1.00	0.07	0.94	0.05	0.92	0.04	0.94	0.04

bootstrap (denoted “boot”) defined in (Hagemann, 2017) for frequentist quantile regression.

The simulation is carried out in R (R Core Team, 2021) as before. “ALD” and “IJ” variance estimates are obtained using Stan (Carpenter et al., 2017) while “boot” variance estimates (Koenker, 2009) are obtained using the `quantreg` R package (Koenker, 2009).

2.4.2 Results

The empirical coverage results when Bayes posterior mean estimates used for constructing intervals are similar to what is presented for the unclustered case in Section 2.3. That is, the under-coverage happens for extreme quantiles (i.e., $\tau = .1/.9$) when the scale parameter is large (i.e., $\sigma = 5$) for the slope parameter (not shown), due to the bias introduced by the Bayesian point estimator, similarly as before for the unclustered case. Instead, we plot



$I = 10, J = 10, \rho = .8$

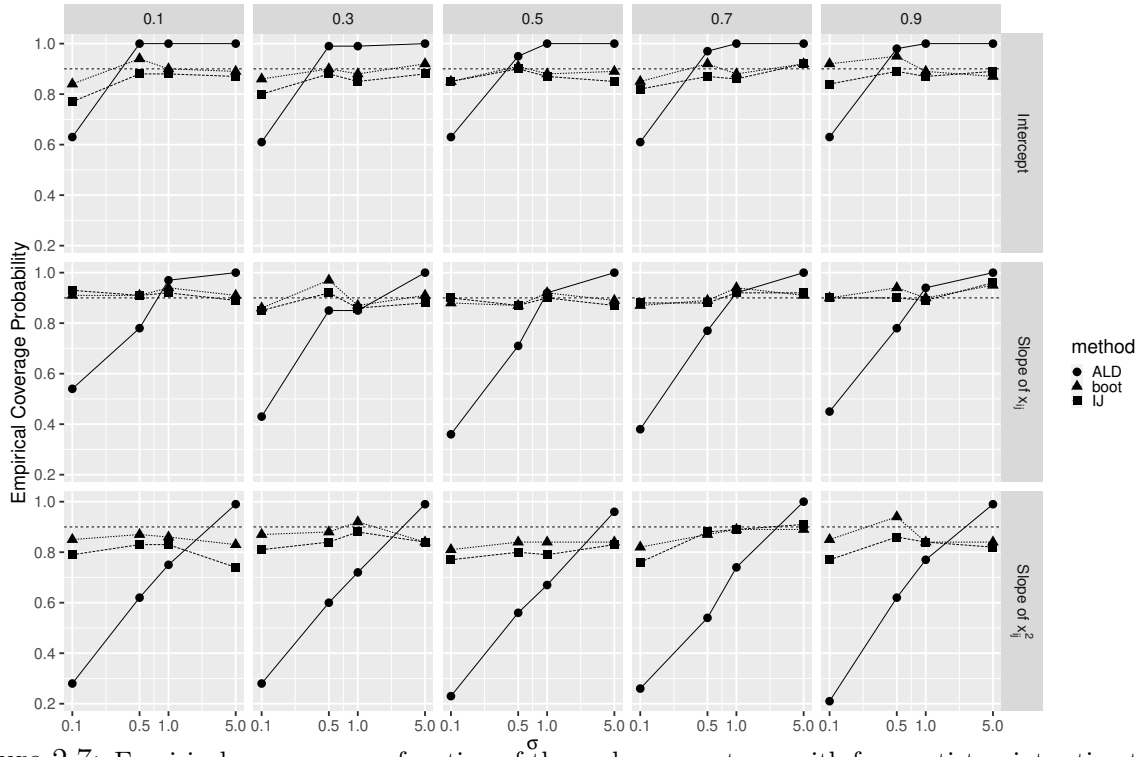
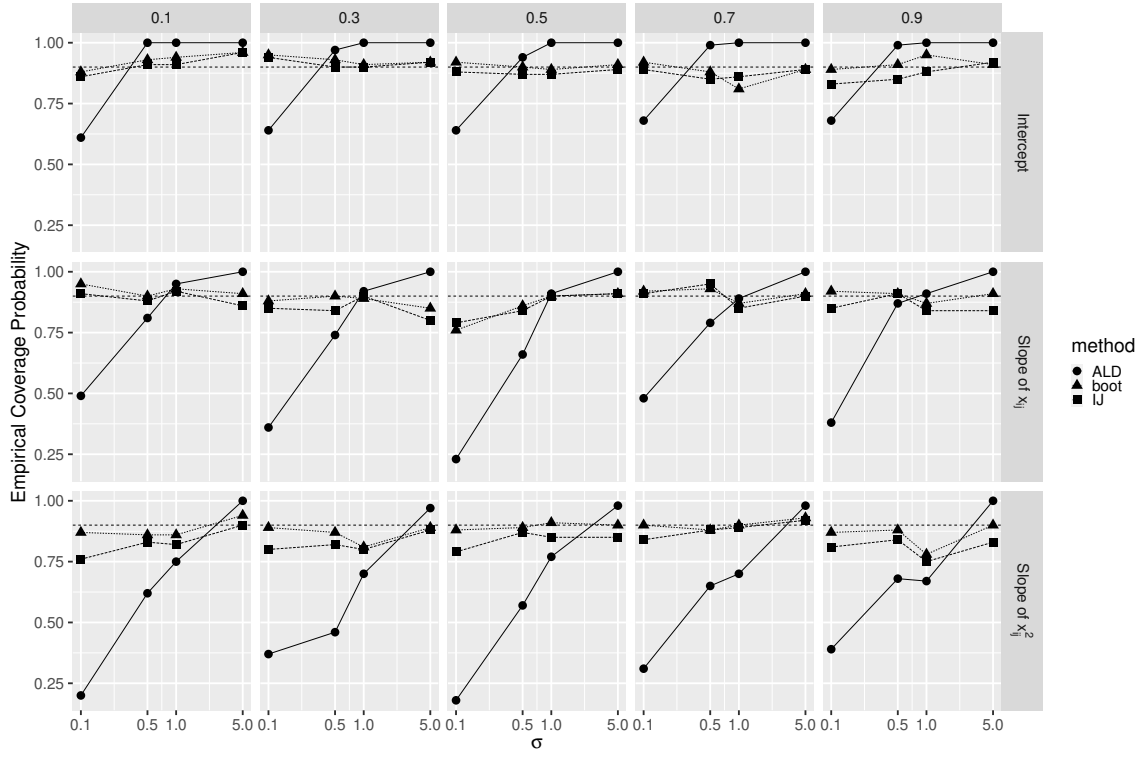


Figure 2.7: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 10, J = 10$.



$I = 10, J = 50, \rho = .8$

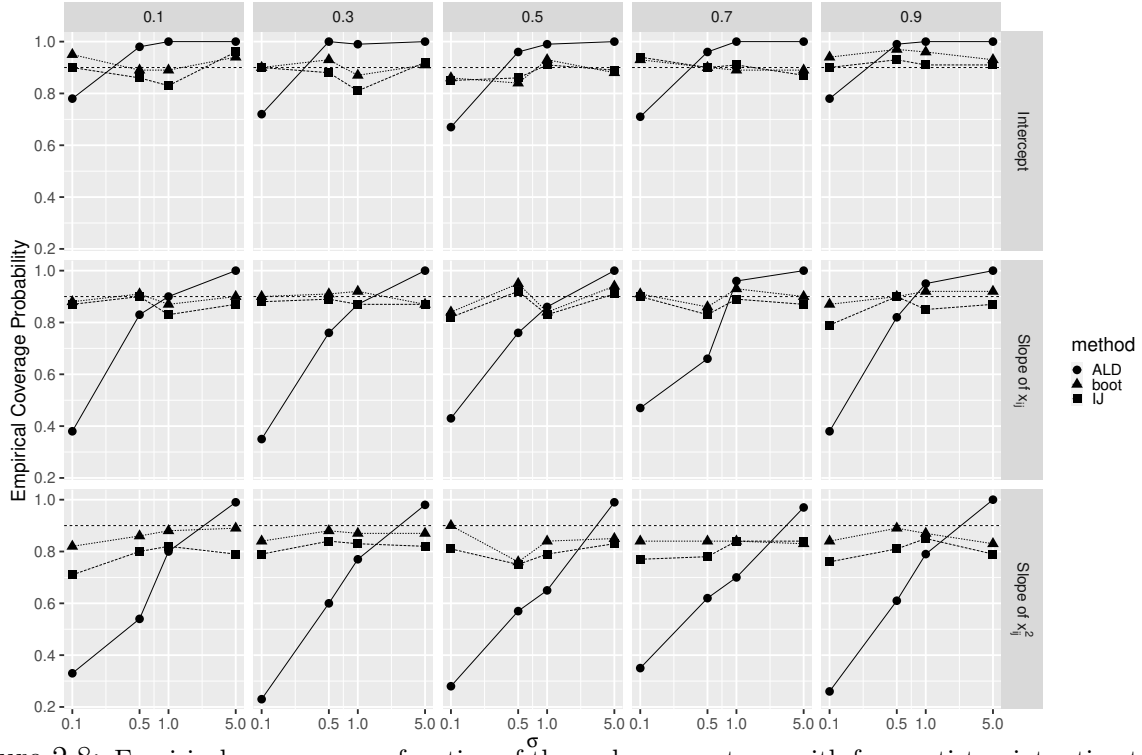
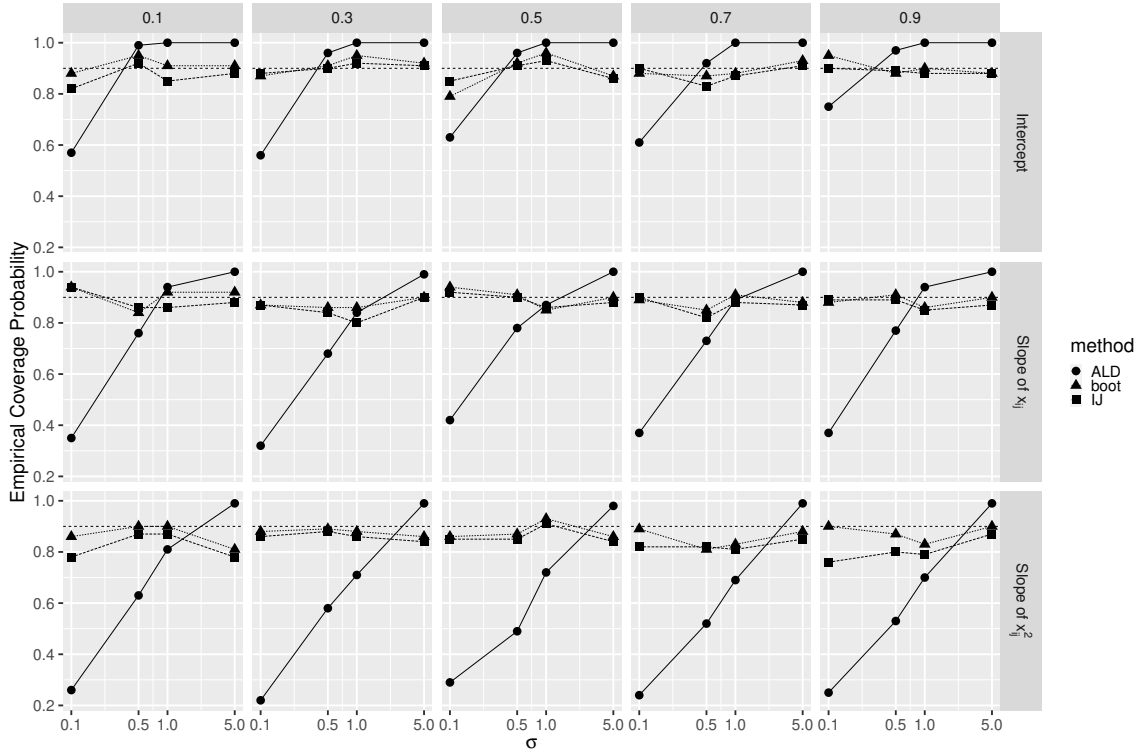


Figure 2.8: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 10, J = 50$.



$I = 10, J = 100, \rho = .8$

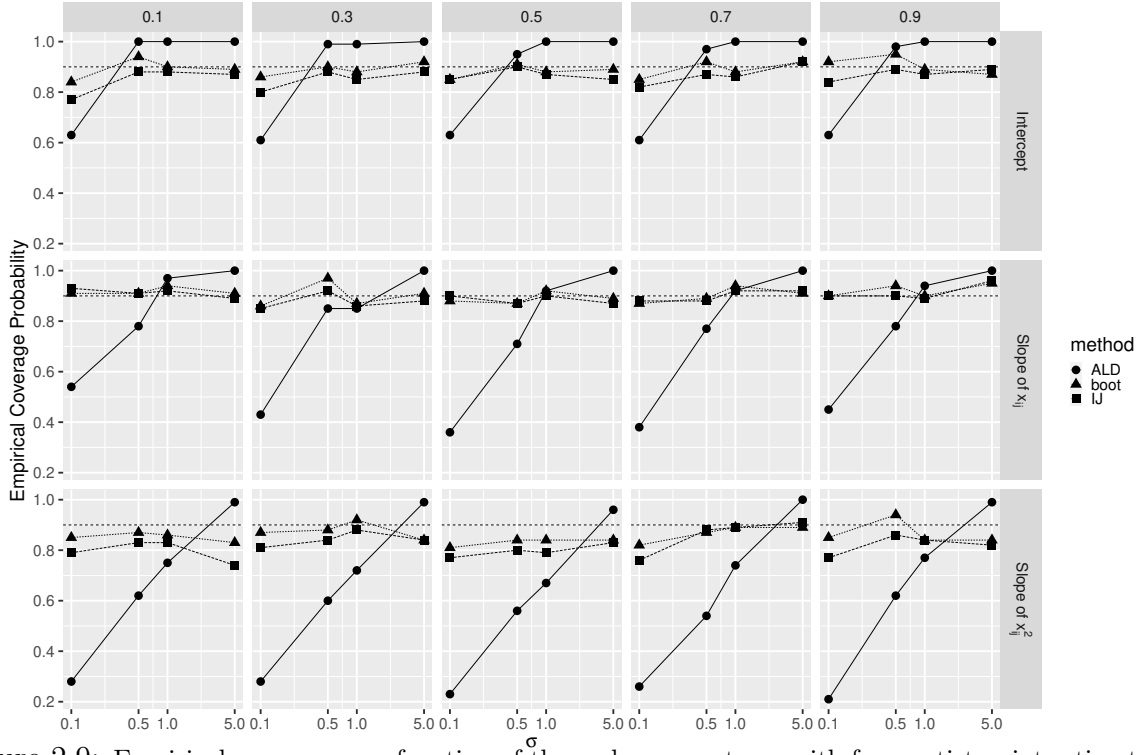
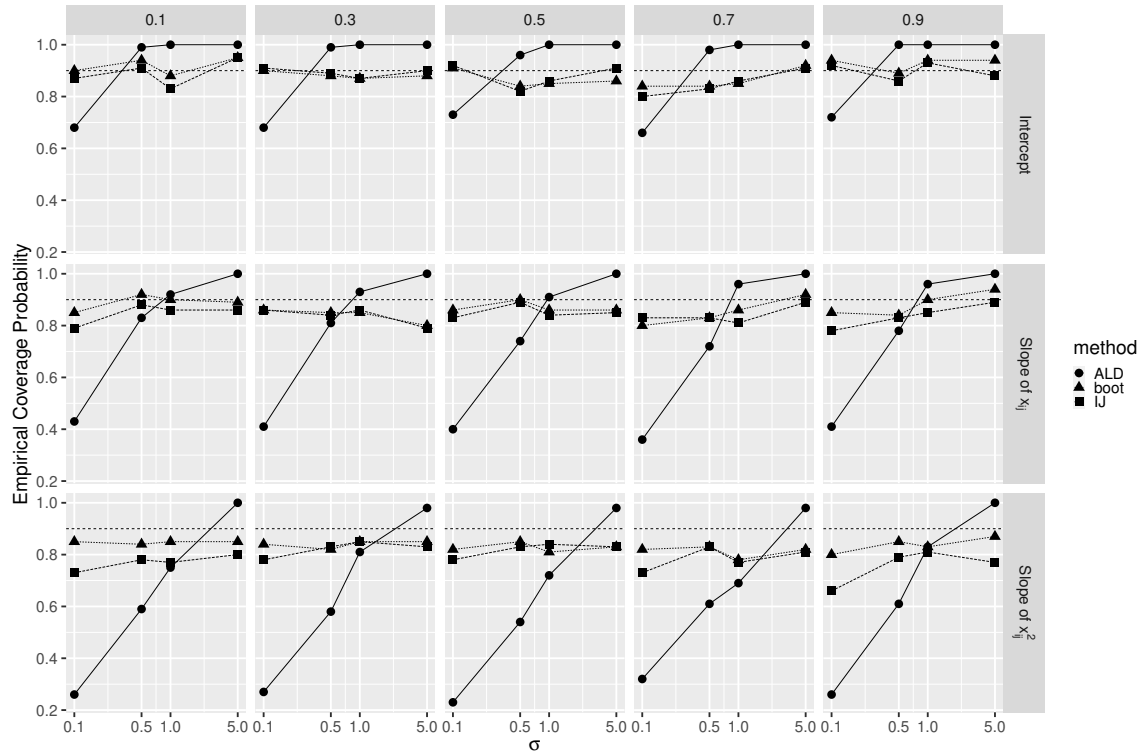


Figure 2.9: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 10, J = 100$.



$I = 30, J = 10, \rho = .8$

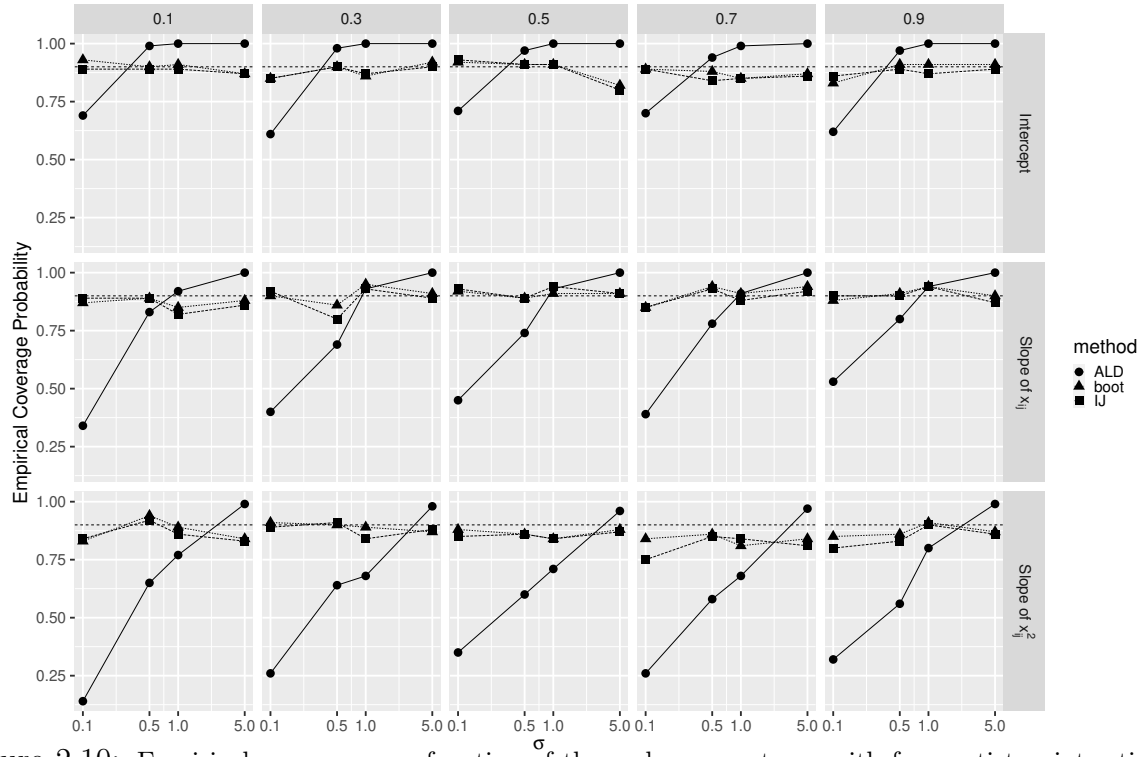
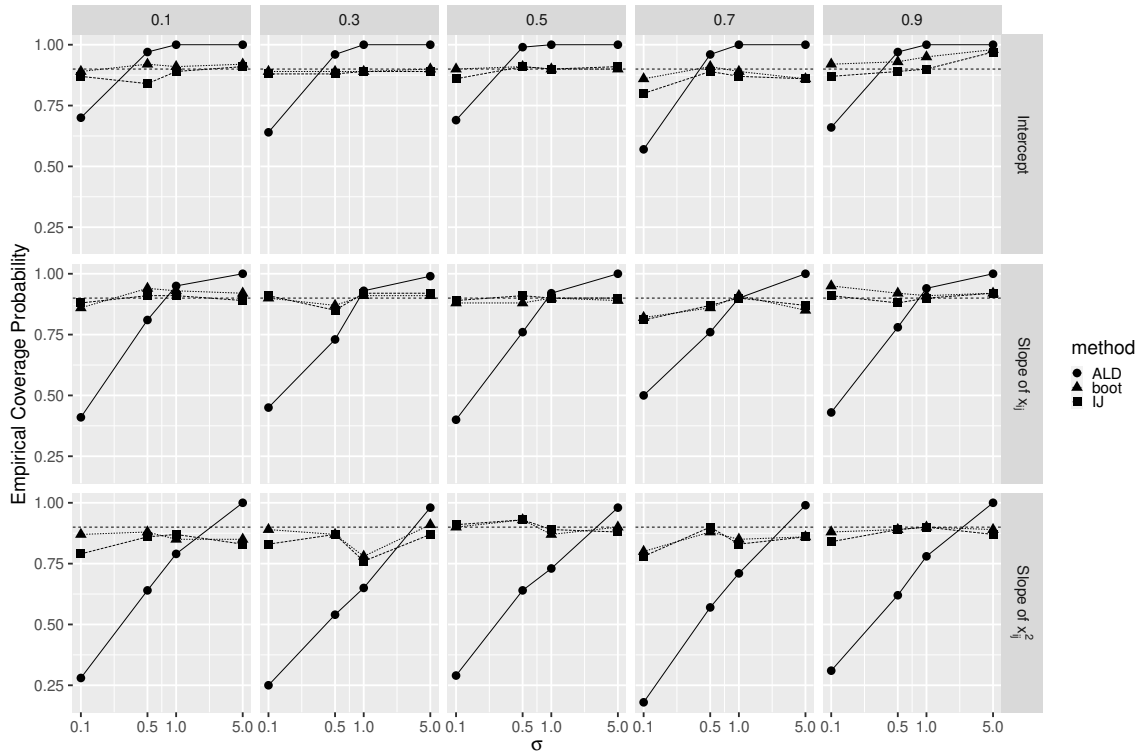


Figure 2.10: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 30, J = 10$.



$I = 30, J = 50, \rho = .8$

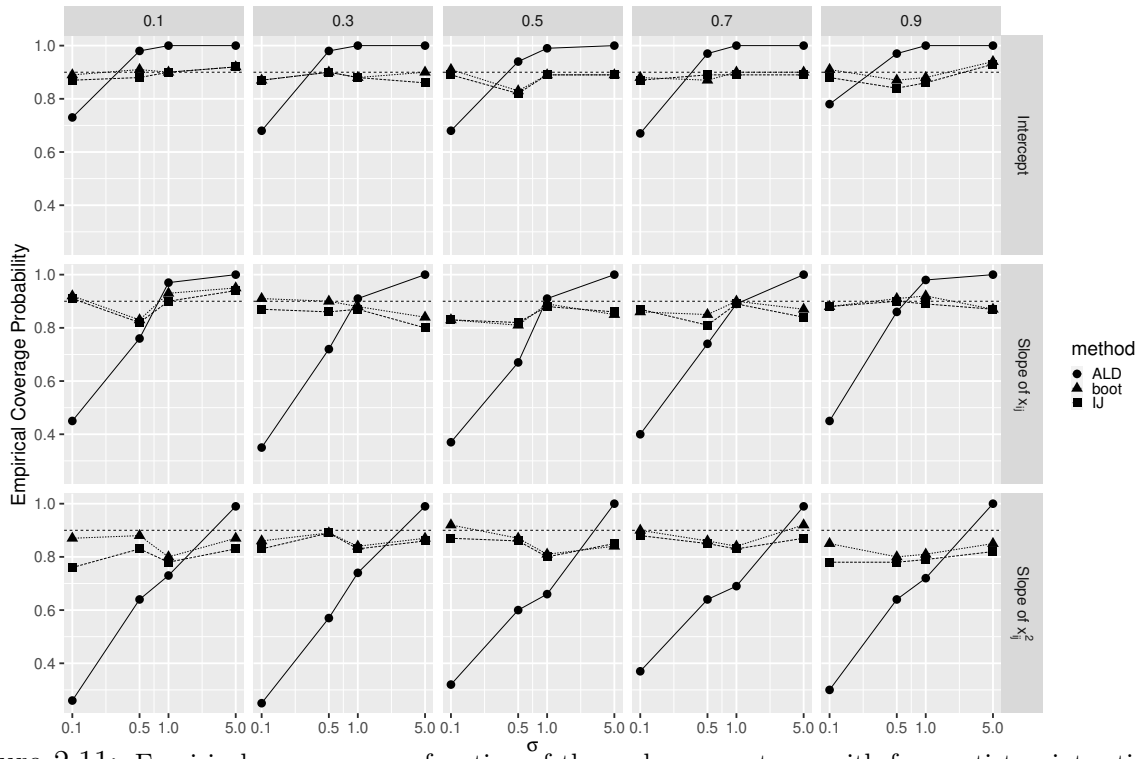
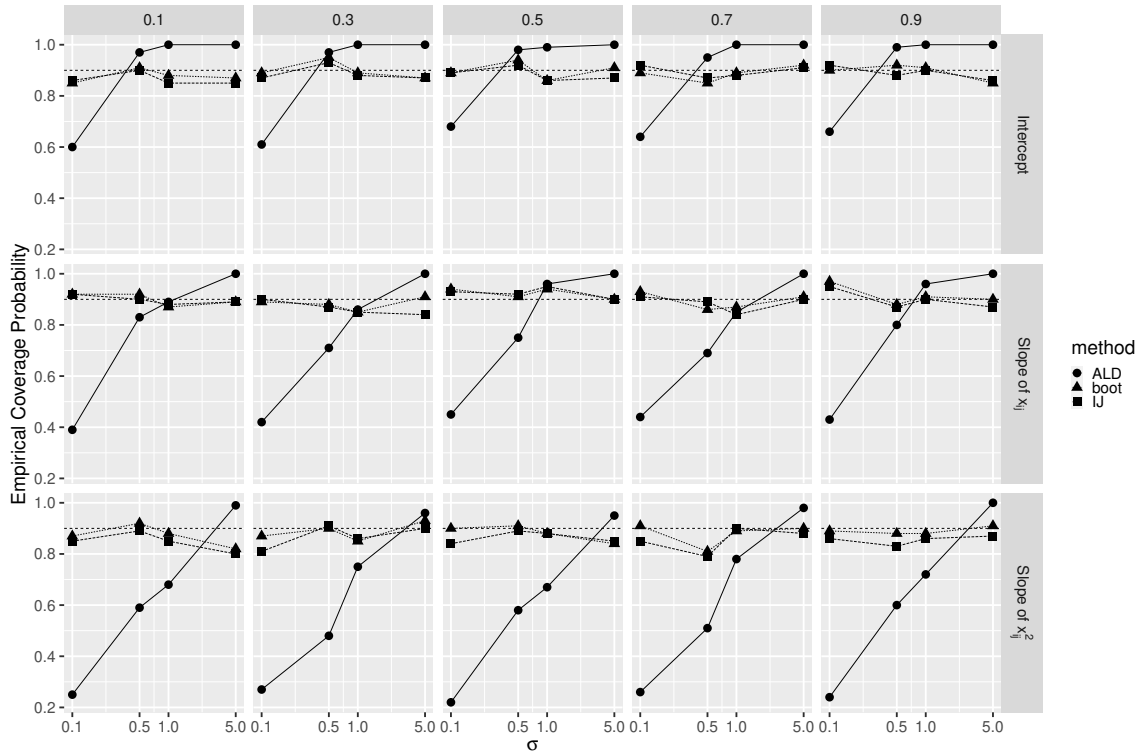


Figure 2.11: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 30, J = 50$.



$I = 30, J = 100, \rho = .8$

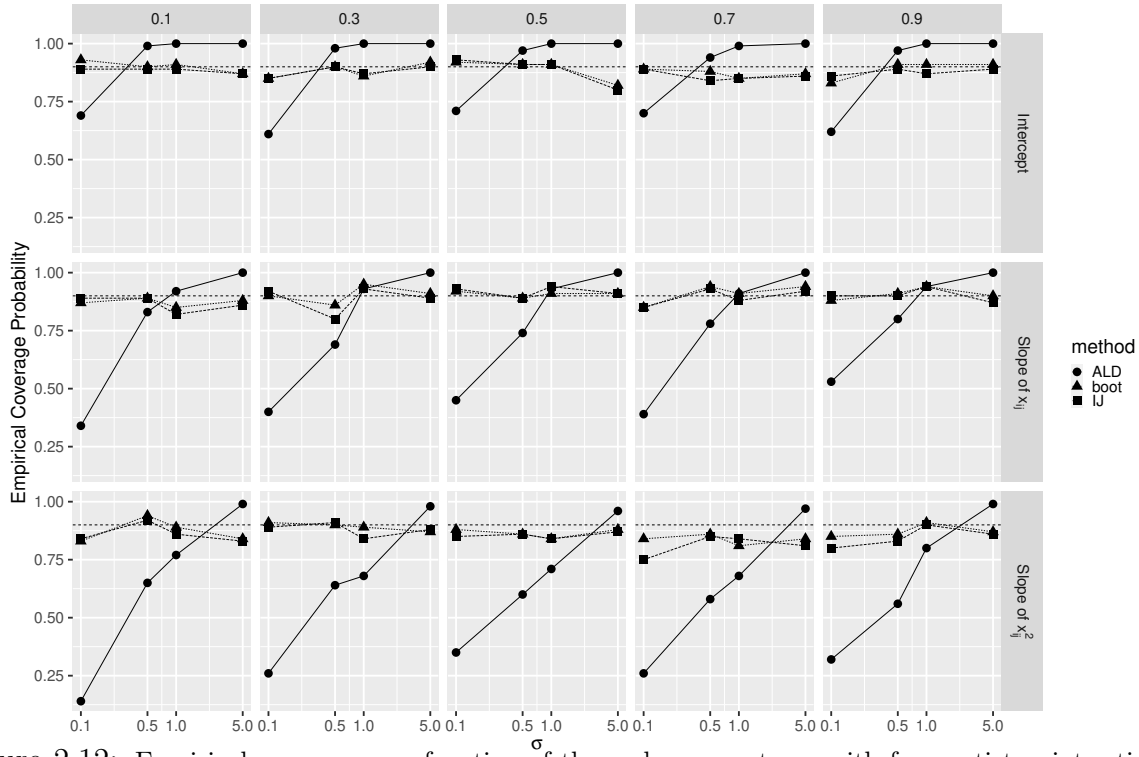


Figure 2.12: Empirical coverage as a function of the scale parameter σ with frequentist point estimates for different quantiles τ (columns) and different parameters (rows) for $I = 30, J = 100$.

Table 2.6: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 5$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	1.00	0.29	0.88	0.12	0.85	0.10	0.87	0.13	1.00	0.29	0.86	0.12	0.76	0.09	0.87	0.12
		500	1.00	0.18	0.85	0.08	0.79	0.07	0.88	0.08	1.00	0.17	0.90	0.08	0.85	0.07	0.88	0.08
		1000	1.00	0.12	0.91	0.06	0.83	0.05	0.94	0.06	1.00	0.12	0.87	0.06	0.81	0.05	0.86	0.06
	0.3	200	1.00	0.19	0.90	0.10	0.87	0.09	0.89	0.10	1.00	0.19	0.92	0.10	0.86	0.08	0.92	0.09
		500	1.00	0.12	0.92	0.06	0.90	0.06	0.93	0.06	0.99	0.12	0.91	0.06	0.90	0.05	0.92	0.06
		1000	1.00	0.08	0.89	0.04	0.88	0.04	0.92	0.04	0.99	0.08	0.90	0.04	0.88	0.04	0.90	0.04
	0.5	200	1.00	0.18	0.89	0.09	0.84	0.08	0.89	0.09	1.00	0.18	0.90	0.10	0.86	0.08	0.91	0.09
		500	1.00	0.11	0.88	0.06	0.87	0.05	0.90	0.06	1.00	0.11	0.95	0.06	0.93	0.05	0.94	0.06
		1000	1.00	0.08	0.93	0.04	0.91	0.04	0.92	0.04	1.00	0.08	0.84	0.04	0.85	0.04	0.87	0.04
	0.7	200	1.00	0.19	0.86	0.10	0.84	0.08	0.89	0.10	1.00	0.19	0.91	0.10	0.87	0.08	0.93	0.09
		500	1.00	0.12	0.91	0.06	0.89	0.06	0.93	0.06	1.00	0.12	0.91	0.06	0.90	0.05	0.92	0.06
		1000	1.00	0.08	0.84	0.04	0.83	0.04	0.84	0.04	1.00	0.09	0.88	0.04	0.88	0.04	0.87	0.04
	0.9	200	1.00	0.29	0.91	0.13	0.81	0.11	0.93	0.14	1.00	0.29	0.89	0.13	0.81	0.10	0.92	0.13
		500	1.00	0.17	0.89	0.08	0.87	0.07	0.93	0.08	1.00	0.17	0.92	0.08	0.87	0.07	0.92	0.08
		1000	1.00	0.12	0.92	0.05	0.88	0.05	0.90	0.06	1.00	0.12	0.92	0.06	0.88	0.05	0.93	0.06
2	0.1	200	1.00	0.29	0.85	0.13	0.84	0.11	0.90	0.14	1.00	0.28	0.92	0.11	0.86	0.09	0.89	0.11
		500	0.99	0.17	0.94	0.08	0.88	0.07	0.93	0.08	1.00	0.16	0.92	0.06	0.91	0.06	0.89	0.07
		1000	1.00	0.12	0.88	0.06	0.82	0.05	0.88	0.06	1.00	0.11	0.92	0.04	0.92	0.04	0.92	0.04
	0.3	200	1.00	0.19	0.85	0.10	0.83	0.09	0.86	0.10	1.00	0.19	0.89	0.09	0.87	0.08	0.86	0.08
		500	1.00	0.12	0.90	0.06	0.88	0.06	0.91	0.06	1.00	0.11	0.90	0.05	0.89	0.05	0.88	0.05
		1000	1.00	0.09	0.90	0.04	0.89	0.04	0.91	0.04	1.00	0.08	0.90	0.03	0.89	0.03	0.90	0.03
	0.5	200	1.00	0.18	0.84	0.10	0.83	0.09	0.89	0.10	1.00	0.17	0.93	0.08	0.90	0.08	0.87	0.07
		500	1.00	0.11	0.86	0.06	0.88	0.06	0.88	0.06	1.00	0.10	0.90	0.05	0.87	0.05	0.85	0.05
		1000	1.00	0.08	0.89	0.04	0.87	0.04	0.91	0.04	1.00	0.07	0.89	0.03	0.89	0.03	0.90	0.03
	0.7	200	1.00	0.20	0.87	0.10	0.82	0.09	0.86	0.10	1.00	0.19	0.94	0.09	0.91	0.08	0.91	0.08
		500	1.00	0.12	0.87	0.06	0.86	0.06	0.89	0.06	1.00	0.11	0.91	0.05	0.89	0.05	0.86	0.05
		1000	1.00	0.08	0.95	0.04	0.94	0.04	0.95	0.04	1.00	0.08	0.90	0.03	0.86	0.03	0.88	0.03
	0.9	200	1.00	0.29	0.92	0.13	0.87	0.11	0.92	0.14	1.00	0.27	0.94	0.12	0.89	0.09	0.92	0.12
		500	1.00	0.18	0.89	0.08	0.86	0.07	0.93	0.08	1.00	0.16	0.91	0.07	0.92	0.06	0.91	0.07
		1000	1.00	0.12	0.93	0.06	0.89	0.05	0.94	0.06	1.00	0.11	0.97	0.05	0.92	0.04	0.94	0.04

the empirical coverage results of the three interval methods (i.e., ALD, IJ, boot) using the frequentist point estimates below for the two cluster sizes, three numbers of clusters, five quantiles, and four scale parameters in Figures 2.7 to 2.12, and summarise the corresponding numerical results in Tables 2.8 to 2.11, along with the empirical mean length (EML) of the intervals.

First, the ALD model-based intervals do not maintain the nominal coverage and are very sensitive to the scale parameter. This result is similar to the unclustered case for Bayesian ALD quantile regression as well. Specifically, across different simulation conditions, the ALD interval has severe undercoverage (as low as 10%) when the scale parameter is .1 and severe overcoverage (as high as 100%) when the scale parameter is 5. The empirical coverage probability increases as the scale parameter increases. This observation holds for all the parameters.

Table 2.7: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 10$.

Case	τ	N	Intercept								Slope							
			ALD		boot		IJ		nid		ALD		boot		IJ		nid	
			ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
1	0.1	200	1.00	0.46	0.91	0.13	0.84	0.10	0.94	0.13	1.00	0.44	0.88	0.13	0.77	0.09	0.87	0.12
		500	1.00	0.26	0.91	0.08	0.86	0.07	0.94	0.08	1.00	0.25	0.85	0.08	0.80	0.06	0.85	0.08
		1000	1.00	0.17	0.91	0.05	0.89	0.05	0.93	0.06	1.00	0.17	0.91	0.06	0.89	0.05	0.93	0.05
	0.3	200	1.00	0.28	0.88	0.10	0.87	0.08	0.91	0.10	1.00	0.29	0.91	0.10	0.83	0.08	0.90	0.09
		500	1.00	0.17	0.90	0.06	0.88	0.05	0.93	0.06	1.00	0.17	0.85	0.06	0.84	0.05	0.87	0.06
		1000	1.00	0.12	0.88	0.04	0.88	0.04	0.91	0.04	1.00	0.12	0.89	0.04	0.87	0.04	0.91	0.04
	0.5	200	1.00	0.26	0.91	0.09	0.87	0.08	0.91	0.09	1.00	0.27	0.87	0.09	0.77	0.08	0.83	0.09
		500	1.00	0.16	0.87	0.06	0.85	0.05	0.91	0.06	1.00	0.16	0.85	0.06	0.81	0.05	0.87	0.06
		1000	1.00	0.11	0.88	0.04	0.88	0.04	0.90	0.04	1.00	0.11	0.92	0.04	0.89	0.04	0.91	0.04
	0.7	200	1.00	0.28	0.84	0.10	0.78	0.08	0.87	0.10	1.00	0.29	0.95	0.10	0.90	0.08	0.96	0.10
		500	1.00	0.17	0.95	0.06	0.93	0.05	0.96	0.06	1.00	0.18	0.85	0.06	0.83	0.05	0.88	0.06
		1000	1.00	0.12	0.86	0.04	0.85	0.04	0.87	0.04	1.00	0.12	0.84	0.04	0.82	0.04	0.85	0.04
	0.9	200	1.00	0.46	0.89	0.13	0.81	0.10	0.90	0.13	1.00	0.43	0.91	0.13	0.81	0.09	0.89	0.12
		500	1.00	0.26	0.87	0.08	0.83	0.07	0.92	0.08	1.00	0.25	0.91	0.08	0.86	0.06	0.89	0.08
		1000	1.00	0.18	0.87	0.05	0.84	0.05	0.87	0.06	1.00	0.17	0.87	0.06	0.83	0.05	0.87	0.05
2	0.1	200	1.00	0.46	0.94	0.14	0.86	0.11	0.95	0.14	1.00	0.43	0.93	0.12	0.84	0.09	0.90	0.11
		500	1.00	0.26	0.96	0.08	0.92	0.07	0.97	0.08	1.00	0.24	0.92	0.07	0.84	0.06	0.89	0.06
		1000	1.00	0.18	0.89	0.06	0.86	0.05	0.91	0.06	1.00	0.16	0.88	0.05	0.81	0.04	0.82	0.04
	0.3	200	1.00	0.28	0.88	0.10	0.85	0.08	0.91	0.10	1.00	0.28	0.90	0.09	0.86	0.08	0.85	0.08
		500	1.00	0.17	0.95	0.06	0.91	0.06	0.96	0.06	1.00	0.17	0.94	0.05	0.90	0.05	0.86	0.05
		1000	1.00	0.12	0.95	0.04	0.92	0.04	0.93	0.04	1.00	0.11	0.91	0.03	0.89	0.03	0.87	0.03
	0.5	200	1.00	0.26	0.88	0.09	0.82	0.08	0.89	0.09	1.00	0.26	0.94	0.08	0.94	0.08	0.95	0.07
		500	1.00	0.16	0.89	0.06	0.86	0.05	0.89	0.06	1.00	0.15	0.93	0.05	0.93	0.05	0.88	0.05
		1000	1.00	0.11	0.91	0.04	0.87	0.04	0.91	0.04	1.00	0.10	0.85	0.03	0.83	0.03	0.86	0.03
	0.7	200	1.00	0.28	0.89	0.10	0.84	0.09	0.88	0.10	1.00	0.28	0.94	0.09	0.91	0.08	0.92	0.08
		500	1.00	0.17	0.91	0.06	0.88	0.06	0.90	0.06	1.00	0.16	0.90	0.05	0.92	0.05	0.88	0.05
		1000	1.00	0.12	0.89	0.04	0.87	0.04	0.88	0.04	1.00	0.11	0.87	0.03	0.88	0.03	0.88	0.03
	0.9	200	1.00	0.46	0.82	0.13	0.76	0.11	0.83	0.14	1.00	0.43	0.89	0.12	0.81	0.09	0.87	0.11
		500	1.00	0.26	0.87	0.08	0.81	0.07	0.88	0.08	1.00	0.24	0.91	0.07	0.87	0.06	0.87	0.07
		1000	1.00	0.17	0.88	0.06	0.87	0.05	0.92	0.06	1.00	0.16	0.94	0.04	0.92	0.04	0.92	0.04

Second, the bootstrap interval has good coverage across simulation conditions and performs fairly well even when the cluster size is small and the number of clusters is small as in Figure 2.7. Slight undercoverage or overcoverage is only seen for extreme quantiles (.1 and .9). These results are consistent with those presented by Hagemann (2017).

Third, as we can see in Figures 2.7 and 2.10, when the number of clusters is small (i.e., 10), the IJ intervals tend to undercover across different quantiles and especially for extreme quantiles (e.g., $\tau = .1$ or $.9$), because of the nature of this interval based on a large-sample approximation. However, the undercoverage does not (or much less severely) happen once we have a moderate number of clusters (i.e., $J = 50$ and $J = 100$) across different choices of the scale parameter, cluster size, and intraclass correlation for all quantiles considered. For non-extreme quantiles, the coverage depends less on the choice of scale parameter.

Fourth, IJ performs consistently better than the model-based ALD intervals across all

Table 2.8: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = .1$.

				Intercept						Slope of x_{ij}						Slope of x_{ij}^2					
				ALD		boot		IJ		ALD		boot		IJ		ALD		boot		IJ	
I	J	ρ	τ	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
10	10	0.3	0.1	0.73	0.03	0.98	0.24	0.90	0.05	0.39	0.05	0.94	0.53	0.76	0.16	0.29	0.05	0.90	0.50	0.63	0.18
			0.3	0.68	0.02	0.94	0.04	0.89	0.03	0.39	0.04	0.93	0.13	0.83	0.13	0.27	0.03	0.87	0.17	0.80	0.16
			0.5	0.68	0.02	0.89	0.03	0.87	0.03	0.38	0.03	0.89	0.10	0.88	0.11	0.26	0.03	0.84	0.14	0.75	0.15
			0.7	0.59	0.02	0.88	0.04	0.81	0.04	0.39	0.04	0.89	0.12	0.78	0.12	0.24	0.03	0.83	0.16	0.68	0.15
		0.8	0.9	0.71	0.03	0.98	1.59	0.85	0.05	0.36	0.05	0.97	4.04	0.80	0.19	0.28	0.05	0.87	3.58	0.67	0.20
			0.1	0.75	0.03	0.98	0.29	0.80	0.05	0.42	0.05	0.90	0.57	0.78	0.12	0.26	0.05	0.84	0.45	0.55	0.13
			0.3	0.76	0.02	0.89	0.03	0.86	0.03	0.47	0.04	0.81	0.10	0.81	0.10	0.28	0.03	0.72	0.12	0.62	0.11
			0.5	0.67	0.02	0.92	0.03	0.87	0.03	0.41	0.03	0.83	0.09	0.80	0.10	0.32	0.03	0.81	0.13	0.74	0.14
		0.9	0.7	0.65	0.02	0.89	0.04	0.85	0.03	0.46	0.04	0.87	0.11	0.81	0.12	0.26	0.03	0.75	0.13	0.65	0.13
			0.9	0.71	0.03	0.97	1.82	0.80	0.05	0.43	0.05	0.86	4.17	0.71	0.14	0.20	0.05	0.83	3.03	0.54	0.14
	50	0.3	0.1	0.61	0.01	0.88	0.02	0.86	0.02	0.49	0.02	0.95	0.07	0.91	0.07	0.20	0.02	0.87	0.09	0.76	0.09
			0.3	0.64	0.01	0.95	0.02	0.94	0.02	0.36	0.02	0.88	0.05	0.85	0.05	0.37	0.01	0.89	0.07	0.80	0.07
			0.5	0.64	0.01	0.92	0.01	0.88	0.01	0.23	0.01	0.76	0.05	0.79	0.05	0.18	0.01	0.88	0.06	0.79	0.06
			0.7	0.68	0.01	0.92	0.01	0.89	0.01	0.48	0.02	0.92	0.05	0.91	0.05	0.31	0.01	0.90	0.07	0.84	0.07
		0.8	0.9	0.68	0.01	0.89	0.02	0.83	0.02	0.38	0.02	0.92	0.07	0.85	0.07	0.39	0.02	0.87	0.09	0.81	0.09
			0.1	0.78	0.01	0.95	0.02	0.90	0.02	0.38	0.02	0.88	0.07	0.87	0.07	0.33	0.02	0.82	0.09	0.71	0.09
			0.3	0.72	0.01	0.90	0.01	0.90	0.01	0.35	0.02	0.90	0.05	0.88	0.05	0.23	0.02	0.84	0.07	0.79	0.07
			0.5	0.67	0.01	0.86	0.01	0.85	0.01	0.43	0.01	0.84	0.04	0.82	0.04	0.28	0.01	0.90	0.06	0.81	0.06
		0.9	0.7	0.71	0.01	0.93	0.01	0.94	0.01	0.47	0.02	0.91	0.05	0.90	0.05	0.35	0.01	0.84	0.07	0.77	0.07
			0.9	0.78	0.01	0.94	0.02	0.90	0.02	0.38	0.02	0.87	0.07	0.79	0.07	0.26	0.02	0.84	0.09	0.76	0.08
	100	0.3	0.1	0.57	0.01	0.88	0.01	0.82	0.01	0.35	0.02	0.94	0.05	0.94	0.05	0.26	0.01	0.86	0.07	0.78	0.06
			0.3	0.56	0.01	0.87	0.01	0.88	0.01	0.32	0.01	0.87	0.04	0.87	0.04	0.22	0.01	0.88	0.05	0.86	0.05
			0.5	0.63	0.01	0.79	0.01	0.85	0.01	0.42	0.01	0.94	0.03	0.92	0.03	0.29	0.01	0.86	0.05	0.85	0.05
			0.7	0.61	0.01	0.88	0.01	0.90	0.01	0.37	0.01	0.89	0.04	0.90	0.04	0.24	0.01	0.89	0.05	0.82	0.05
		0.8	0.9	0.75	0.01	0.95	0.01	0.90	0.01	0.37	0.02	0.88	0.05	0.89	0.05	0.25	0.02	0.90	0.07	0.76	0.07
			0.1	0.63	0.01	0.84	0.01	0.77	0.01	0.54	0.02	0.91	0.05	0.93	0.05	0.28	0.01	0.85	0.06	0.79	0.06
			0.3	0.61	0.01	0.86	0.01	0.80	0.01	0.43	0.01	0.86	0.03	0.85	0.04	0.28	0.01	0.87	0.05	0.81	0.05
			0.5	0.63	0.01	0.85	0.01	0.85	0.01	0.36	0.01	0.88	0.03	0.90	0.03	0.23	0.01	0.81	0.05	0.77	0.05
		0.9	0.7	0.61	0.01	0.85	0.01	0.82	0.01	0.38	0.01	0.87	0.04	0.88	0.04	0.26	0.01	0.82	0.05	0.76	0.05
			0.9	0.63	0.01	0.92	0.01	0.84	0.01	0.45	0.01	0.90	0.05	0.90	0.04	0.21	0.01	0.85	0.06	0.77	0.05
30	10	0.3	0.1	0.68	0.02	0.90	0.03	0.87	0.02	0.43	0.03	0.85	0.10	0.79	0.09	0.26	0.03	0.85	0.13	0.73	0.11
			0.3	0.68	0.01	0.90	0.02	0.91	0.02	0.41	0.02	0.86	0.06	0.86	0.06	0.27	0.02	0.84	0.08	0.78	0.08
			0.5	0.73	0.01	0.91	0.02	0.92	0.02	0.40	0.02	0.86	0.06	0.83	0.06	0.23	0.02	0.82	0.09	0.78	0.09
			0.7	0.66	0.01	0.84	0.02	0.80	0.02	0.36	0.02	0.80	0.06	0.83	0.07	0.32	0.02	0.82	0.09	0.73	0.10
		0.8	0.9	0.72	0.02	0.94	0.04	0.92	0.03	0.41	0.03	0.85	0.13	0.78	0.08	0.26	0.03	0.80	0.15	0.66	0.11
			0.1	0.71	0.02	0.90	0.03	0.88	0.03	0.48	0.03	0.80	0.07	0.80	0.07	0.28	0.03	0.78	0.09	0.69	0.09
			0.3	0.68	0.01	0.88	0.02	0.89	0.02	0.50	0.02	0.82	0.05	0.82	0.06	0.26	0.02	0.76	0.07	0.68	0.07
			0.5	0.65	0.01	0.88	0.02	0.88	0.02	0.35	0.02	0.76	0.05	0.77	0.05	0.33	0.02	0.75	0.07	0.74	0.07
		0.9	0.7	0.71	0.01	0.82	0.02	0.84	0.02	0.48	0.02	0.89	0.06	0.87	0.06	0.29	0.02	0.70	0.07	0.70	0.08
			0.9	0.68	0.02	0.88	0.03	0.78	0.02	0.40	0.03	0.86	0.07	0.78	0.07	0.23	0.03	0.71	0.09	0.60	0.09
	50	0.3	0.1	0.70	0.01	0.89	0.01	0.87	0.01	0.41	0.01	0.86	0.04	0.88	0.04	0.28	0.01	0.87	0.05	0.79	0.05
			0.3	0.64	0.00	0.89	0.01	0.88	0.01	0.45	0.01	0.90	0.03	0.91	0.03	0.25	0.01	0.89	0.04	0.83	0.04
			0.5	0.69	0.00	0.90	0.01	0.86	0.01	0.40	0.01	0.88	0.03	0.89	0.03	0.29	0.01	0.90	0.04	0.91	0.04
			0.7	0.57	0.00	0.86	0.01	0.80	0.01	0.50	0.01	0.82	0.03	0.81	0.03	0.18	0.01	0.80	0.04	0.78	0.04
		0.8	0.9	0.66	0.01	0.92	0.01	0.87	0.01	0.43	0.01	0.95	0.04	0.91	0.04	0.31	0.01	0.88	0.06	0.84	0.05
			0.1	0.73	0.01	0.89	0.01	0.87	0.01	0.45	0.01	0.92	0.04	0.91	0.03	0.26	0.01	0.87	0.05	0.76	0.04
			0.3	0.68	0.00	0.87	0.01	0.87	0.01	0.35	0.01	0.91	0.03	0.87	0.03	0.25	0.01	0.86	0.04	0.83	0.04
			0.5	0.68	0.00	0.91	0.01	0.89	0.01	0.37	0.01	0.83	0.03	0.83	0.03	0.32	0.01	0.92	0.04	0.87	0.04
		0.9	0.7	0.67	0.00	0.88	0.01	0.87	0.01	0.40	0.01	0.86	0.03	0.87	0.03	0.37	0.01	0.90	0.04	0.88	0.04
			0.9	0.78	0.01	0.91	0.01	0.88	0.01	0.45	0.01	0.88	0.04	0.88	0.04	0.30	0.01	0.85	0.05	0.78	0.05
100	0.3	0.1	0.60	0.00	0.85	0.01	0.86	0.01	0.39	0.01	0.92	0.03	0.92	0.03	0.25	0.01	0.87	0.04	0.85	0.04	
		0.3	0.61	0.00	0.89	0.01	0.87	0.01	0.42	0.01	0.89	0.02	0.90	0.02	0.27	0.01	0.87	0.03	0.81	0.03	
		0.5	0.68	0.00	0.89	0.01	0.89	0.01	0.45	0.01	0.94	0.02	0.93	0.02	0.22	0.01	0.90	0.03	0.84	0.03	
		0.7	0.64	0.00	0.89	0.01	0.92	0.01	0.44	0.01	0.93	0.02	0.91	0.02	0.26	0.01	0.91	0.03	0.85	0.03	
	0.8	0.9	0.66	0.00	0.90	0.01	0.92	0.01	0.43	0.01	0.97	0.03	0.95	0.03	0.24	0.01	0.89	0.04	0.86	0.04	
		0.1	0.69	0.00	0.93	0.01	0.89	0.01	0.34	0.01	0.87	0.03	0.89	0.03	0.14	0.01	0.83	0.04	0.84	0.03	
		0.3	0.61	0.00	0.85	0.01	0.85	0.01	0.40	0.01	0.90	0.02	0.92	0.02	0.26	0.01	0.91	0.03	0.89	0.03	
		0.5	0.71	0.00	0.92	0.01	0.93	0.01	0.45	0.01	0.92	0.02	0.93	0.02	0.35	0.01	0.88	0.03	0.85	0.03	
	0.9	0.7	0.70	0.00	0.89	0.01	0.89	0.01	0.39	0.01	0.85	0.02	0.85	0.02	0.26	0.01	0.84	0.03	0.75	0.02	
		0.9	0.62	0.00	0.83	0.01	0.86	0.01	0.53	0.01	0.88	0.03	0.90	0.03	0.32	0.01	0.85	0.04	0.80	0.04	

Table 2.9: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = .5$.

				Intercept						Slope of x_{ij}						Slope of x_{ij}^2					
				ALD		boot		IJ		ALD		boot		IJ		ALD		boot		IJ	
I	J	ρ	τ	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
10	10	0.3	0.1	0.99	0.08	0.98	0.26	0.80	0.05	0.89	0.12	0.98	0.62	0.90	0.16	0.59	0.11	0.91	0.56	0.69	0.18
			0.3	0.98	0.04	0.89	0.04	0.87	0.03	0.80	0.08	0.91	0.11	0.82	0.11	0.58	0.07	0.87	0.15	0.83	0.14
			0.5	0.97	0.04	0.94	0.03	0.91	0.03	0.72	0.08	0.89	0.11	0.86	0.11	0.52	0.07	0.88	0.15	0.83	0.14
			0.7	0.98	0.05	0.98	0.04	0.93	0.04	0.70	0.08	0.83	0.11	0.78	0.10	0.55	0.08	0.83	0.17	0.81	0.15
		0.8	0.9	0.94	0.08	0.94	1.01	0.76	0.05	0.85	0.12	0.90	2.89	0.78	0.13	0.56	0.11	0.83	2.92	0.64	0.15
			0.1	0.99	0.08	0.98	0.27	0.87	0.05	0.86	0.13	0.94	0.61	0.84	0.14	0.73	0.13	0.87	0.54	0.68	0.16
			0.3	0.98	0.05	0.94	0.04	0.91	0.04	0.78	0.09	0.83	0.11	0.78	0.10	0.54	0.08	0.72	0.14	0.68	0.12
			0.5	0.97	0.04	0.90	0.03	0.84	0.03	0.77	0.08	0.82	0.10	0.79	0.09	0.64	0.08	0.75	0.12	0.76	0.11
		0.9	0.7	0.97	0.05	0.89	0.04	0.84	0.03	0.77	0.08	0.83	0.10	0.74	0.09	0.62	0.08	0.75	0.12	0.64	0.10
			0.9	0.99	0.08	0.96	1.60	0.82	0.05	0.92	0.13	0.92	3.80	0.86	0.13	0.63	0.12	0.89	3.21	0.64	0.15
50	0.3	0.1	1.00	0.03	0.93	0.02	0.91	0.02	0.81	0.05	0.90	0.07	0.88	0.07	0.62	0.05	0.86	0.09	0.83	0.09	
			0.3	0.97	0.02	0.93	0.01	0.90	0.01	0.74	0.03	0.90	0.05	0.84	0.05	0.46	0.03	0.87	0.07	0.82	0.07
			0.5	0.94	0.02	0.90	0.01	0.87	0.01	0.66	0.03	0.86	0.05	0.84	0.05	0.57	0.03	0.89	0.07	0.87	0.07
			0.7	0.99	0.02	0.88	0.01	0.85	0.01	0.79	0.04	0.93	0.05	0.95	0.05	0.65	0.03	0.88	0.07	0.88	0.07
		0.8	0.9	0.99	0.03	0.91	0.02	0.85	0.02	0.87	0.05	0.91	0.07	0.91	0.07	0.68	0.05	0.88	0.09	0.84	0.09
			0.1	0.98	0.03	0.89	0.02	0.86	0.02	0.83	0.05	0.91	0.07	0.90	0.06	0.54	0.05	0.86	0.09	0.80	0.08
			0.3	1.00	0.02	0.93	0.01	0.88	0.01	0.76	0.04	0.91	0.05	0.89	0.05	0.60	0.03	0.88	0.07	0.84	0.07
			0.5	0.96	0.02	0.84	0.01	0.86	0.01	0.76	0.03	0.95	0.05	0.92	0.04	0.57	0.03	0.76	0.06	0.75	0.06
		0.9	0.7	0.96	0.02	0.90	0.01	0.90	0.01	0.66	0.04	0.86	0.05	0.83	0.05	0.62	0.03	0.84	0.06	0.78	0.06
			0.9	0.99	0.03	0.97	0.02	0.93	0.02	0.82	0.05	0.90	0.07	0.90	0.06	0.61	0.05	0.89	0.10	0.81	0.09
100	0.3	0.1	0.99	0.02	0.95	0.01	0.92	0.01	0.76	0.03	0.84	0.05	0.86	0.05	0.63	0.03	0.90	0.06	0.87	0.06	
			0.3	0.96	0.01	0.91	0.01	0.90	0.01	0.68	0.02	0.86	0.04	0.84	0.03	0.58	0.02	0.89	0.05	0.88	0.05
			0.5	0.96	0.01	0.92	0.01	0.91	0.01	0.78	0.02	0.91	0.03	0.90	0.03	0.49	0.02	0.87	0.05	0.85	0.05
			0.7	0.92	0.01	0.87	0.01	0.83	0.01	0.73	0.02	0.85	0.04	0.82	0.04	0.52	0.02	0.81	0.05	0.82	0.05
		0.8	0.9	0.97	0.02	0.88	0.01	0.89	0.01	0.77	0.04	0.91	0.05	0.89	0.05	0.53	0.03	0.87	0.06	0.80	0.06
			0.1	1.00	0.02	0.94	0.01	0.88	0.01	0.78	0.03	0.91	0.05	0.91	0.04	0.62	0.03	0.87	0.07	0.83	0.06
			0.3	0.99	0.01	0.90	0.01	0.88	0.01	0.85	0.02	0.97	0.04	0.92	0.03	0.60	0.02	0.88	0.05	0.84	0.05
			0.5	0.95	0.01	0.91	0.01	0.90	0.01	0.71	0.02	0.87	0.03	0.87	0.03	0.56	0.02	0.84	0.05	0.80	0.05
		0.9	0.7	0.97	0.01	0.92	0.01	0.87	0.01	0.77	0.02	0.89	0.04	0.88	0.04	0.54	0.02	0.87	0.05	0.88	0.05
			0.9	0.98	0.02	0.95	0.01	0.89	0.01	0.78	0.04	0.94	0.05	0.90	0.04	0.62	0.03	0.94	0.06	0.86	0.06
30	10	0.3	0.1	0.99	0.04	0.94	0.03	0.91	0.03	0.83	0.07	0.92	0.09	0.88	0.08	0.59	0.06	0.84	0.12	0.78	0.11
			0.3	0.99	0.02	0.88	0.02	0.89	0.02	0.81	0.05	0.85	0.06	0.84	0.06	0.58	0.04	0.82	0.08	0.83	0.09
			0.5	0.96	0.02	0.84	0.02	0.82	0.02	0.74	0.04	0.90	0.06	0.89	0.06	0.54	0.04	0.85	0.08	0.83	0.08
			0.7	0.98	0.02	0.84	0.02	0.83	0.02	0.72	0.05	0.83	0.06	0.83	0.06	0.61	0.04	0.83	0.09	0.83	0.09
		0.8	0.9	1.00	0.04	0.89	0.07	0.86	0.02	0.78	0.07	0.84	0.22	0.83	0.09	0.61	0.06	0.85	0.23	0.79	0.11
			0.1	0.99	0.04	0.92	0.03	0.91	0.02	0.80	0.07	0.78	0.07	0.77	0.07	0.64	0.07	0.75	0.09	0.69	0.08
			0.3	0.98	0.02	0.84	0.02	0.86	0.02	0.83	0.05	0.80	0.05	0.79	0.05	0.66	0.05	0.77	0.07	0.77	0.07
			0.5	0.98	0.02	0.94	0.02	0.94	0.02	0.78	0.05	0.82	0.05	0.83	0.05	0.51	0.04	0.74	0.07	0.78	0.07
		0.9	0.7	1.00	0.02	0.84	0.02	0.84	0.02	0.81	0.05	0.80	0.05	0.78	0.05	0.55	0.05	0.75	0.07	0.71	0.07
			0.9	1.00	0.04	0.86	0.03	0.81	0.02	0.83	0.07	0.78	0.08	0.73	0.07	0.74	0.07	0.79	0.10	0.75	0.08
50	0.3	0.1	0.97	0.01	0.92	0.01	0.84	0.01	0.81	0.03	0.94	0.04	0.91	0.04	0.64	0.03	0.88	0.05	0.86	0.05	
			0.3	0.96	0.01	0.89	0.01	0.88	0.01	0.73	0.02	0.87	0.03	0.85	0.03	0.54	0.02	0.87	0.04	0.87	0.04
			0.5	0.99	0.01	0.91	0.01	0.91	0.01	0.76	0.02	0.88	0.03	0.91	0.03	0.64	0.02	0.93	0.04	0.93	0.04
			0.7	0.96	0.01	0.91	0.01	0.89	0.01	0.76	0.02	0.86	0.03	0.87	0.03	0.57	0.02	0.88	0.04	0.90	0.04
		0.8	0.9	0.97	0.01	0.93	0.01	0.89	0.01	0.78	0.03	0.92	0.04	0.88	0.04	0.62	0.03	0.89	0.05	0.89	0.05
			0.1	0.98	0.02	0.91	0.01	0.88	0.01	0.76	0.03	0.83	0.04	0.82	0.03	0.64	0.03	0.88	0.05	0.83	0.04
			0.3	0.98	0.01	0.90	0.01	0.90	0.01	0.72	0.02	0.90	0.03	0.86	0.03	0.57	0.02	0.89	0.04	0.89	0.04
			0.5	0.94	0.01	0.83	0.01	0.82	0.01	0.67	0.02	0.81	0.03	0.82	0.03	0.60	0.02	0.87	0.04	0.86	0.04
		0.9	0.7	0.97	0.01	0.87	0.01	0.89	0.01	0.74	0.02	0.85	0.03	0.81	0.03	0.64	0.02	0.86	0.04	0.85	0.04
			0.9	0.97	0.01	0.87	0.01	0.84	0.01	0.86	0.03	0.91	0.04	0.90	0.04	0.64	0.03	0.80	0.05	0.78	0.05
100	0.3	0.1	0.97	0.01	0.91	0.01	0.90	0.01	0.83	0.02	0.92	0.03	0.90	0.03	0.59	0.02	0.92	0.04	0.89	0.04	
			0.3	0.97	0.01	0.95	0.01	0.93	0.01	0.71	0.01	0.88	0.02	0.87	0.02	0.48	0.01	0.90	0.03	0.91	0.03
			0.5	0.98	0.01	0.94	0.01	0.92	0.01	0.75	0.01	0.91	0.02	0.92	0.02	0.58	0.01	0.91	0.03	0.89	0.03
			0.7	0.95	0.01	0.85	0.01	0.87	0.01	0.69	0.01	0.86	0.02	0.89	0.02	0.51	0.01	0.81	0.03	0.79	0.03
		0.8	0.9	0.99	0.01	0.92	0.01	0.88	0.01	0.80	0.02	0.88	0.03	0.87	0.03	0.60	0.02	0.88	0.04	0.83	0.04
			0.1	0.99	0.01	0.90	0.01	0.89	0.01	0.83	0.02	0.89	0.03	0.89	0.03	0.65	0.02	0.94	0.03	0.92	0.03
			0.3	0.98	0.01	0.90	0.01	0.90	0.01	0.69	0.01	0.86	0.02	0.80	0.02	0.64	0.01	0.90	0.03	0.91	0.03
			0.5	0.97	0.01	0.91	0.01	0.91	0.01	0.74	0.01	0.89	0.02	0.89	0.02	0.60	0.01	0.86	0.03	0.86	0.03
		0.9	0.7	0.94	0.01	0.88	0.01	0.84	0.01	0.78	0.01	0.94	0.02	0.93	0.02	0.58	0.01	0.86	0.03	0.85	0.03
			0.9	0.97	0.01	0.91	0.01	0.89	0.01	0.80	0.02	0.91	0.03	0.90	0.03	0.56	0.02	0.86	0.04	0.83	0.04

Table 2.10: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 1$.

I	J	ρ	τ	Intercept				Slope of x_{ij}						Slope of x_{ij}^2					
				ALD		boot		IJ		ALD		boot		IJ		ALD		boot	
				ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
10	10	0.3	0.1	0.99	0.13	0.95	0.21	0.86	0.05	0.96	0.18	0.95	0.44	0.81	0.14	0.78	0.16	0.86	0.41
			0.3	1.00	0.07	0.87	0.04	0.85	0.03	0.92	0.12	0.91	0.12	0.87	0.11	0.75	0.11	0.80	0.15
			0.5	1.00	0.06	0.83	0.03	0.83	0.03	0.89	0.11	0.85	0.11	0.84	0.10	0.68	0.10	0.80	0.15
			0.7	1.00	0.07	0.87	0.04	0.88	0.04	0.85	0.12	0.83	0.11	0.80	0.11	0.73	0.11	0.84	0.14
		0.8	0.9	0.99	0.13	0.97	1.26	0.86	0.06	0.90	0.18	0.95	3.47	0.78	0.14	0.84	0.16	0.94	3.42
			0.1	0.99	0.13	0.96	0.27	0.90	0.06	0.93	0.20	0.90	0.58	0.78	0.14	0.85	0.18	0.82	0.46
			0.3	1.00	0.07	0.90	0.04	0.90	0.03	0.94	0.12	0.86	0.10	0.80	0.09	0.76	0.12	0.74	0.13
			0.5	1.00	0.06	0.90	0.03	0.91	0.03	0.97	0.11	0.81	0.09	0.76	0.08	0.79	0.11	0.70	0.11
		0.9	0.7	1.00	0.07	0.89	0.04	0.90	0.03	0.92	0.12	0.78	0.09	0.74	0.09	0.78	0.12	0.75	0.13
			0.9	1.00	0.13	0.98	2.14	0.90	0.05	0.93	0.19	0.94	4.33	0.80	0.12	0.81	0.18	0.85	3.30
	50	0.3	0.1	1.00	0.04	0.94	0.02	0.91	0.02	0.95	0.07	0.93	0.07	0.92	0.06	0.75	0.07	0.86	0.09
			0.3	1.00	0.03	0.91	0.01	0.90	0.01	0.92	0.05	0.89	0.05	0.90	0.05	0.70	0.05	0.81	0.07
			0.5	1.00	0.02	0.89	0.01	0.87	0.01	0.91	0.05	0.90	0.05	0.90	0.05	0.77	0.04	0.91	0.07
			0.7	1.00	0.03	0.81	0.01	0.86	0.01	0.89	0.05	0.87	0.05	0.85	0.05	0.70	0.05	0.90	0.07
		0.8	0.9	1.00	0.04	0.95	0.02	0.88	0.02	0.91	0.07	0.87	0.07	0.84	0.06	0.67	0.06	0.78	0.09
			0.1	1.00	0.04	0.89	0.02	0.83	0.02	0.90	0.07	0.87	0.07	0.83	0.06	0.80	0.07	0.88	0.09
			0.3	0.99	0.03	0.87	0.01	0.81	0.01	0.87	0.05	0.92	0.05	0.87	0.05	0.77	0.05	0.87	0.07
			0.5	0.99	0.02	0.93	0.01	0.91	0.01	0.86	0.05	0.84	0.05	0.83	0.04	0.65	0.04	0.84	0.06
		0.9	0.7	1.00	0.03	0.89	0.01	0.91	0.01	0.96	0.05	0.93	0.05	0.89	0.05	0.70	0.05	0.84	0.07
			0.9	1.00	0.04	0.96	0.02	0.91	0.02	0.95	0.07	0.92	0.07	0.85	0.06	0.79	0.07	0.87	0.09
	100	0.3	0.1	1.00	0.03	0.91	0.01	0.85	0.01	0.94	0.05	0.92	0.05	0.86	0.04	0.81	0.05	0.90	0.06
			0.3	1.00	0.02	0.95	0.01	0.92	0.01	0.84	0.03	0.86	0.04	0.80	0.03	0.71	0.03	0.88	0.05
			0.5	1.00	0.02	0.96	0.01	0.93	0.01	0.87	0.03	0.85	0.03	0.86	0.03	0.72	0.03	0.93	0.05
			0.7	1.00	0.02	0.88	0.01	0.87	0.01	0.89	0.03	0.91	0.04	0.88	0.03	0.69	0.03	0.83	0.05
		0.8	0.9	1.00	0.03	0.90	0.01	0.88	0.01	0.94	0.05	0.86	0.05	0.85	0.04	0.70	0.05	0.83	0.06
			0.1	1.00	0.03	0.90	0.01	0.88	0.01	0.97	0.05	0.94	0.05	0.92	0.05	0.75	0.05	0.86	0.07
			0.3	0.99	0.02	0.88	0.01	0.85	0.01	0.85	0.03	0.87	0.04	0.86	0.03	0.72	0.03	0.92	0.05
			0.5	1.00	0.02	0.88	0.01	0.87	0.01	0.92	0.03	0.92	0.03	0.90	0.03	0.67	0.03	0.84	0.05
		0.9	0.7	1.00	0.02	0.88	0.01	0.86	0.01	0.92	0.03	0.94	0.03	0.92	0.03	0.74	0.03	0.89	0.05
			0.9	1.00	0.03	0.89	0.01	0.87	0.01	0.94	0.05	0.90	0.05	0.89	0.04	0.77	0.05	0.84	0.06
30	10	0.3	0.1	1.00	0.06	0.88	0.03	0.83	0.03	0.92	0.10	0.90	0.09	0.86	0.08	0.75	0.09	0.85	0.12
			0.3	1.00	0.04	0.87	0.02	0.87	0.02	0.93	0.07	0.85	0.06	0.86	0.06	0.81	0.06	0.85	0.08
			0.5	1.00	0.03	0.85	0.02	0.86	0.02	0.91	0.06	0.86	0.06	0.84	0.06	0.72	0.06	0.81	0.08
			0.7	1.00	0.03	0.85	0.02	0.86	0.02	0.96	0.07	0.86	0.06	0.81	0.06	0.69	0.06	0.78	0.08
		0.8	0.9	1.00	0.06	0.94	0.03	0.93	0.03	0.96	0.10	0.90	0.11	0.85	0.08	0.83	0.09	0.83	0.21
			0.1	1.00	0.06	0.89	0.03	0.84	0.02	0.96	0.10	0.88	0.07	0.81	0.07	0.75	0.10	0.72	0.10
			0.3	1.00	0.04	0.87	0.02	0.86	0.02	0.93	0.07	0.79	0.05	0.79	0.05	0.70	0.07	0.72	0.07
			0.5	1.00	0.03	0.89	0.01	0.92	0.02	0.92	0.06	0.73	0.05	0.74	0.05	0.76	0.06	0.71	0.06
		0.9	0.7	1.00	0.04	0.86	0.02	0.88	0.02	0.88	0.07	0.82	0.05	0.83	0.05	0.76	0.06	0.71	0.06
			0.9	1.00	0.06	0.85	0.09	0.85	0.02	0.93	0.10	0.85	0.27	0.82	0.07	0.75	0.09	0.66	0.23
	50	0.3	0.1	1.00	0.02	0.91	0.01	0.89	0.01	0.95	0.04	0.93	0.04	0.91	0.04	0.79	0.04	0.85	0.05
			0.3	1.00	0.01	0.89	0.01	0.89	0.01	0.93	0.03	0.91	0.03	0.92	0.03	0.65	0.03	0.78	0.04
			0.5	1.00	0.01	0.90	0.01	0.90	0.01	0.92	0.03	0.90	0.03	0.90	0.03	0.73	0.02	0.87	0.04
			0.7	1.00	0.01	0.89	0.01	0.87	0.01	0.90	0.03	0.91	0.03	0.90	0.03	0.71	0.03	0.85	0.04
		0.8	0.9	1.00	0.02	0.95	0.01	0.90	0.01	0.94	0.04	0.91	0.04	0.90	0.04	0.78	0.04	0.90	0.05
			0.1	1.00	0.02	0.90	0.01	0.90	0.01	0.97	0.04	0.93	0.04	0.90	0.03	0.73	0.04	0.80	0.05
			0.3	1.00	0.01	0.88	0.01	0.88	0.01	0.91	0.03	0.88	0.03	0.87	0.03	0.74	0.03	0.84	0.04
			0.5	0.99	0.01	0.89	0.01	0.89	0.01	0.91	0.03	0.89	0.03	0.88	0.03	0.66	0.03	0.81	0.03
		0.9	0.7	1.00	0.01	0.90	0.01	0.89	0.01	0.89	0.03	0.90	0.03	0.89	0.03	0.69	0.03	0.84	0.04
			0.9	1.00	0.02	0.88	0.01	0.86	0.01	0.98	0.04	0.92	0.04	0.89	0.04	0.72	0.04	0.81	0.05
	100	0.3	0.1	1.00	0.01	0.88	0.01	0.85	0.01	0.89	0.03	0.87	0.03	0.88	0.03	0.68	0.03	0.88	0.04
			0.3	1.00	0.01	0.89	0.01	0.88	0.01	0.86	0.02	0.85	0.02	0.85	0.02	0.75	0.02	0.85	0.03
			0.5	0.99	0.01	0.86	0.01	0.86	0.01	0.96	0.02	0.94	0.02	0.95	0.02	0.67	0.02	0.88	0.03
			0.7	1.00	0.01	0.89	0.01	0.88	0.01	0.85	0.02	0.87	0.02	0.84	0.02	0.78	0.02	0.89	0.03
		0.8	0.9	1.00	0.01	0.91	0.01	0.90	0.01	0.96	0.03	0.91	0.03	0.90	0.03	0.72	0.03	0.88	0.04
			0.1	1.00	0.01	0.91	0.01	0.89	0.01	0.92	0.03	0.85	0.03	0.82	0.02	0.77	0.03	0.89	0.03
			0.3	1.00	0.01	0.86	0.01	0.87	0.01	0.93	0.02	0.95	0.02	0.93	0.02	0.68	0.02	0.89	0.03
			0.5	1.00	0.01	0.91	0.01	0.91	0.01	0.93	0.02	0.91	0.02	0.94	0.02	0.71	0.02	0.84	0.03
		0.9	0.7	0.99	0.01	0.85	0.01	0.85	0.01	0.91	0.02	0.91	0.02	0.88	0.02	0.68	0.02	0.81	0.03
			0.9	1.00	0.01	0.91	0.01	0.87	0.01	0.94	0.03	0.94	0.03	0.94	0.03	0.80	0.03	0.91	0.04

Table 2.11: Empirical mean length (EML) and empirical coverage probability (ECP) for the intervals when $\sigma = 5$.

I	J	ρ	τ	Intercept				Slope of x_{ij}						Slope of x_{ij}^2					
				ALD		boot		IJ		ALD		boot		IJ		ALD		boot	
ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML	ECP	EML
10	10	0.3	0.1	1.00	0.43	0.99	0.21	0.90	0.06	1.00	0.55	0.93	0.53	0.88	0.15	1.00	0.46	0.89	0.51
			0.3	1.00	0.21	0.93	0.04	0.94	0.05	1.00	0.30	0.87	0.12	0.81	0.10	0.99	0.28	0.85	0.16
			0.5	1.00	0.18	0.99	0.03	0.99	0.04	0.99	0.27	0.87	0.11	0.85	0.10	1.00	0.25	0.88	0.15
			0.7	1.00	0.21	0.93	0.04	0.98	0.05	1.00	0.30	0.92	0.12	0.88	0.11	1.00	0.28	0.86	0.18
			0.9	1.00	0.44	0.98	1.76	0.87	0.06	1.00	0.57	0.93	4.93	0.85	0.17	0.99	0.47	0.95	4.74
		0.8	0.1	1.00	0.47	0.97	0.24	0.94	0.08	1.00	0.62	0.92	0.51	0.94	0.22	1.00	0.56	0.89	0.45
			0.3	1.00	0.22	0.94	0.04	0.98	0.05	1.00	0.32	0.87	0.11	0.85	0.10	1.00	0.29	0.76	0.14
			0.5	1.00	0.19	0.89	0.03	0.91	0.04	1.00	0.29	0.82	0.09	0.79	0.08	1.00	0.28	0.73	0.11
			0.7	1.00	0.22	0.87	0.04	0.91	0.04	1.00	0.33	0.78	0.10	0.80	0.10	1.00	0.33	0.84	0.14
			0.9	1.00	0.45	0.98	1.90	0.93	0.08	1.00	0.61	0.92	4.41	0.97	0.20	1.00	0.54	0.90	4.32
			0.1	1.00	0.12	0.96	0.02	0.96	0.02	1.00	0.17	0.91	0.07	0.86	0.06	1.00	0.16	0.94	0.09
			0.3	1.00	0.07	0.92	0.02	0.92	0.02	1.00	0.12	0.85	0.05	0.80	0.05	0.97	0.11	0.89	0.07
			0.5	1.00	0.06	0.91	0.01	0.89	0.01	1.00	0.11	0.91	0.05	0.91	0.04	0.98	0.10	0.90	0.07
			0.7	1.00	0.07	0.89	0.02	0.89	0.02	1.00	0.11	0.91	0.05	0.90	0.05	0.98	0.11	0.93	0.07
			0.9	1.00	0.12	0.91	0.02	0.92	0.02	1.00	0.17	0.91	0.07	0.84	0.06	1.00	0.16	0.90	0.10
		0.8	0.1	1.00	0.12	0.94	0.02	0.96	0.02	1.00	0.18	0.90	0.07	0.87	0.06	0.99	0.16	0.89	0.09
			0.3	1.00	0.07	0.91	0.01	0.92	0.01	1.00	0.12	0.87	0.05	0.87	0.05	0.98	0.11	0.87	0.07
			0.5	1.00	0.06	0.88	0.01	0.89	0.01	1.00	0.11	0.94	0.05	0.91	0.04	0.99	0.10	0.85	0.06
			0.7	1.00	0.07	0.89	0.01	0.87	0.01	1.00	0.12	0.90	0.05	0.87	0.05	0.97	0.11	0.83	0.07
			0.9	1.00	0.12	0.93	0.02	0.91	0.02	1.00	0.18	0.92	0.07	0.87	0.06	1.00	0.16	0.83	0.09
		100	0.3	0.1	1.00	0.07	0.91	0.01	0.88	0.01	1.00	0.12	0.92	0.05	0.88	0.04	0.99	0.11	0.81
			0.3	1.00	0.04	0.92	0.01	0.91	0.01	0.99	0.08	0.90	0.03	0.90	0.03	0.99	0.07	0.86	0.05
			0.5	1.00	0.04	0.87	0.01	0.86	0.01	1.00	0.07	0.90	0.03	0.88	0.03	0.98	0.07	0.86	0.05
			0.7	1.00	0.04	0.93	0.01	0.91	0.01	1.00	0.08	0.88	0.04	0.87	0.03	0.99	0.07	0.88	0.05
			0.9	1.00	0.07	0.88	0.01	0.88	0.01	1.00	0.12	0.90	0.05	0.87	0.05	0.99	0.11	0.90	0.07
		0.8	0.1	1.00	0.07	0.89	0.01	0.87	0.01	1.00	0.12	0.91	0.05	0.89	0.04	0.99	0.11	0.83	0.06
			0.3	1.00	0.04	0.92	0.01	0.88	0.01	1.00	0.08	0.91	0.03	0.88	0.03	0.99	0.07	0.84	0.05
			0.5	1.00	0.04	0.89	0.01	0.85	0.01	1.00	0.07	0.89	0.03	0.87	0.03	0.96	0.07	0.84	0.05
			0.7	1.00	0.04	0.92	0.01	0.92	0.01	1.00	0.08	0.91	0.04	0.92	0.03	1.00	0.08	0.89	0.05
			0.9	1.00	0.07	0.87	0.01	0.89	0.01	1.00	0.12	0.95	0.05	0.96	0.04	0.99	0.11	0.84	0.07
		30	10	0.3	0.1	1.00	0.17	0.95	0.03	0.95	0.03	1.00	0.25	0.89	0.10	0.86	0.08	1.00	0.22
			0.3	1.00	0.09	0.88	0.02	0.90	0.02	1.00	0.16	0.80	0.06	0.79	0.06	0.98	0.15	0.85	0.09
			0.5	1.00	0.08	0.86	0.02	0.91	0.02	1.00	0.14	0.86	0.06	0.85	0.05	0.98	0.13	0.83	0.08
			0.7	1.00	0.09	0.92	0.02	0.91	0.02	1.00	0.15	0.92	0.06	0.89	0.06	0.98	0.14	0.82	0.09
			0.9	1.00	0.17	0.94	0.03	0.88	0.03	1.00	0.25	0.94	0.09	0.89	0.08	1.00	0.22	0.87	0.18
		0.8	0.1	1.00	0.18	0.95	0.03	0.99	0.04	1.00	0.27	0.77	0.08	0.86	0.09	1.00	0.25	0.81	0.10
			0.3	1.00	0.10	0.84	0.02	0.89	0.02	1.00	0.16	0.83	0.05	0.79	0.05	0.99	0.16	0.70	0.07
			0.5	1.00	0.08	0.86	0.01	0.93	0.02	1.00	0.15	0.67	0.04	0.70	0.04	0.99	0.15	0.59	0.06
			0.7	1.00	0.09	0.90	0.02	0.89	0.02	1.00	0.17	0.80	0.05	0.80	0.05	1.00	0.16	0.77	0.07
			0.9	1.00	0.18	0.86	0.03	0.96	0.04	1.00	0.27	0.78	0.07	0.87	0.08	0.99	0.24	0.69	0.09
		50	0.3	0.1	1.00	0.05	0.92	0.01	0.91	0.01	1.00	0.09	0.92	0.04	0.89	0.04	1.00	0.09	0.85
			0.3	1.00	0.03	0.90	0.01	0.89	0.01	0.99	0.06	0.91	0.03	0.92	0.03	0.98	0.06	0.91	0.04
			0.5	1.00	0.03	0.90	0.01	0.91	0.01	1.00	0.06	0.89	0.03	0.90	0.03	0.98	0.06	0.90	0.04
			0.7	1.00	0.03	0.86	0.01	0.86	0.01	1.00	0.06	0.85	0.03	0.87	0.03	0.99	0.06	0.86	0.04
			0.9	1.00	0.05	0.98	0.01	0.97	0.01	1.00	0.09	0.92	0.04	0.92	0.04	1.00	0.09	0.89	0.05
		0.8	0.1	1.00	0.05	0.92	0.01	0.92	0.01	1.00	0.09	0.95	0.04	0.94	0.04	0.99	0.09	0.87	0.05
			0.3	1.00	0.03	0.90	0.01	0.86	0.01	1.00	0.07	0.84	0.03	0.80	0.03	0.99	0.06	0.87	0.04
			0.5	1.00	0.03	0.89	0.01	0.89	0.01	1.00	0.06	0.85	0.03	0.86	0.03	1.00	0.06	0.84	0.04
			0.7	1.00	0.03	0.90	0.01	0.89	0.01	1.00	0.06	0.87	0.03	0.84	0.03	0.99	0.06	0.92	0.04
			0.9	1.00	0.06	0.94	0.01	0.93	0.01	1.00	0.09	0.87	0.04	0.87	0.03	1.00	0.09	0.85	0.05
		100	0.3	0.1	1.00	0.04	0.87	0.01	0.85	0.01	1.00	0.06	0.89	0.03	0.89	0.02	0.99	0.06	0.82
			0.3	1.00	0.02	0.87	0.01	0.87	0.01	1.00	0.05	0.91	0.02	0.84	0.02	0.96	0.04	0.93	0.03
			0.5	1.00	0.02	0.91	0.01	0.87	0.01	1.00	0.04	0.90	0.02	0.90	0.02	0.95	0.04	0.84	0.03
			0.7	1.00	0.02	0.92	0.01	0.91	0.01	1.00	0.05	0.91	0.02	0.90	0.02	0.98	0.04	0.90	0.03
			0.9	1.00	0.04	0.85	0.01	0.86	0.01	1.00	0.07	0.90	0.03	0.87	0.03	1.00	0.06	0.91	0.04
		0.8	0.1	1.00	0.04	0.87	0.01	0.87	0.01	1.00	0.07	0.88	0.03	0.86	0.03	0.99	0.06	0.84	0.04
			0.3	1.00	0.02	0.92	0.01	0.90	0.01	1.00	0.05	0.91	0.02	0.89	0.02	0.98	0.04	0.87	0.03
			0.5	1.00	0.02	0.82	0.01	0.80	0.01	1.00	0.04	0.91	0.02	0.91	0.02	0.96	0.04	0.88	0.03
			0.7	1.00	0.02	0.87	0.01	0.86	0.01	1.00	0.05	0.94	0.02	0.92	0.02	0.97	0.04	0.84	0.03
			0.9	1.00	0.04	0.91	0.01	0.89	0.01	1.00	0.07	0.90	0.03	0.87	0.03	0.99	0.06	0.87	0.04

conditions and performs similarly to the bootstrap interval, although for the coefficient for the second order term (i.e., $\beta_2(\tau)$), there is certain level of undercoverage compared to bootstrap even when $I = 10$, $J = 50$ and when the intraclass correlation is high (i.e., $\rho = .8$). We believe that this is due to the IJ interval being a first-order approximation.

Fifth, we found that the coverage for the second order term $\beta_2(\tau)$ is in general worse for all three intervals than for the intercept $\alpha(\tau)$ and slope $\beta_1(\tau)$, although the coverage becomes better when $J = 100$.

2.5 Concluding Remarks

Bayesian inference has become increasingly popular in applied settings in recent years across many different disciplines. Reasons for this trend include the flexibility in modeling complex phenomena and ease of implementation thanks to the availability of modern probabilistic programming frameworks (e.g., Stan, Carpenter et al., 2017).

Bayesian quantile regression using ALD has been applied widely in practice but is lacking simple and valid methods for estimating standard errors. Credible intervals based on model-based standard errors have poor coverage and the popular adjustment (Yang et al., 2016) has been shown unable to address the problem (Lee, 2020). We applied the Bayesian infinite jackknife idea to two Bayesian quantile ALD regression settings for both independent and clustered data and demonstrated its validity and limitations through a series of simulation studies.

It is worth pointing out that the general idea used here to obtain approximated resampling-based standard errors can have many more applications besides the quantile regression setting and based beyond on the IJ approximation considered in the current chapter. One immediate use case would be to incorporate sampling weights in Bayesian inference for survey data, an issue that has been a challenge in the literature (e.g., Gelman, 2007). Although methods for incorporating survey weights have been proposed recently (e.g., Si et al., 2015), implementation requires great effort and poses severe challenges to practitioners.

As noted before, we chose the the IJ approximation because of its practical convenience. However, there are many alternative approaches to obtain Bayesian resampling estimates hence resampling-based variance estimates. Lee et al. (2017) proposed to use importance sampling for approximating resampling Bayesian estimates as a general way to assess frequentist standard errors for Bayesian estimates. The naive importance sampling idea has been improved in a series of recent papers (Vehtari et al., 2017; Paananen et al., 2021) in the context of Bayesian model evaluation and assessment. Such methods can also be applied to Bayesian ALD quantile regression as well as in more general contexts (e.g., for other more complicated models, models with sampling weights, or misspecified models) for variance estimation. The comparison between these two lines of research is worth exploring further. We leave this to future investigations.

2.6 References

- Alaa, A. and Van Der Schaar, M. (2020). Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*, pages 165–174. PMLR.
- Andrews, D. F. and Hampel, F. R. (2015). Robust estimates of location. In *Robust Estimates of Location*. Princeton University Press.
- Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:1–43.
- Benoit, D. F. and Van den Poel, D. (2012). Binary quantile regression: a bayesian approach based on the asymmetric laplace distribution. *Journal of Applied Econometrics*, 27(7):1174–1188.
- Benoit, D. F. and Van den Poel, D. (2017). bayesqr: A bayesian approach to quantile regression. *Journal of Statistical Software*, 76(1):1–32.
- Beran, R. (1984). Jackknife approximations to bootstrap estimates. *The Annals of Statistics*, 12(1):101–118.
- Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: When can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*.
- Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.
- Buzas, J. S. (1997). Fast estimators of the jackknife. *The American Statistician*, 51(3):235–240.
- Cain, K. C. and Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, 40(2):493–499.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge university press.
- Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3):531–545.
- Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.
- Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.
- Efron, B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association*, 85(409):79–89.
- Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):83–111.
- Efron, B. (2015). Frequentist accuracy of bayesian estimates. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 77(3):617.
- Efron, B. and Hastie, T. (2021). *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, volume 6. Cambridge University Press.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Fang, Y. (2011). Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. *Journal of Data Science*, 9(1):15–21.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.
- Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.
- Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.
- Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479.
- Giordano, R. (2020). Effortless frequentist covariances of posterior expectations in stan. In *StanCon 2020 Contributed Talks*.

- Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational bayes. *Journal of Machine Learning Research*, 19(51):1–49.
- Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019). A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.
- Goldstein, H. (2011). Bootstrapping in multilevel models. *Handbook of advanced multilevel analysis*, pages 163–171.
- Hagemann, A. (2017). Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association*, 112(517):446–456.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons.
- Hanley, J. A. and MacGibbon, B. (2006). Creating non-parametric bootstrap samples using poisson frequencies. *Computer Methods and Programs in Biomedicine*, 83(1):57–62.
- Hinkley, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika*, 65(1):13–21.
- Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.
- Jaekel, L. A. (1972). *The infinitesimal jackknife*. Bell Telephone Laboratories.
- Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.
- Koenker, R. (2009). quantreg: Quantile regression. <http://CRAN.R-project.org/package=quantreg>.
- Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.
- Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.
- Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.
- Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319.

- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.
- Lee, D., Carroll, R. J., and Sinha, S. (2017). Frequentist standard errors of bayes estimators. *Computational Statistics*, 32(3):867–888.
- Lee, J.-H. (2020). *Essays on Treatment Effect Heterogeneity in Education Policy Interventions*. PhD thesis, University of California, Berkeley.
- Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3):802–829.
- Paananen, T., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):1–19.
- Parente, P. M. and Silva, J. M. S. (2016). Quantile regression with clustered data. *Journal of Econometric Methods*, 5(1):1–15.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.
- Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reich, B. J., Bondell, H. D., and Wang, H. J. (2010). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, 11(2):337–352.
- Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, 106(493):6–20.
- Si, Y., Pillai, N. S., and Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10(3):605–625.
- Sriram, K. (2015). A sandwich likelihood correction for bayesian quantile regression based on the misspecified asymmetric laplace density. *Statistics & Probability Letters*, 107:18–26.
- Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614.
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.
- Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314.

- White, H. (1984). *Asymptotic theory for econometricians*. Academic press.
- Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40(2):1102–1131.
- Yang, Y. and Tokdar, S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, 112(519):1107–1120.
- Yang, Y., Wang, H. J., and He, X. (2016). Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. *International Statistical Review*, 84(3):327–344.
- Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.
- Yu, K. and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, 34(9-10):1867–1879.

Chapter 3

A Leave-one-out Information Criterion for Multilevel Models

3.1 Introduction

Cross validation and information criteria (IC) are two sets of methods to assist with model selection. Here we focus on one of the most popular IC, the Akaike information criterion (AIC) proposed by Akaike (1974) that shares the goal of cross validation to identify a model with the best out-of-sample predictive accuracy. Although the AIC and cross-validation are presented in different forms and originated from different literatures, many formal equivalence relations have been established. For example, asymptotically, AIC is equivalent to leave-one-out (LOO) cross validation (Stone, 1977).

AIC, like most of other IC, consists of two parts. One part is an in-sample estimate of prediction error and another part is “Optimism”, a penalty term that corrects for using the same data for both model fitting and error estimation. The AIC was initially proposed for independent data (also referred to as unclustered data in the following text). In many realistic applications, data are clustered (e.g., patients nested in hospitals, students nested in classrooms, repeated observations on individuals in longitudinal data) and naively applying the AIC for unclustered data may not be justified. Statistical models that include random effects or latent variables to induce dependence among units belonging to the same cluster, such as linear mixed models (Laird and Ware, 1982), or in general, latent variable models (Skron dal and Rabe-Hesketh, 2004), pose additional challenges because the out-of-sample prediction error can be defined conditionally, given the random effects/latent variables for the clusters in the data, or marginally.

Vaida and Blanchard (2005) propose the conditional AIC (cAIC) and discuss the difference between marginal AIC (mAIC) and cAIC in terms of different purposes of model comparison. For linear mixed models, they provide analytic formulas for the penalty term of

the cAIC involving estimates of variance-covariance parameters. Unfortunately, such derivations are not straightforward for other generalized linear mixed-effects models and would require new derivations whenever a new model is proposed. Another downside is that predictive error is defined as the log density (or log likelihood), whereas other scoring rules may be more appropriate. Gneiting (2011) provides a survey on the use of different scoring rules, such as the squared error and absolute error, that may be more appropriate in practice depending on the nature of the models and research goals.

LOO cross validation, on the other hand, can easily be applied for new models and any kind of scoring rule, such as mean squared error or log predictive density (for an overview, see, e.g., Arlot and Celisse, 2010; Gelman et al., 2014). However, because it requires refitting the model n times, where n is the sample size, LOO can be practically prohibitive. For clustered data, another issue with LOO is how to decide whether to leave individual *units* out or whether to leave entire clusters out, an issue closely related to the decision whether to use cAIC or mAIC.

How to approximate LOO in a computationally-efficient way has been extensively studied in the literature (e.g., Golub et al., 1979; Pregibon, 1981; Verweij and Van Houwelingen, 1993; Xiang and Wahba, 1996; Cawley and Talbot, 2003, 2004, 2008; Vehtari et al., 2017; Wang et al., 2018; Giordano et al., 2019b,a; Rad and Maleki, 2020). In this chapter, we aim to explore the middle ground between LOO and AIC using *approximate* LOO parameter estimates to obtain an alternative approximation to the target quantity for which the AIC is an approximation. We derive approximations for both unclustered data and clustered data, that require much less computational cost compared to exact calculation while being more flexible, more intuitive, and potentially more precise than the AIC under certain conditions. We conduct a series of simulation studies to empirically evaluate the approximation and compare its use for model selection with existing methods. We emphasize the applicability of the approach to different models and scoring rules. The approximation is practically feasible to implement and can be tailored for different inferential goals for new models.

Section 3.2 discusses the connection between AIC and LOO cross validation for unclustered data, making use of results in Akaike (1974), Stone (1977), and Anderson and Burnham (2004). In Section 3.3 we then discuss approximating the LOO parameter estimates based on Giordano et al. (2019b)’s Swiss army jackknife approximation, which dates back to the infinitesimal jackknife by Jaeckel (1972). This naturally leads to an approximate LOO-based information criterion that we refer to as LOOIC (for LOO information criterion, an acronym previously used by Vehtari et al. (2017) in the Bayesian model evaluation literature). In Section 3.4, we perform simulations to compare both LOOIC and AIC with the target quantity that AIC approximates in terms of bias, mean-squared error and model selection performance. In Section 3.5 we discuss two different versions of AIC for clustered data (cAIC and mAIC) targeting different inferential goals and introduce corresponding extensions of LOOIC, an approximated leave-one-*unit*-out information criterion (LOuOIC) and an approximated leave-one-*cluster*-out information criterion (LOcOIC). Section 3.6 compares LOcOIC

with its exact counterpart and mAIC for multilevel linear models through a simulation study. In Section 3.7, we conclude with possible future directions, especially the opportunity and challenges of using LOuOIC.

3.2 Connection between AIC and LOO

3.2.1 AIC target quantity

To assess a model's performance, Akaike (1974) defines a good model to be a candidate model $f_1(\theta_1)$ (belonging to a family of models $\mathcal{F}$) that is “close” to the true model g in terms of the Kullback-Leibler (KL) distance. The KL distance can be written in terms of the densities $g(y)$ and $f_1(y|\theta_1)$ of datasets y for the true and candidate models, respectively, as

$$\begin{aligned} \text{KL}(g, f_1(\theta_1)) &= \int g(y) \log \left(\frac{g(y)}{f_1(y | \theta_1)} \right) dy \\ &= \underbrace{\int g(y) \log g(y) dy}_{\text{constant}} - \underbrace{\int g(y) \log f_1(y | \theta_1) dy}_{\text{relative KL information}}. \end{aligned} \quad (3.1)$$

The KL distance is non-negative and is zero if and only if $g = f_1$ almost everywhere, so a smaller KL distance means a better candidate model. Of the models in $\mathcal{F}$, the best model is the one that minimizes the KL distance, and the corresponding KL distance may or may not be zero, depending on whether the true model g is in $\mathcal{F}$. The first term of equation (3.1) depends on the true model and is hence not estimable. However, because this term does not depend on the candidate model, we can focus on the second term, often called “relative KL information” and define the best model as the one that minimizes the relative KL information, which can also be written as $-E_y \log f(y | \theta)$. Note that, for a misspecified candidate model f_1 , where there is no such thing as the “true” parameter values, the “pseudo” true values are defined as the parameter values that minimize the relative KL information. The maximum likelihood estimator of the parameters $\hat{\theta}_1$ of a candidate model f_1 minimizes the relative KL information asymptotically.

This relative KL distance enables us to compare models within a model family and across model families. If we have two candidate models $f_1(\theta_1)$ and $f_2(\theta_2)$ (with “pseudo” true values θ_1 and θ_2), then the difference between their KL distances to the true model, $\text{KL}(g, f_1(\theta_1)) - \text{KL}(g, f_2(\theta_2))$, is equal to the difference in the corresponding relative KL information, $E_y \log f_2(y | \theta_2) - E_y \log f_1(y | \theta_1)$. A negative difference indicates that f_1 is a better approximation to the true model than f_2 . When comparing multiple models, we define the best model to be the one with the smallest relative KL information, $-E_y \log(f(y | \theta))$.

In practice, we do not know the pseudo true parameter values, but for given data y , we can obtain maximum likelihood estimates $\hat{\theta}_1$ for a given model f_1 . The quantity used for

model comparison could therefore be defined as

$$-E_y \log f(y \mid \hat{\theta}(y)).$$

However, if f_1 is nested in f_2 , it is well-known that the larger model will always have a larger maximized log-likelihood than the smaller model, so the criterion above would always select the larger model. The problem is that the same data is used twice, once for estimating parameters and then for model selection. To prevent this from happening, we can instead consider estimating parameters from other data y' independent from y , and taking the expectation, $-E_{y'} E_y \log f(y \mid \hat{\theta}(y'))$. The target quantity for the AIC is twice this expected relative KL information,

$$-2E_{y'} E_y \log f(y \mid \hat{\theta}(y')). \quad (3.2)$$

However, since the expectations are over the true distribution $g(y)$, it is not possible to obtain this quantity in practice. One approximation of this quantity is the AIC.

3.2.2 AIC as an approximation of the target quantity

Following the derivation steps in Chapter 7 of Anderson and Burnham (2004), we now derive an approximation for the expected relative KL information in (3.2). Consider a second-order Taylor expansion of $\log f(y \mid \hat{\theta}(y'))$ around θ_0 , the “pseudo” true parameter values of the candidate model discussed in the previous section (i.e., the parameter values minimizing the KL divergence to the truth, or we could call it the KL minimizer, for a given candidate model):

$$\begin{aligned} \log f(y \mid \hat{\theta}(y')) &\approx \log f(y \mid \theta_0) + \left(\frac{\partial \log f(y \mid \theta_0)}{\partial \theta} \right)^\top (\hat{\theta}(y') - \theta_0) \\ &\quad + \frac{1}{2} (\hat{\theta}(y') - \theta_0)^\top \frac{\partial^2 \log f(y \mid \theta_0)}{\partial \theta \partial \theta^\top} (\hat{\theta}(y') - \theta_0) \end{aligned} \quad (3.3)$$

To get closer to the target quantity in equation (3.2), we first take the expectation over y , recognizing that the second term in the expansion above drops out because the expectation of the first partial derivatives of the log-likelihood evaluated at the “pseudo” true parameters θ_0 are zero:

$$\begin{aligned} E_y \log f(y \mid \hat{\theta}(y')) &\approx E_y \left\{ \log f(y \mid \theta_0) + \frac{1}{2} (\hat{\theta}(y') - \theta_0)^\top \frac{\partial^2 \log f(y \mid \theta_0)}{\partial \theta \partial \theta^\top} (\hat{\theta}(y') - \theta_0) \right\} \\ &= E_y \log f(y \mid \theta_0) - \frac{1}{2} (\hat{\theta}(y') - \theta_0)^\top I(\theta_0) (\hat{\theta}(y') - \theta_0), \end{aligned} \quad (3.4)$$

where $I(\theta_0)$ is minus the expectation of the negative Hessian evaluated at the “pseudo” true parameter values (hence the sign switch from the first step to the second step). To get even

closer to the target quantity in equation (3.2), we again take the expectation of (3.4), this time over y' :

$$\begin{aligned}
 E_{y'} E_y \log f(y \mid \hat{\theta}(y')) &\approx E_{y'} \left[E_y \log f(y \mid \theta_0) - \frac{1}{2} \left(\hat{\theta}(y') - \theta_0 \right)^\top I(\theta_0) \left(\hat{\theta}(y') - \theta_0 \right) \right] \\
 &= E_y \log f(y \mid \theta_0) - \frac{1}{2} \text{trace} \left\{ I(\theta_0) \underbrace{E_{y'} \left[\left(\hat{\theta}(y') - \theta_0 \right)^\top \left(\hat{\theta}(y') - \theta_0 \right) \right]}_{=\Sigma} \right\} \\
 &= E_y \log f(y \mid \theta_0) - \frac{1}{2} \text{trace} \{ I(\theta_0) \Sigma \}, \tag{3.5}
 \end{aligned}$$

where Σ is the sandwich covariance matrix (i.e., the large sample covariance matrix of the MLE of θ for the candidate model under possible model misspecification).

Now consider a second-order Taylor expansion of $\log f(y \mid \theta_0)$ around $\hat{\theta}(y)$,

$$\begin{aligned}
 \log f(y \mid \theta_0) &\approx \log f(y \mid \hat{\theta}(y)) + \left(\frac{\partial \log f(y \mid \hat{\theta}(y))}{\partial \theta} \right)^\top (\theta_0 - \hat{\theta}(y)) \\
 &\quad + \frac{1}{2} (\theta_0 - \hat{\theta}(y))^\top \frac{\partial^2 \log f(y \mid \hat{\theta}(y))}{\partial \theta \partial \theta^\top} (\theta_0 - \hat{\theta}(y)). \tag{3.6}
 \end{aligned}$$

Taking the expectation over y , recognizing that the second term is zero because the partial derivatives of the log-likelihood are zero at the maximum likelihood estimates, gives

$$\begin{aligned}
 E_y \log f(y \mid \theta_0) &\approx E_y \left\{ \log f(y \mid \hat{\theta}(y)) + \frac{1}{2} (\theta_0 - \hat{\theta}(y))^\top \frac{\partial^2 \log f(y \mid \hat{\theta}(y))}{\partial \theta \partial \theta^\top} (\theta_0 - \hat{\theta}(y)) \right\} \\
 &= E_y \log f(y \mid \hat{\theta}(y)) - \frac{1}{2} \text{trace} \left\{ E_y \left[I(\hat{\theta}(y)) \left(\hat{\theta}(y) - \theta_0 \right)^\top \left(\hat{\theta}(y) - \theta_0 \right) \right] \right\} \\
 &\approx E_y \log f(y \mid \hat{\theta}(y)) - \frac{1}{2} \text{trace} \left\{ I(\theta_0) \underbrace{E_y \left[\left(\hat{\theta}(y) - \theta_0 \right)^\top \left(\hat{\theta}(y) - \theta_0 \right) \right]}_{=\Sigma} \right\} \\
 &= E_y \log f(y \mid \hat{\theta}(y)) - \frac{1}{2} \text{trace} \{ I(\theta_0) \Sigma \}. \tag{3.7}
 \end{aligned}$$

Substituting (3.7) into the first term of equation (3.5), we obtain

$$\begin{aligned}
 E_{y'} E_y \log f(y \mid \hat{\theta}(y')) &\approx E_y \log f(y \mid \hat{\theta}(y)) - \text{trace} \{ I(\theta_0) \Sigma \} \\
 &\approx \log f(y \mid \hat{\theta}(y)) - K, \tag{3.8}
 \end{aligned}$$

where K is the number of model parameters,

This gives the AIC formula as an estimator of twice the expected relative KL information

$$-2E_{y'}E_y \log f(y \mid \hat{\theta}(y')) \approx -2\log f(y \mid \hat{\theta}(y)) + 2K = \text{AIC}. \quad (3.9)$$

Note that increasing the dimension K leads to more complex models that tend to lead to a decrease in $\log f(y \mid \hat{\theta}(y))$, or so called in-sample error, but the penalty term $2K$ counteracts this decrease to prevent selection of a model that over-fits the data.

It is instructive to take a closer look at the trace term in (3.8) that is approximated by K in the last line. Inside the trace, $I(\theta_0)$ is the inverse of the model-based covariance matrix and is correct only when there is no misspecification, while Σ is the true sandwich covariance matrix (robust to model misspecification). The second approximation is therefore exact when the candidate model is correct so that $I(\theta_0)\Sigma$ is a $K \times K$ identity matrix, with trace equal to K . We can therefore see that the approximation assumes that we have a “good” candidate model (in the KL sense). In defense of using the approximation for model comparison, a “bad” model (in the KL sense) will likely have a larger AIC than well-fitting models based in its dominant (large) first term even though the second term is poorly approximated.

If we directly estimate the trace term (e.g., using resampling methods such as bootstrap or jackknife) instead of using K , we obtain Takeuchi’s information criterion (TIC, Takeuchi, 1976; Shibata, 1989; Anderson and Burnham, 2004), which is a generalization of AIC. The analogy here is similar to relying on the model-based standard errors versus using the sandwich estimator (or resampling methods). For clustered data, a similar analogy can be drawn with model-based standard errors, cluster resampling methods, as well as cluster-robust standard errors.

3.2.3 Exact LOO information criterion (eLOOIC)

The target quantity for the AIC, or twice the expected relative KL information, given in equation (3.2),

$$-2E_{y'}E_y \log f(y \mid \hat{\theta}(y')),$$

can be viewed also from a prediction perspective if we think of y' as the training data used for estimating the parameters and of y as future data or test data used to evaluate the predictions.

In the absence of independent training and test data, LOO cross validation can be used to approximate this quantity. LOO treats the data with the i th unit removed, $y_{(-i)}$, as the “training data” y' for parameter estimation and y_i as the “test data” y for each unit’s log-likelihood contribution. The target quantity is therefore approximated by

$$-2E_{y_{(-i)}}E_{y_i} \sum_{i=1}^n \log f(y_i \mid \hat{\theta}(y_{(-i)})), \quad (3.10)$$

where the reason for the sum is that

$$\log f(y|\theta) = \sum_{i=1}^N \log f(y_i|\theta). \quad (3.11)$$

Correspondingly, we can define the LOO approximation to the target quantity as eLOOIC (“e” for “exact” in anticipation of an approximation introduced below):

$$\text{eLOOIC} = -2 \sum_{i=1}^n \log f(y_i|\hat{\theta}_{(-i)}). \quad (3.12)$$

The eLOOIC was referred to as “cross-validated log-likelihood (cvl)” in the equation (1) by Verweij and Van Houwelingen (1993). The function we apply here for each unit is the log density (log likelihood). However, many other functions can be used for different purposes as discussed by Gneiting and Raftery (2007) and Gneiting (2011).

Throughout the text, we implicitly condition on covariates x for regression models and omit covariates x from our notation for convenience – see recent work by Bates et al. (2021) for further discussion of random predictors.

3.2.4 Asymptotic equivalence of AIC and eLOOIC

Here we follow Stone (1977) in showing the asymptotic equivalence of AIC and eLOOIC.

The goal is to derive an expression for $\log f(y_i|\hat{\theta}_{(-i)})$ in (3.12). We start by taking a first-order Taylor expansion of $\log f(y_i|\hat{\theta}_{(-i)})$ around $\hat{\theta}$,

$$\log f(y_i|\hat{\theta}_{(-i)}) \approx \log f(y_i|\hat{\theta}) + (\hat{\theta}_{(-i)} - \hat{\theta})^\top \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta}. \quad (3.13)$$

Summing over the units i and from (3.11), we obtain

$$\begin{aligned} \sum_{i=1}^n \log f(y_i|\hat{\theta}_{(-i)}) &\approx \sum_{i=1}^n \log f(y_i|\hat{\theta}) + \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta})^\top \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta} \\ \log f(y|\hat{\theta}_{(-i)}) &\approx \log f(y|\hat{\theta}) + \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta})^\top \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta}. \end{aligned} \quad (3.14)$$

To derive an expression for $(\hat{\theta}_{(-i)} - \hat{\theta})^\top$ in (3.14), we take a first-order Taylor expansion of $\frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta}$ around $\hat{\theta}$,

$$\frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta} \approx \frac{\partial \log f(y|\hat{\theta})}{\partial \theta} + \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} (\hat{\theta}_{(-i)} - \hat{\theta})$$

$$= \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} (\hat{\theta}_{(-i)} - \hat{\theta}), \quad (3.15)$$

where the first term disappears because the gradients of the full-sample log-likelihood are zero at the full-sample MLEs. Therefore,

$$\begin{aligned} (\hat{\theta}_{(-i)} - \hat{\theta})^\top &\approx \left(\left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} \frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta} \right)^\top \\ &= \left(\frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta} \right)^\top \left(\left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} \right)^\top \\ &= \left(\frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta} \right)^\top \left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1}. \end{aligned} \quad (3.16)$$

For the last equality we made use of

$$\frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta} = \frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta}. \quad (3.17)$$

This equality can be derived by noting from (3.11) that

$$\log f(y_{(-i)}|\theta) = \log f(y|\theta) - \log f(y_i|\theta).$$

Taking derivatives and evaluating them at $\theta = \hat{\theta}_{(-i)}$, we obtain

$$\frac{\partial \log f(y_{(-i)}|\hat{\theta}_{(-i)})}{\partial \theta} = \frac{\partial \log f(y|\hat{\theta}_{(-i)})}{\partial \theta} - \frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta}.$$

The left-hand-side of the above equation is zero because the gradient of the leave-one-out log-likelihood is zero at the leave-one-out MLEs, so we obtain the equality in (3.17).

Following Stone (1977), we substitute equation (3.16) into equation (3.14):

$$\begin{aligned} \log f(y|\hat{\theta}_{(-i)}) &= \sum_{i=1}^n \log f(y_i|\hat{\theta}_{(-i)}) \\ &\approx \log f(y|\hat{\theta}) + \sum_{i=1}^n \left(\frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta} \right)^\top \left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta}. \end{aligned} \quad (3.18)$$

Asymptotically and under weak conditions, both $\hat{\theta}$ and $\hat{\theta}_{(-i)}$ converge to the “pseudo” true value θ_0 , $\frac{1}{n} \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} = \frac{1}{n} \frac{\partial^2 \sum_{i=1}^n \log f(y_i|\hat{\theta})}{\partial \theta \partial \theta^\top}$ converges in probability to $E \left\{ \frac{\partial^2 \log f(y_i|\theta_0)}{\partial \theta \partial \theta^\top} \right\} = -\frac{1}{n} I(\theta_0)$

and

$\frac{1}{n} \sum_i^n \frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta} \left(\frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta} \right)^\top$ converges in probability to $E \left\{ \frac{\partial \log f(y_i|\theta_0)}{\partial \theta} \left(\frac{\partial \log f(y_i|\theta_0)}{\partial \theta} \right)^\top \right\} = \frac{1}{n} J(\theta_0)$, so

$$\log f(y|\hat{\theta}_{(-i)}) \approx \log f(y|\hat{\theta}) + \text{trace} \left(-J(\theta_0) I(\theta_0)^{-1} \right) \quad (3.19)$$

$$= \log f(y|\hat{\theta}) - \text{trace} (I(\theta_0) \Sigma). \quad (3.20)$$

Here we used the property of trace and the asymptotic results to obtain (3.19) and the sandwich equality $\Sigma = I(\theta_0)^{-1} J(\theta_0) I(\theta_0)^{-1}$ (see, e.g., (7.14) in Anderson and Burnham, 2004) to obtain equation (3.20). The trace part is identical to that in equation (3.8) derived from the KL perspective. We have hence shown that twice the eLOOIC is asymptotically equivalent to the Takeuchi's Information Criterion (TIC).

The same logic as in Section 3.2.2 can be followed to approximate the trace by K by assuming that the candidate model is close to the true model. With this assumption, it follows from (3.20) that

$$-2 \log f(y|\hat{\theta}_{(-i)}) \approx -2 \log f(y|\hat{\theta}) - 2K = \text{AIC}. \quad (3.21)$$

3.3 Approximation for eLOOIC

From equation (3.16), it is easy to see that we derived an expression of a convenient approximation for the LOO estimate $\hat{\theta}_{(-i)}$. Rearranging terms and writing it below again for convenience,

$$\hat{\theta}_{(-i)} \approx \hat{\theta} + \left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} \left(\frac{\partial \log f(y_i|\hat{\theta}_{(-i)})}{\partial \theta} \right). \quad (3.22)$$

The approximation is exact for linear Gaussian models. For example, consider a multiple linear regression model,

$$y_i = x_i^\top \beta + \epsilon_i,$$

where $x_i \in \mathbb{R}^K$ is the covariate vector (including a constant for the intercept) for unit i and $\beta \in \mathbb{R}^K$ is the associated regression coefficient vector, and the independently identically distributed error terms are drawn from $N(0, 1)$ for each of the observation. It is easy to verify that equation (3.22) can be simplified to (for its regression coefficient vector)

$$\hat{\beta}_{(-i)} = \hat{\beta} + \left(\sum_i^n x_i x_i^\top \right)^{-1} x_i \hat{\epsilon}_{(-i)} = \hat{\beta} + \left(\sum_i^n x_i x_i^\top \right)^{-1} x_i \frac{\hat{\epsilon}_i}{1 - h_{ii}},$$

where $\hat{\epsilon}_{(-i)} = y_i - x_i^\top \beta_{(-i)}$ is the LOO predicted residual, and $h_{ii} = x_i^\top (\sum_i^n x_i x_i^\top)^{-1} x_i$ is the i^{th} diagonal element of the hat matrix, often referred to as leverage scores, a classic result in the literature (e.g., Lemma 3.2, Miller Jr, 1974).

Since the goal, for general cases, is to avoid estimating $\hat{\theta}_{(-i)}$, we cannot use $\hat{\theta}_{(-i)}$ within the approximation in (3.22). However, we know from the last section that under regularity conditions both estimates converges to θ_0 and also that the estimates will be similar when the sample size is large, so we replace $\hat{\theta}_{(-i)}$ with $\hat{\theta}$,

$$\hat{\theta}_{(-i)} \approx \hat{\theta} + \left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} \left(\frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta} \right). \quad (3.23)$$

The approximation in equation (3.23) and various extensions of it have been well known for a long time in the literature with numerous applications ranging from influence statistics (Cook and Weisberg, 1982), survival analysis (equation (5), Verweij and Van Houwelingen, 1993), ridge regressions (equation (6), Meijer and Goeman, 2013), to traditional machine learning (Cawley and Talbot, 2004), and it has been utilized and extended in multiple recent papers in various contexts as well.

Most relevant to our work, Commenges et al. (2015) propose a general approximate cross-validation criterion (UACVR) based on a similar approximation (together with equation (3.18) and (3.19)), for various estimation schemes and loss functions, and demonstrate the exact equivalence of the UACVR with the normalized TIC and approximate equivalence with the normalized AIC (i.e., $\text{UACVR} = \frac{1}{2n} \text{TIC} \approx \frac{1}{2n} \text{AIC}$) when MLE is used for estimation and the log-likelihood is used for model evaluation. However, it seems more meaningful to stay on the original scale of the AIC and the link between LOO and AIC seems crucial to discuss as the primary foundation and reasoning basis for model evaluation and comparison. So below we provide an approximated LOO criterion (LOOIC) starting with the approximate LOO parameter estimates instead of directly approximating the information criterion. With similar reasoning, Lumley and Scott (2015) extend AIC to the context of survey sampling.

Giordano et al. (2019b)'s "Swiss army Infinitesimal Jackknife (IJ) approximation" can also be viewed as an extension of equation (3.23) to more general settings with different weighting schemes when evaluating machine learning models with the derivatives calculated using modern auto-differentiation methods (Baydin et al., 2018), although it was motivated from a direct approximation for $\hat{\theta}(w)$, the parameter estimate for a weighted dataset using the weighting vector w , around $\hat{\theta}$ (using their notation):

$$\hat{\theta}^{\text{IJ}}(w) := \hat{\theta}(1_n) + \left\{ \frac{\partial^2 \log f(y|\hat{\theta})}{\partial \theta \partial \theta^\top} \right\}^{-1} S(w - 1_n), \quad (3.24)$$

where $S \in \mathbb{R}^{K \times n}$ is a convenient notation for a matrix whose i th column contains the derivatives $\frac{\partial \log f(y_i|\hat{\theta})}{\partial \theta}$ of unit i 's contribution to the objective function with respect to the

parameters, evaluated at the full sample estimate $\hat{\theta}$. 1_n is an all one vector. It is easy to verify that equation (3.24) reduces to (3.23) in this scenario with weighting scheme being set to LOO (i.e., set $w = 1_{(-i)}$, where $1_{(-i)}$ is an all one vectors except 0 for unit i). We use the notation $\hat{\theta}_{(-i)}^{\text{IJ}} := \hat{\theta}^{\text{IJ}}(1_{(-i)})$ for the Infinitesimal Jackknife (IJ) approximation to the Loo estimator. As noted in Giordano et al. (2019b), the approximation works under a wide range of conditions, as we will illustrate for a simple example below.

We can use the IJ approximation in (3.24) to obtain an approximate version of the eLOOIC in equation (3.12). Whereas eLOOIC is the “exact” version, we will call the approximate version simply LOOIC,

$$\text{LOOIC} = -2 \sum_i^N \log f(y_i | \hat{\theta}_{(-i)}^{\text{IJ}}). \quad (3.25)$$

Compared to Commenges et al. (2015) who directly target and generalize TIC, the benefit of the above definition is its simplicity and the intuition that comes from targeting eLOOIC and utilizing the asymptotic equivalence between eLOOIC and AIC, both of which are widely known concepts for applied researchers. The name of “LOOIC” (as well as following extensions for clustered data) is inspired by the Bayesian approximated LOO counterpart in Vehtari et al. (2017), which uses Bayesian LOO posterior estimates based on importance sampling. This Bayesian counterpart has a close relation with the Watanabe–Akaike information criterion (WAIC, Watanabe and Oppen, 2010), as discussed by Merkle et al. (2019). Another advantage of this definition is that it allows us to define approximate information criteria for models that account for more complex data structure beyond simple unclustered data, as we will discuss in Section 3.5.

As a side note, the utility of the LOOIC depends on how well the IJ approximation works. We use the IJ approximation mainly for its convenience in implementation, and there are many more approximation methods available across disciplines (e.g., Obuchi and Kabashima, 2016; Wang et al., 2018; Giordano et al., 2019a; Rad and Maleki, 2020) that differ in terms of assumptions, computational burden, and accuracy.

Simulation to evaluate IJ approximation of LOO estimates We simulated 100 observations from a Gaussian distribution with $\mu = 1$ and $\sigma = 2$ and obtained both the exact LOO estimates, $\hat{\mu}_{(-i)}$ and $\hat{\sigma}_{(-i)}$, of the mean and standard deviation, and the corresponding IJ approximations, $\hat{\mu}_{(-i)}^{\text{IJ}}$ and $\hat{\sigma}_{(-i)}^{\text{IJ}}$. The left column of Figure 1 shows the results for the mean and the right column for the standard deviation. In the second row, we show results for leaving *two* units out. It is easy to see that, for this simple case, if we consider reweighting schemes that leave more than one observation out (the scenario more similar to the clustered data case that will be discussing later), the IJ approximation still works quite well.

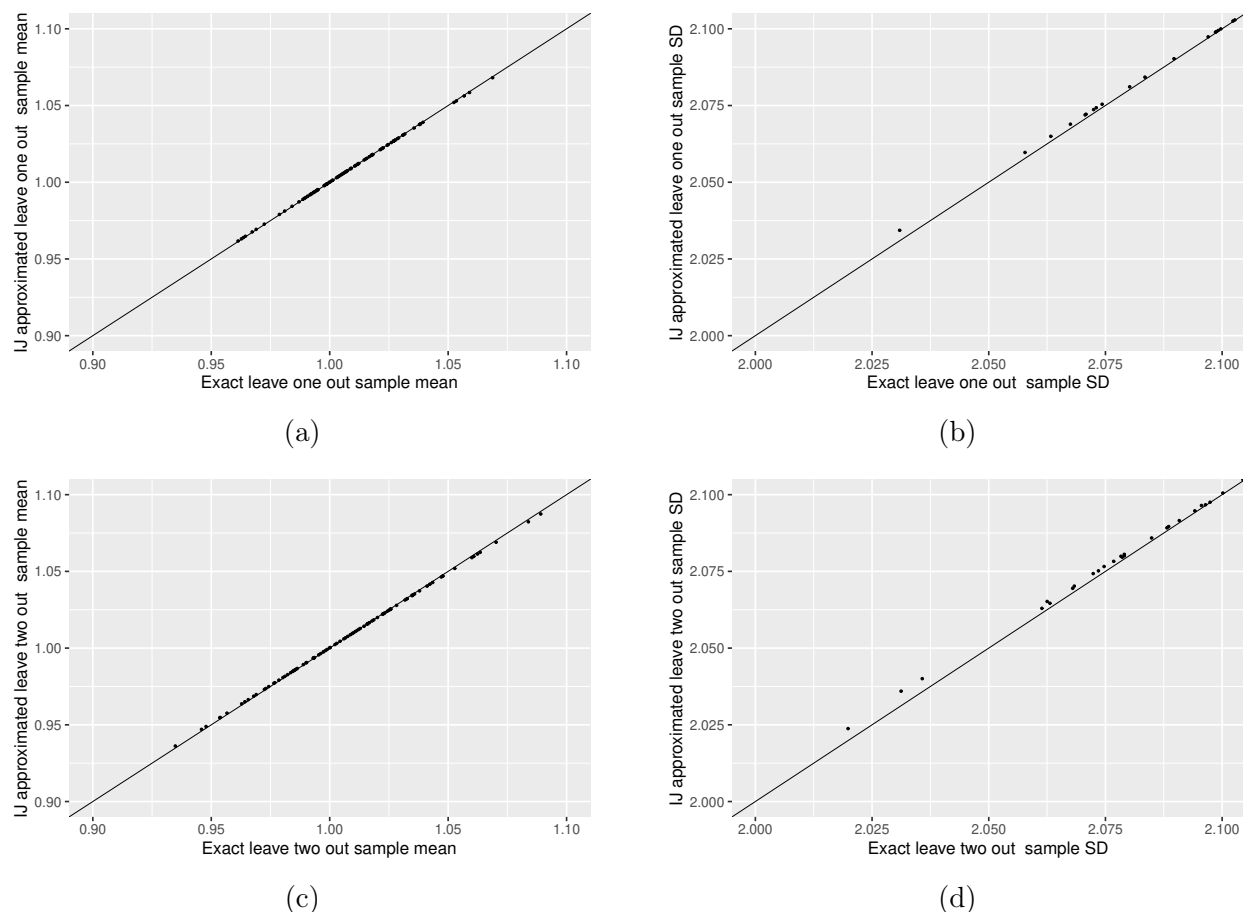


Figure 3.1: (a) Comparison between exact LOO sample mean and its IJ approximation; (b) Comparison between exact LOO sample standard deviation and its IJ approximation; (c) Comparison between exact leave-two-out sample mean and its IJ approximation; (d) Comparison between exact leave-two-out sample standard deviation and its IJ approximation.

3.4 Study I: comparing LOOIC, eLOOIC and AIC with the target quantity

There is an extensive literature on empirically assessing model comparison and selection performance using different types of cross validation and information criteria. In this simulation, we do not seek to provide a comprehensive picture of how well LOOIC performs, but aim to show that it has a comparable performance with AIC in some situations and therefore could be part of the toolkit, also potentially for more complex models beyond the unclustered data case.

We follow the simulation design from Example 1 in Section 2 of Kuha (2004) which is

also similar in spirit to Shao (1993).

3.4.1 Design

The data-generating model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

with $\epsilon_i \sim N(0, 1)$, and different choices, such as (0,1,0) or (-1,0,1) for the intercept and two regression coefficients. Here the scalar x_i was kept fixed across replications and was set to $\{0, \frac{1}{N-1}, \dots, \frac{N-2}{N-1}, 1\}$, for sample sizes $N = 50, 100, 500, 1000$. For each simulation condition, two models were estimated, one null model with only an intercept (denoted M1) and another model with an intercept as well as one predictor x_i (denoted M2).

We use Monte Carlo integration with 1,000,000 samples for the target quantity, twice the expected relative KL information, $-2E_{y'}E_y \log f(y | \theta(y'))$. Specifically, we simulated a million pairs of datasets, y and y' , of size N and took the average of $-2 \log f(y | \hat{\theta}(y'))$. This allows us to estimate the bias and root mean squared error of LOOIC, eLOOIC, and AIC as estimates of the target quantity. We also compare empirical model selection performance (i.e., proportion of replicates where the correct model is selected). The correct model is defined as the true model when the true model is among the candidate models or as the better performing model in terms of the target quantity when the candidate models do not include the true model.

We simulate 500 replicates for each simulation condition. The sample mean of the AIC, eLOOIC, and LOOIC, as well as the true target quantity are summarised in Table 3.1. Here the four data-generating models are defined by the values of $(\beta_1, \beta_2, \beta_3)$ given under “Parameters.” The corresponding estimated bias and estimated root mean squared errors are summarised in Figures 3.2 and 3.3.

3.4.2 Results

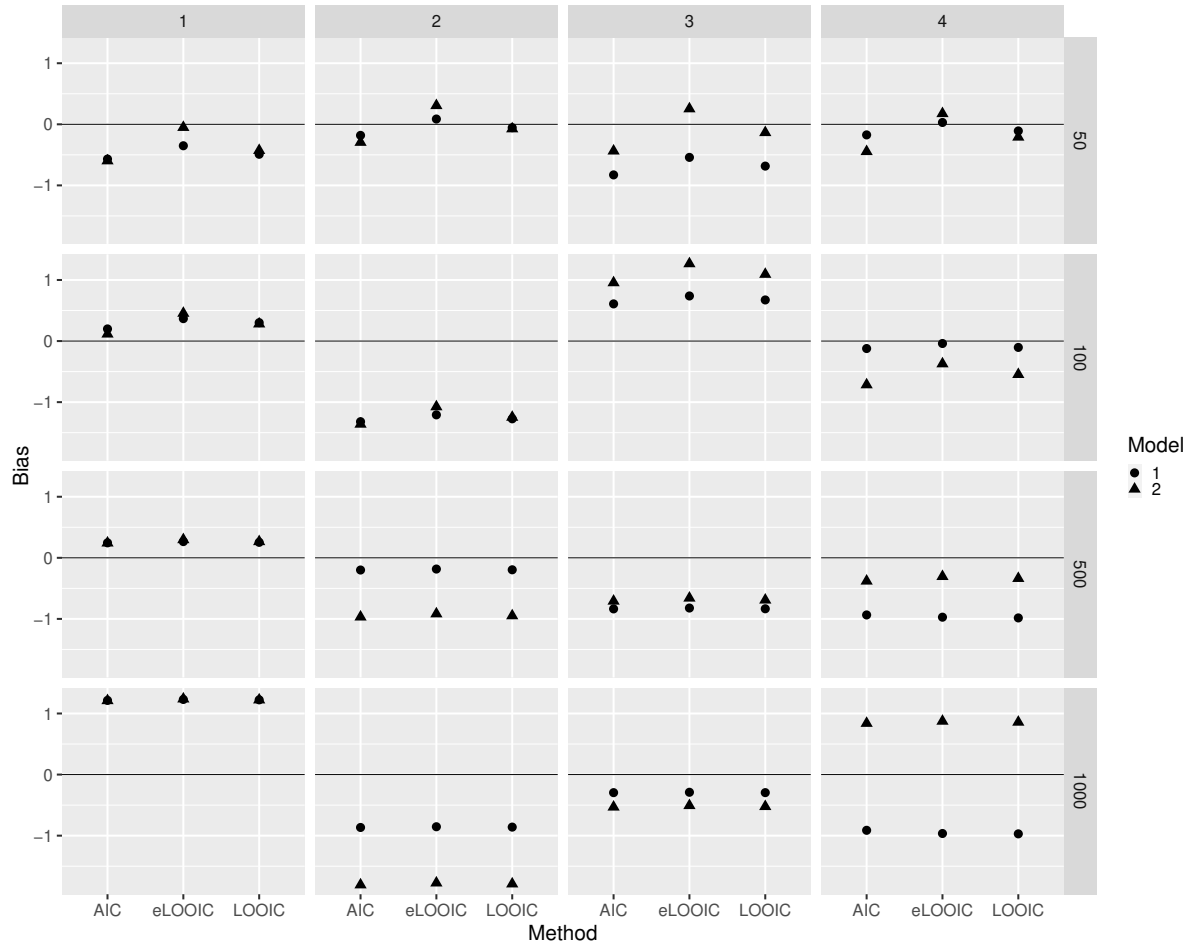


Figure 3.2: Finite sample bias estimates for AIC, eLOOIC, and LOOIC. Each row represents a different sample size while each column represents a different data generating processes.

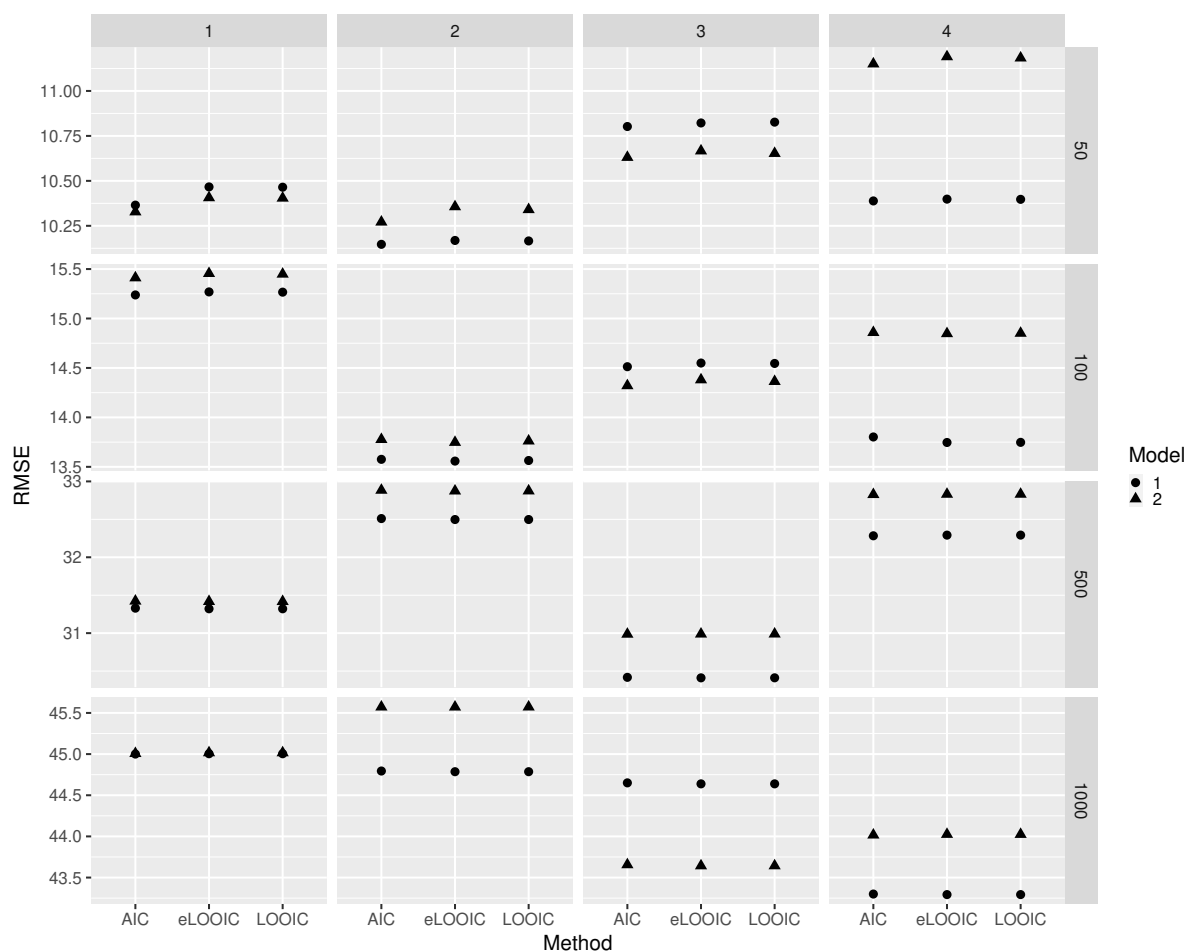


Figure 3.3: Root mean squared error estimates for AIC, eLOOIC, and LOOIC. Each row represents a different sample size while each column represents a different data generating processes.

Table 3.1: Mean AIC, eLOOIC, and LOOIC estimates, proportion correct, and target quantity.

Sample Size	Parameters	AIC			eLOOIC			LOOIC			Target Quantity		
		M1	M2	Correct	M1	M2	Correct	M1	M2	Correct	M1	M2	Model [†]
N = 50	1 (0, 0, 0)	143.477	144.516	0.852	143.695	145.063	0.872	143.556	144.686	0.850	144.046	145.114	M1
	2 (0, 1, 0)	147.903	144.796	0.714	148.171	145.398	0.694	148.029	145.017	0.716	148.084	145.091	M2
	3 (-1, 0, 1)	147.549	144.960	0.706	147.836	145.651	0.656	147.693	145.262	0.680	148.377	145.397	M2
	4 (-1, 1, 1)	158.606	144.945	0.996	158.809	145.567	0.994	158.671	145.182	0.996	158.779	145.391	M2
N = 100	1 (0, 0, 0)	286.081	287.031	0.826	286.250	287.371	0.836	286.184	287.196	0.832	285.883	286.915	M1
	2 (0, 1, 0)	292.601	285.521	0.940	292.713	285.806	0.924	292.648	285.635	0.928	293.921	286.882	M2
	3 (-1, 0, 1)	295.075	288.423	0.910	295.205	288.734	0.904	295.140	288.561	0.910	294.467	287.468	M2
	4 (-1, 1, 1)	315.097	286.740	1	315.179	287.084	1	315.115	286.909	1	315.219	287.456	M2
N = 500	1 (0, 0, 0)	1421.181	1422.185	0.834	1421.201	1422.239	0.838	1421.189	1422.207	0.838	1420.935	1421.942	M1
	2 (0, 1, 0)	1460.805	1420.964	1	1460.821	1421.015	1	1460.809	1420.984	1	1461.004	1421.932	M2
	3 (-1, 0, 1)	1462.854	1424.112	1	1462.868	1424.162	1	1462.855	1424.130	1	1463.690	1424.821	M2
	4 (-1, 1, 1)	1566.153	1424.374	1	1566.117	1424.449	1	1566.105	1424.417	1	1567.089	1424.755	M2
N = 1000	1 (0, 0, 0)	2841.082	2842.078	0.848	2841.098	2842.106	0.844	2841.092	2842.090	0.844	2839.868	2840.867	M1
	2 (0, 1, 0)	2919.062	2839.026	1	2919.074	2839.057	1	2919.068	2839.041	1	2919.928	2840.831	M2
	3 (-1, 0, 1)	2924.813	2845.880	1	2924.819	2845.905	1	2924.813	2845.889	1	2925.109	2846.412	M2
	4 (-1, 1, 1)	3130.957	2847.216	1	3130.905	2847.251	1	3130.899	2847.235	1	3131.870	2846.377	M2

[†] Best model according to target quantity

For the smallest sample size, $N = 50$, the estimated bias is smaller for LOOIC and eLOOIC than for AIC across different data-generating processes. This could be due to the fact that the AIC relies on more asymptotic approximations. However, for this sample size, the estimated root mean squared error is smaller for AIC than for eLOOIC and LOOIC across different data-generating processes, suggesting that the AIC has a smaller variance. This could be because the second term of the AIC, namely $2K$, does not vary at all. Across conditions, as long as the sample size is modestly large $N \geq 100$, the difference between the approximations is almost negligible in terms of bias and root mean squared error.

In terms of model selection performance, we can see that across sample sizes and data generating processes, the performance for all three selection criteria is similar with a maximum discrepancy of at most 2 percentage points for the small sample condition, $N = 50$. The discrepancy decreases as the sample size increases, with the maximum difference between the three criteria being 1.6, .4, and .4 percentage points for $N = 100, 500, 1000$, respectively. When there is no true model included among the candidate models (i.e., data-generating processes 3 and 4), the probability of AIC, eLOOIC and LOOIC choosing the better model (in terms of the expected relative KL information) is nearly 1 when the sample size is 100 for data-generating process 4 as well as for both data-generating processes 3 and 4 once sample size goes above 100.

3.4.3 Section discussion

Study one shows that the proposed LOOIC performs similarly to eLOOIC and AIC in terms of how well it approximates twice the expected relative KL information and how frequently it selects the best model. While the estimated bias is smaller than that of the AIC for the smallest sample size of $N = 50$, the estimated root mean squared error is larger for that sample size. These differences become negligible once $N \geq 100$. However, the LOOIC has the advantage of being more general than the AIC because it can be derived for other scoring rules. We will return to this point later. In the next section, we will discuss an easy-to-implement extension to account for clustered data and show how it can be useful when evaluating multilevel models.

3.5 Evaluation of multilevel models for clustered data

Multilevel models are models with random effects that represent heterogeneity between clusters and induce dependence among units within the same cluster. Such models are widely used across many academic disciplines, due to their flexibility in modeling between-cluster heterogeneity. General modeling frameworks (Muthén, 2002; Skrondal and Rabe-Hesketh, 2004; Bartholomew et al., 2011) have also been proposed to subsume variants of random effects and latent variable models across different disciplines (e.g., discrete choice models (e.g.,

Train, 2009) in econometrics/business, psychometric models in psychology/education (e.g., Embretson and Reise, 2013), and many longitudinal data analysis models (e.g., Verbeke, 1997) in biostatistics). For an overview, see Chapters 2 and 3 in Skrondal and Rabe-Hesketh (2004).

Since multilevel models can be used to predict the responses for new units from either a new cluster or from a cluster that already exists in the data (Skrondal and Rabe-Hesketh, 2009), cross-validation can be designed for choosing the best model for new clusters or for the clusters in the data. This additional complexity motivates two different target quantities for model comparison for clustered data. We will use a two-level model for the following discussion but the extension to more levels should be straightforward.

3.5.1 Marginal and conditional perspectives on multilevel models

A two-level linear model for response y_{ij} of unit i in cluster j can be written as

$$y_{ij} = x_{ij}^T \beta + z_{ij}^T \zeta_j + \varepsilon_{ij},$$

where x_{ij} is a covariate vector with corresponding fixed coefficient vector β , z_{ij} is a covariate vector with random coefficient vector ζ_j and ε_{ij} is a level-1 error. It is usually assumed that ζ_j has a multivariate normal distribution with zero means and constant unstructured covariance matrix Ψ and that ε_{ij} is normally distributed with zero mean and constant variance σ^2 .

To facilitate discussion below, we can write the “cluster-wise” model for cluster j as

$$y_j = X_j \beta + Z_j \zeta_j + \varepsilon_j,$$

where y_j is the vector of responses for the n_j units in cluster j and ε_j is the corresponding vector of level-1 errors, X_j is an $n_j \times d$ matrix with rows x_{ij}^T , and Z_j an $n_j \times k$ matrix with rows z_{ij}^T . We refer to all the (unique) parameters of the distribution of ζ_j as ψ , and assemble the remaining fixed parameters (regression coefficients and σ) in ω .

To accommodate many types of outcomes, we can generalize the model to a multilevel generalized linear model, with the linear predictor η_{ij} defined as

$$\eta_{ij} := h^{-1}[E(y_{ij}|\zeta_j, x_{ij}, z_{ij})] = x_{ij}^T \beta + z_{ij}^T \zeta_j,$$

where $h^{-1}(\cdot)$ is the link function. Equivalently, we can write the conditional expectation μ_{ij} as

$$\mu_{ij} := E(y_{ij}|\zeta_j, x_{ij}, z_{ij}) = h(x_{ij}^T \beta + z_{ij}^T \zeta_j).$$

This includes all the outcomes whose conditional distributions given the covariates and random effects are in the exponential family of distributions. For example, for a continuous outcome, the link function is usually the identity; for a binary response, the link function is often a logit function, $\log \frac{\mu_{ij}}{1-\mu_{ij}}$; for count data assumed to follow a Poisson distribution,

the link function is usually $\log(\mu_{ij})$, and so on. The likelihood can be defined accordingly depending on which conditional distribution is specified.

The likelihood used for maximum likelihood estimation is marginal over the random effects. However, when it comes to prediction, cross validation, and information criteria, we can decide whether we want to use the marginal likelihood or whether we want to condition on the random effects. Not showing conditioning on covariates (and conditioning on parameters as before even though they are not random) for clarity, these likelihoods (for subsets of the data) can be written as:

Conditional likelihood

$$f_c(y_{ij} \mid \omega, \zeta_j),$$

the density/probability that we would use for predicting the response of a unit in cluster j if we knew ζ_j , the cluster-specific random effects. Here cluster j must be in the training data, and our interest must be in making predictions for this cluster.

Marginal likelihood

$$f_m(y_j \mid \omega, \psi) = \int f_c(y_j \mid \omega, \zeta_j) g(\zeta_j \mid \psi) d\zeta_j,$$

the density/probability we would use to predict the responses for a cluster j with unknown random effects. Here cluster j is assumed to be a *new* cluster from the same population as the clusters in the training data.

Taking a random-intercept model as an example, the model can be written as,

$$\begin{aligned} \zeta_j &\sim N(0, \psi) \\ y_{ij} \mid \zeta_j &\sim N(x_{ij}^\top \beta + \zeta_j, \sigma^2), \end{aligned} \tag{3.26}$$

where ζ_j , the random intercept for cluster j , is generated first and y_{ij} is then generated for each unit, given ζ_j . In this example, $g(\zeta_j \mid \psi)$ and $f_c(y_{ij} \mid \omega, \zeta_j)$ are normal densities as shown in (3.26), and $f_m(y_j \mid \omega, \psi)$ is multivariate normal with means $x_{ij}^\top \beta$ and exchangeable covariance matrix, with variances $\psi + \sigma^2$ and covariances ψ . While $f_m(y_j \mid \omega, \psi)$ has a closed form for all linear mixed models, the integration must be performed numerically (e.g. Rabe-Hesketh et al., 2002, 2005) for other generalized linear mixed models. For fully Bayesian models, we also need to specify priors for ω and hyperpriors for ψ .

3.5.2 Asymptotic equivalence of cAIC/mAIC and eLOuOIC/eLOcOIC

Based on the two versions of the likelihood, models can be evaluated for their two types of predictive performance. Vaida and Blanchard (2005) proposed conditional AIC (cAIC)

along with marginal AIC (mAIC) for multilevel linear models, defined as:

$$\begin{aligned} \text{cAIC} &= -2 \log f_c(y \mid \hat{\omega}, \hat{\zeta}) + 2p_{\text{Ac}} = -2 \sum_{j=1}^J \sum_{i=1}^{n_j} \log f_c(y_{ij} \mid \hat{\omega}, \hat{\zeta}_j) + 2p_{\text{Ac}} \\ \text{mAIC} &= -2 \log f_m(y \mid \hat{\omega}, \hat{\psi}) + 2p_{\text{Am}} = -2 \sum_{j=1}^J \log f_m(y_j \mid \hat{\omega}, \hat{\psi}) + 2p_{\text{Am}}. \end{aligned}$$

Most of the notation was defined in the previous section, and we also use $\log f_c(y \mid \hat{\omega}, \hat{\zeta})$ to denote the sum of the conditional log likelihood contributions of all units for a dataset y with corresponding estimated/predicted random effects conditioned on. Similarly, we use $\log f_m(y \mid \hat{\omega}, \hat{\psi})$ to denote the sum of the marginal log likelihood contributions of all clusters (hence all units). The penalty terms are defined as

$$\begin{aligned} p_{\text{Ac}} &= \hat{\rho} + 1 \\ p_{\text{Am}} &= K, \end{aligned}$$

where $\hat{\rho}$ is the trace of the estimated hat matrix which projects the observations onto the predicted outcome based on the conditional likelihood and K is equal to the number of all parameters in ω and ψ , with the random effects being treated as latent variables, missing data, or nuisance parameters. Notice that the marginal AIC is essentially the same as that for unclustered data because the the first term in mAIC is just minus twice the marginal log likelihood that is used for parameter estimation, although how to define the number of parameters is less obvious in the Bayesian context (e.g., Merkle et al., 2019).

Similar to equation (3.2), the target quantity of the conditional version of AIC (cAIC), provided in Vaida and Blanchard (2005)'s equation (8), is

$$\text{conditional expected predictive log density} = -2E_{y', \zeta} E_{y|\zeta} \log f \left\{ y \mid \hat{\omega}(y'), \hat{\zeta}(y') \right\} \quad (3.27)$$

and the target quantity for mAIC, provided in equation (1) in Vaida and Blanchard (2005), is

$$\text{marginal expected predictive log density} = -2E_{y'} E_y \log f \left\{ y \mid \hat{\omega}(y'), \hat{\psi}(y') \right\}. \quad (3.28)$$

It turns out that, similar to the correspondence between AIC and eLOOIC for unclustered data (Stone, 1977), Fang (2011) shows that asymptotically, cAIC corresponds to leave one unit out (LOuO) while mAIC corresponds to leave one cluster out (LOuO). Similar arguments have been made also for Bayesian latent variable models in Merkle et al. (2019). Specifically, we can refer the predictive log density for unit i in cluster j when leaving one

unit out as $f_c(y_{ij} | \hat{\omega}_{(-ij)}, \hat{\zeta}_{j(-ij)})$ and define the corresponding exact LOuO information criterion (eLOuOIC) as

$$\text{eLOuOIC} = -2 \sum_{j=1}^J \sum_{i=1}^{n_j} \log f_c(y_{ij} | \hat{\omega}_{(-ij)}, \hat{\zeta}_{j(-ij)}). \quad (3.29)$$

Here $\hat{\omega}_{(-ij)}$ and $\hat{\zeta}_{j(-ij)}$ are the parameter estimates/predicted random effects based on the dataset with unit i in cluster j removed. The predictive log density for all units in cluster j when leaving one cluster out is

$$f_m(y_j | \hat{\omega}_{(-j)}, \hat{\psi}_{(-j)}) = \int f_c(y_j | \hat{\omega}_{(-j)}, \zeta_j) g(\zeta_j | \hat{\psi}_{(-j)}) d\zeta_j$$

and the corresponding eLOcOIC is

$$\text{eLOcOIC} = -2 \sum_{j=1}^J \log f_m(y_j | \hat{\omega}_{(-j)}, \hat{\psi}_{(-j)}). \quad (3.30)$$

Here $\hat{\omega}_{(-j)}$ and $\hat{\psi}_{(-j)}$ are parameter estimates based on the dataset with all units in cluster j removed.

3.5.3 Approximations for eLOuOIC and eLOcOIC

Analogous to the LOOIC defined in equation (3.25) for unclustered data, we now define the IJ approximations for the eLOuOIC and eLOcOIC for clustered data.

For LOuOIC, we consider the original data after deleting one unit, or equivalently we set the weight for one unit i in one cluster j to 0 and the remaining weights to 1, with corresponding weight vector denoted $1_{(-ij)}$. Following the definition in the unclustered case, we define

$$\text{LOuOIC} = -2 \log f(y | \hat{\omega}_{(-ij)}^{\text{IJ}}, \hat{\zeta}_{j(-ij)}^{\text{IJ}}) = -2 \sum_{j=1}^J \sum_{i=1}^{n_j} \log f_c(y_{ij} | \hat{\omega}_{(-ij)}^{\text{IJ}}, \hat{\zeta}_{j(-ij)}^{\text{IJ}}).$$

For LOcOIC, we consider the original data after deleting all units for one cluster j , or equivalently assign weights equal to 0 for all n_j units in the cluster and equal to 1 for all other units, with corresponding weight vector denoted $1_{(-j)}$,

$$\text{LOcOIC} = -2 \log f(y | \hat{\omega}_{(-j)}^{\text{IJ}}, \hat{\psi}_{(-j)}^{\text{IJ}}) = \sum_j \log f_m(y_j | \hat{\omega}_{(-j)}^{\text{IJ}}, \hat{\psi}_{(-j)}^{\text{IJ}}).$$

In the current chapter, we will mainly focus on mAIC and LOcOIC because most of time we are interested in a model's predictive performance when applied to new clusters.

3.6 Study II: Comparing LOcOIC, eLOCOIC and mAIC with the target quantity for clustered data

In this section, we extend Study I to clustered data and examine how closely LOcOIC, eLOcOIC and mAIC agree with each other and with the target quantity for multilevel models. The central question that we would like to answer is how well LOcOIC estimates the target quantity and how well model selection performs based on LOcOIC.

3.6.1 Design

Extending the simulation setup in Study I, we use a two-level random-intercept model,

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \zeta_j + \epsilon_{ij},$$

to generate the data, with $\epsilon_{ij} \sim N(0, 1)$ and $\zeta_j \sim N(0, 1)$. The scalar covariate x_{ij} is set to $\{0, \frac{1}{I-1}, \dots, \frac{I-2}{I-1}, 1\}$, for each cluster, where I is the cluster size. We vary the number of clusters ($J \in \{10, 50, 100\}$) and (constant) cluster size ($I \in \{10, 30\}$), and set the regression coefficient vectors to different values across simulation conditions. For candidate models, we use one variance-components model (no covariate included, denoted M1) and one random-intercept model with one predictor x_{ij} (denoted M2).

Similar to Study I, we use Monte Carlo integration with 200,000 samples to obtain a good estimate for the target quantity, twice the expected relative KL information $-2E_{y'}E_y \log f_m(y | \hat{\omega}(y'), \hat{\psi}(y'))$ for both M1 and M2. We estimate the bias and root mean squared error of LOcOIC, eLOcOIC, and mAIC, as estimates of the target quantity. We also compare empirical model selection performance (i.e., proportion of the time the correct model is selected). The correct model is defined as the true model when the true model is among the candidate models or as the better performing model, in terms of the target quantity, when the candidate models do not include the true model.

We generate 500 replicates for each simulation condition. The sample mean of each type of estimate across 500 replications as well as the true target quantity are summarised in Table 3.2. The estimated bias and estimated root mean squared errors are shown in Figure 3.4 and Figure 3.5.

3.6.2 Results

Similar to Study I, we found that across all conditions, the three estimators of the target quantity do not differ much in terms of bias, with a few exceptions. For example, for the smallest sample data ($I = J = 10$), there is a slight advantage in using eLOcOIC over LOcOIC and mAIC while LOcOIC has worse performance compared to mAIC. This could

be due to the fact that we leave out more data than the unclustered case when approximating the leave-one-cluster-out estimates by the infinitesimal jackknife, so the accuracy of the approximation decreases. There is sometimes downward bias (e.g., $I = 10, J = 100$ and $I = 30, J = 100$) and sometimes upward bias (e.g., $I = 30, J = 50$ and $I = 30, J = 100$) in estimating the target quantity. However, the empirical model selection performance of LOcOIC is close to that of eLOcOIC, as can be seen in Table 3.2.

Second, the estimated root mean squared errors are similar for all three estimators, with a slight edge of mAIC over the other two estimators when sample size is small $I = J = 10$. This finding is consistent with our conclusion in Study I and is due to the variance being smaller for mAIC compared to LOcOIC and eLOcOIC. When the number of cluster is small and cluster size is small ($I = 10, J = 10/30$), mAIC does a better job selecting the correct model due to its smaller variance.

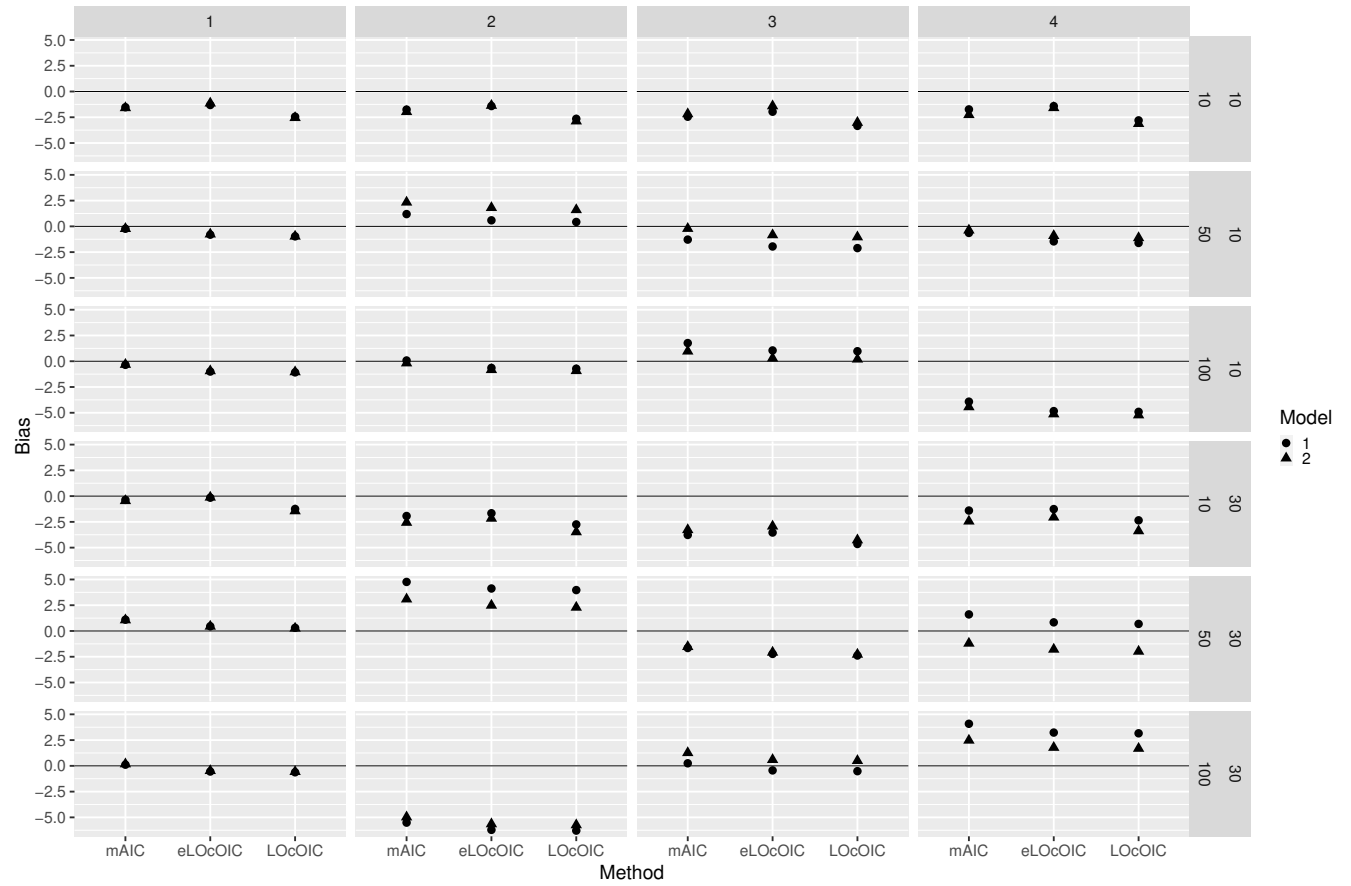


Figure 3.4: Finite sample bias estimates for mAIC, eLOcOIC, and LOcOIC. Each row represents a different sample size while each column represents a different data generating processes.

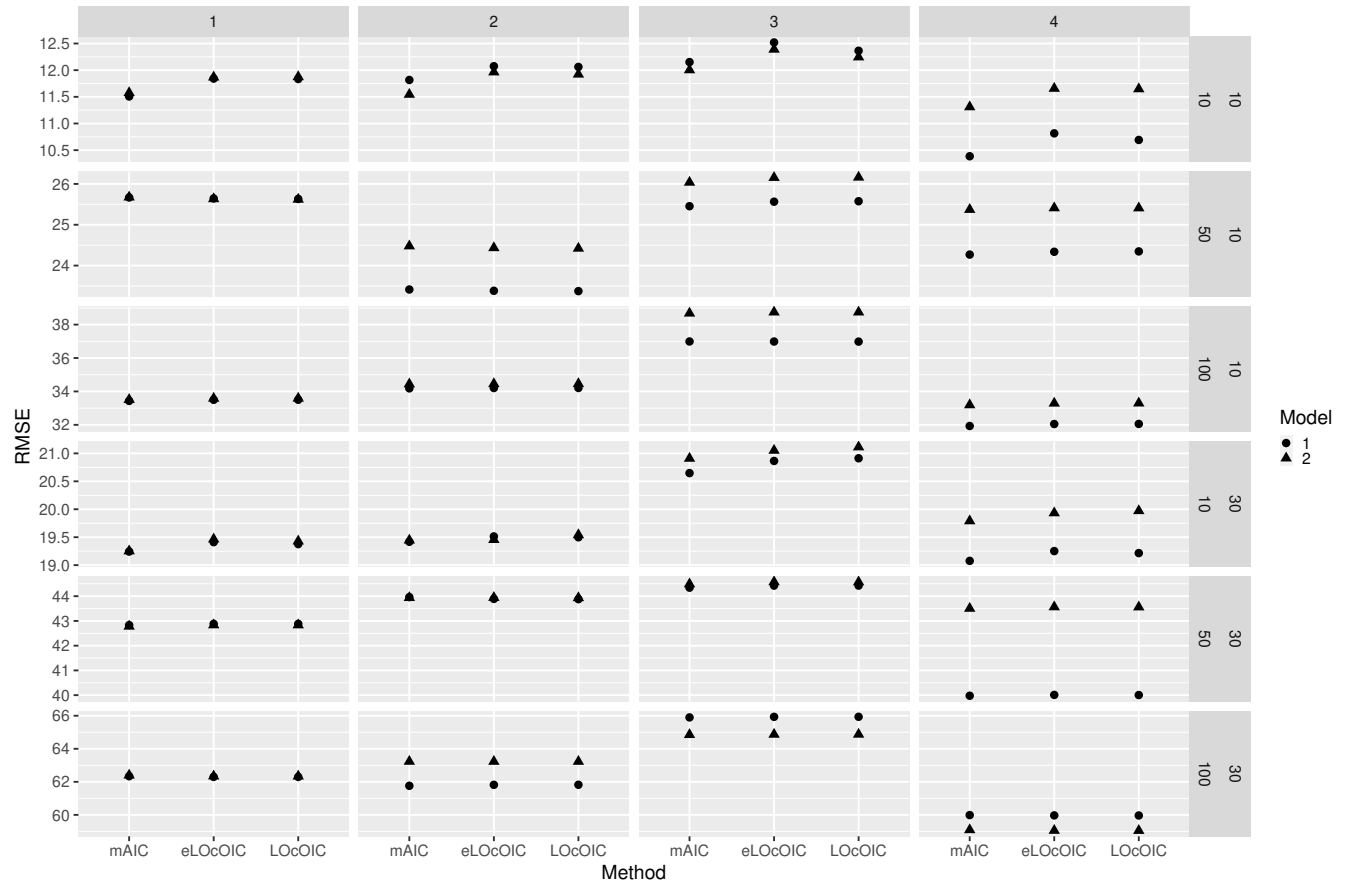


Figure 3.5: Root mean squared error estimates for mAIC, eLOcOIC, and LOcOIC. Each row represents a different sample size while each column represents different data generating processes.

Table 3.2: Mean mAIC, eLOcOIC, and LOcOIC estimates, proportion correct, and target quantity.

I	J	DGP	Parameter	mAIC			eLOcOIC			LOcOIC			Target Quantity		
				M1	M2	Correct	M1	M2	Correct	M1	M2	Correct	M1	M2	Model
10	10	1	(0, 0, 0)	310.784	311.795	0.842	310.999	312.245	0.828	309.865	310.835	0.80	312.311	313.382	M1
10	10	2	(0, 1, 0)	320.182	311.416	0.952	320.526	312.022	0.944	319.289	310.507	0.948	321.938	313.382	M2
10	10	3	(-1, 0, 1)	320.206	311.996	0.956	320.694	312.767	0.928	319.320	311.160	0.948	322.655	314.173	M2
10	10	4	(-1, -1, 1)	344.590	311.915	1	344.909	312.590	1	343.525	311.050	1	346.327	314.173	M2
10	50	1	(0, 0, 0)	1541.828	1542.846	0.850	1541.252	1542.313	0.846	1541.088	1542.103	0.844	1542.058	1543.070	M1
10	50	2	(0, 1, 0)	1591.486	1545.408	1	1590.880	1544.881	1	1590.720	1544.674	1	1590.293	1543.070	M2
10	50	3	(-1, 0, 1)	1592.606	1546.850	1	1591.938	1546.232	1	1591.783	1546.030	1	1593.890	1547.065	M2
10	50	4	(-1, -1, 1)	1712.116	1546.691	1	1711.295	1546.157	1	1711.144	1545.953	1	1712.753	1547.065	M2
10	100	1	(0, 0, 0)	3080.427	3081.461	0.838	3079.775	3080.840	0.834	3079.696	3080.738	0.830	3080.776	3081.779	M1
10	100	2	(0, 1, 0)	3177.335	3081.600	1	3176.616	3080.951	1	3176.541	3080.852	1	3177.263	3081.779	M2
10	100	3	(-1, 0, 1)	3186.225	3090.750	1	3185.509	3090.072	1	3185.434	3089.973	1	3184.461	3089.780	M2
10	100	4	(-1, -1, 1)	3418.356	3085.343	1	3417.439	3084.654	1	3417.367	3084.556	1	3422.278	3089.780	M2
30	10	1	(0, 0, 0)	889.858	890.792	0.806	890.071	891.131	0.794	888.978	889.803	0.776	890.225	891.247	M1
30	10	2	(0, 1, 0)	913.856	888.686	1	914.123	889.083	1	913.040	887.767	1	915.786	891.247	M2
30	10	3	(-1, 0, 1)	913.750	889.864	1	913.984	890.210	1	912.873	888.870	1	917.519	893.133	M2
30	10	4	(-1, -1, 1)	981.144	890.694	1	981.284	891.082	1	980.192	889.751	1	982.543	893.133	M2
30	50	1	(0, 0, 0)	4432.822	4433.797	0.824	4432.182	4433.183	0.822	4432.031	4432.989	0.818	4431.726	4432.732	M1
30	50	2	(0, 1, 0)	4564.263	4435.822	1	4563.624	4435.217	1	4563.471	4435.021	1	4559.501	4432.732	M2
30	50	3	(-1, 0, 1)	4566.511	4440.656	1	4565.938	4440.112	1	4565.782	4439.914	1	4568.163	4442.185	M2
30	50	4	(-1, -1, 1)	4895.122	4440.983	1	4894.350	4440.400	1	4894.199	4440.203	1	4893.513	4442.185	M2
30	100	1	(0, 0, 0)	8860.208	8861.271	0.864	8859.543	8860.620	0.854	8859.477	8860.526	0.852	8860.094	8861.094	M1
30	100	2	(0, 1, 0)	9110.132	8856.120	1	9109.429	8855.451	1	9109.356	8855.357	1	9115.641	8861.094	M2
30	100	3	(-1, 0, 1)	9133.184	8881.237	1	9132.494	8880.567	1	9132.421	8880.472	1	9132.936	8879.970	M2
30	100	4	(-1, -1, 1)	9787.804	8882.446	1	9786.951	8881.748	1	9786.879	8881.654	1	9783.721	8879.970	M2

[†] Best model according to target quantity

3.7 Concluding Remarks

In this chapter, we discussed the potential use of approximated information criteria for both unclustered (LOOIC) and clustered data (LOcOIC) and empirically evaluated their performance in terms of estimating the target quantity as well as selecting the correct models. We also compared the LOOIC and LOcOIC with their AIC counterparts. We found that across the different data-generating processes considered, LOOIC and LOcOIC perform similarly in terms of both estimation of the target quantity and model selection with their AIC counterparts, except for very small dataset where they perform a little worse.

In our simulation studies, we only considered the cases under minor model misspecification. However, because LOO-based information criteria correspond to TIC, the generalization of AIC, so we expect that a better performance for the LOOIC and LOcOIC than their AIC counterparts can be observed when there is a substantial difference between the true and candidate models. We will leave it as our future work.

When evaluating multilevel models, we mainly considered the LOcOIC which targets the predictive performance of new data from new clusters because this corresponds to how most researchers choose models most of time. However, there could be applications for considering LOuOIC, when we are interested in predictive performance of the models for new units from existing clusters. An example would be when the goal is to make forecasts for subjects (the clusters) based on longitudinal data. Deriving the IJ approximation for this case and evaluating the performance will be an interesting project for the future. For generalized linear mixed models, LOcOIC may add more complexity related to obtaining IJ approximated LOO predictions for the random effects. Potential problems could also arise due to certain violations of appropriate regularity conditions (Millar, 2018). However we leave this question for future investigation. As noted before, alternative scoring rules (Gneiting, 2011) could also be used and investigation of the theoretical and empirical properties can be pursued in the future.

As noted before, we chose the IJ approximation mainly due to its simplicity. However, since there are many other approximation methods available (e.g., Cain and Lange, 1984; Wang et al., 2018; Giordano et al., 2019a; Alaa and Van Der Schaar, 2020), it may be interesting to contrast and compare their performance in terms of estimating the target quantity as well as model selection for both unclustered and clustered data.

3.8 References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Alaa, A. and Van Der Schaar, M. (2020). Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*, pages 165–174. PMLR.
- Anderson, D. and Burnham, K. (2004). *Model selection and multi-model inference*. Second. NY: Springer-Verlag.
- Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.
- Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.
- Bates, S., Hastie, T., and Tibshirani, R. (2021). Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*.
- Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18:1–43.
- Cain, K. C. and Lange, N. T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics*, 40(2):493–499.
- Cawley, G. C. and Talbot, N. L. (2003). Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11):2585–2592.
- Cawley, G. C. and Talbot, N. L. (2004). Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17(10):1467–1475.
- Cawley, G. C. and Talbot, N. L. (2008). Efficient approximate leave-one-out cross-validation for kernel logistic regression. *Machine Learning*, 71(2):243–264.
- Commenges, D., Proust-Lima, C., Samieri, C., and Lique, B. (2015). A universal approximate cross-validation criterion for regular risk functions. *The International Journal of Biostatistics*, 11(1):51–67.
- Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

- Fang, Y. (2011). Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. *Journal of Data Science*, 9(1):15–21.
- Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.
- Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.
- Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019b). A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.
- Jaekel, L. A. (1972). *The infinitesimal jackknife*. Bell Telephone Laboratories.
- Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.
- Lumley, T. and Scott, A. (2015). Aic and bic for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18.
- Meijer, R. J. and Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155.
- Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3):802–829.
- Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using waic and cross-validation. *Statistics and Computing*, 28(2):375–385.
- Miller Jr, R. G. (1974). An unbalanced jackknife. *The Annals of Statistics*, 2(5):880–891.

- Muthén, B. O. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, 29(1):81–117.
- Obuchi, T. and Kabashima, Y. (2016). Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(5):053304.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301–323.
- Rad, K. R. and Maleki, A. (2020). A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.
- Shibata, R. (1989). Statistical aspects of model selection. In *From data to model*, pages 215–240. Springer.
- Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.
- Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):659–687.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.
- Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *suri-kagaku (mathematical sciences)* 153 12-18.
- Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.
- Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370.
- Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432.

- Verbeke, G. (1997). Linear mixed models for longitudinal data. In *Linear mixed models in practice*, pages 63–153. Springer.
- Verweij, P. J. and Van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314.
- Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. (2018). Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*, pages 5228–5237. PMLR.
- Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).
- Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, 6(3):675–692.