

UNIFIED FRAMEWORK FOR POST-SELECTION INFERENCE

Arun Kumar Kuchibhotla

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Lawrence D. Brown

Lawrence D. Brown, Professor of Statistics

Andreas Buja

Andreas Buja, The Liem Sioe Liong/First Pacific Company Professor of Statistics

Dissertation Committee

Andreas Buja, The Liem Sioe Liong/First Pacific Company Professor of Statistics.

Edward I. George, Universal Furniture Professor of Statistics.

Dylan Small, Professor of Statistics

UNIFIED FRAMEWORK FOR POST-SELECTION INFERENCE

© COPYRIGHT

2020

Arun Kumar Kuchibhotla

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

Dedicated to my parents, Prabhakara Kumar and Sita Kumari

and

my academic parents (late) Lawrence D. Brown and Linda H. Zhao.

ACKNOWLEDGEMENT

Firstly, I thank you, the reader, for picking my thesis to read. Hope you find it interesting and useful for your purposes.

Proceeding to thank the people that stood by me throughout my life and particularly, my life as a PhD student, I start with my parents (Prabhakara Kumar and Sita Kumari) and brother (Sarma) for standing by me patiently over the years while I am away on studies.

I consider myself lucky in many ways. Although out of my league, I was fortunate to have been admitted to the Department of Statistics at the University of Pennsylvania. Furthermore, Larry (Lawrence D. Brown) was kind enough to take me on as a student. Since then I have been working with Larry as a part of the PoSI group including, such great minds as Andreas Buja, Edward I. George, Eric J. Tchetgen Tchetgen, Richard Berk, Linda Zhao, and Junhui Cai. Larry being such a kind person he was, he supported me as an RA during my first two years. Unfortunately, I and the whole statistics community lost Larry on 21 February 2018. Linda has been kind enough to take care of me as a son, supporting me over the last three years.

Larry and Andreas are my supervisors. I have learnt some much from both of them over the past few years. In particular, I very much appreciate Andreas' efforts to correct my pronunciation and writing. It is so enriching to be a part of the PoSI group¹. Most of my papers have started appearing after Larry passed away and Andreas has been great at taking care of my writing to ensure the readability. I sincerely hope I can retain some of Andreas' wisdom in writing my future papers; I doubt I will but I can hope all I can. (Old habits die hard, right?). I thank Eric, Richard, Ed, and Linda, for their support, questions, and ideas that have led me to improve my presentation over many years.

¹which is now renamed as the Larry's group.

A special shout-out to Junhui Cai, without whom this thesis would not be complete. He has been my technical advisor and computational collaborator for the past three years. Almost all of the simulations in my papers and in this thesis were produced by Junhui Cai.

Prof. Todd Kuffner of Washington University deserves a special mention in this thesis. He is the organizer of the workshop on higher order asymptotics and post-selection inference (WHOA-PSI) since 2016. He supported me as a friend over the years and inviting me to present at his workshop. More than anything, I thank him for organizing the workshop which is a good forum for meeting people. Another special mention is Prof. Ayanendranath Basu of Indian Statistical Institute (ISI), Kolkata. He was my master's thesis advisor at ISI but he is much more than that. He has supported in almost all respects while I was doing research at ISI.

I should also thank my friends/well-wishers from CMU: Larry Wasserman, Alessandro Rinaldo, Aaditya Ramdas, and Ryan Tibshirani for their support.

I also thank my friends Gajjala Chalapathi Charan, Cecilia Balocchi, and Mo Huang for, well, being friends. Also, I should thank the statistics staff: Noelle, Carol, Tanya, Carol and Adam for their support.

Thank you all,
Arun Kumar Kuchibhotla.

ABSTRACT

UNIFIED FRAMEWORK FOR POST-SELECTION INFERENCE

Arun Kumar Kuchibhotla

Lawrence D. Brown

Andreas Buja

The development of the classical inferential theory of mathematical statistics is based on the philosophy that all the models to fit, all the hypotheses to test and all the parameters to do inference for are fixed prior to the collection of data. Interestingly and in fact, more concerningly, this is not how the practice of statistics is. The practice of statistics often explores (if not tortures) the data to find the “right” model to fit to the data, “right” hypothesis to test and so on. Quoting [Tulloch \(2001, page 205\)](#)

As Ronald Coase says, ”if you torture the data long enough it will confess”.

The young researcher, convinced he knows the truth will make changes in his specifications and very likely produce significant results. In some cases this is correct; his original specification was wrong and his new one is right. Nevertheless, this procedure reduces the significance of the significance test.

Once the data is explored to find the hypothesis or model, the classical theory is (bluntly speaking) useless for inference and can in fact be very misleading.

The current thesis focuses on the problem of providing Valid Inference after Data Exploration (VIDE). Although a unified framework is provided for such a goal, the framework is explained through the problem of inference with the ordinary least squares linear regression estimator when the data is explored to find the “right” subset of covariates to be used in the regression model.

Valid post-selection inference has been a topic of research interest at least since 1960's but has received increasing attention in the recent times. Invalidity of classical inference in post-selection problems may not only be due to the selection but also due to misspecification of model. Misspecification is a very natural outcome of model selection since the selected model cannot always be guaranteed to match the truth. If such a guarantee exists, then the post-selection problem does not require further study. Most of the literature on valid post-selection inference has concentrated on the assumption of a true parametric model.

In this thesis, valid post-selection inference is provided under no parametric assumptions. The simplest setting in this thesis is when the observations are independent satisfying certain moment restrictions (and no further model/distributional assumptions). Extensions to various dependent settings are also given. Throughout, the total number of covariates available is allowed to grow with the sample size and can be almost exponential in the sample size.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iv
ABSTRACT	vi
LIST OF TABLES	xi
LIST OF ILLUSTRATIONS	xiv
1 Motivating Examples	1
1.1 Practice of Statistics in Textbooks/Education	3
1.2 Practice of Statistics in Literature	7
1.3 Some Observations and Formulation of the Problem	12
2 Outline of the Thesis	16
2.1 General Introduction to VIDE and Some Review	16
2.2 Simultaneous Inference Approach to VIDE	21
2.3 Selective Inference Approach to VIDE	23
2.4 Sample Splitting Approach to VIDE	23
2.5 Outline of the Thesis	25
3 Assumption-lean Framework for Linear Regression	30
3.1 Target of Estimation	32
3.2 Linear Regression with Fixed Covariates	38
3.3 Linear Regression with Random Covariates	41
3.4 Unified Framework for Linear Regression	44
3.5 Variance Estimation and Bootstrap in Unified Framework	48

3.6	Hypothesis Testing in the Unified Framework	57
3.7	Conclusions on Assumptions for Linear Regression	60
3.A	Semiparametric Efficiency	62
4	Post-selection Inference in Linear Regression	71
4.1	Notation and Problem Formulation	73
4.2	Equivalence of Post-selection and Simultaneous Inference	79
4.3	First Approach for Post-Selection Inference	83
4.4	Computation by Multiplier Bootstrap	93
4.5	A Generalization for Linear Regression-type Problems	97
4.6	Connection to High-dimensional Regression	99
4.7	Pros and Cons of Approach 1	102
4.8	Numerical examples	105
4.A	Proof of Lemma 4	109
4.B	Proof of Lemma 6	110
4.C	Proof of Theorem 8	111
4.D	High-dimensional CLT and Bootstrap Consistency	113
4.E	Bounds on $\ \hat{\Omega}_n - \Omega_n\ _\infty$ under Dependence	123
5	Unified Framework for Post-selection Inference	140
5.1	General Recipe for Valid PoSI	141
5.2	Application to Linear Regression	154
5.3	On the Shape of Intervals for Valid Post-selection Inference	162
5.4	Simulations illustrating the Power of HPoSI	176
5.A	Proof of Theorem 13	178
5.B	Proof of Theorem 14	181
5.C	Proof of Theorem 15	184

5.D	Proof of Theorem 16	186
5.E	Proof of Lemma 10	193
5.F	Proof of Theorem 17	195
6	Real Data Examples	202
6.1	Boston housing data	202
6.2	Telomere length example	207
7	Conclusions	211

LIST OF TABLES

6.1	Variables in the Boston housing data by Harrison Jr and Rubinfeld (1978).	203
6.2	Significance at 0.05 level of variables in the final model with and without adjustment.	205
6.3	Unadjusted confidence interval for the final model.	205
6.4	HPoSI (simultaneous) for the final model.	206
6.5	HPoSI (marginal) for the final model.	207
6.6	Variables in the Telomere Length (TL) analysis by Nersisyan et al. (2019).	208
6.7	Significance at 0.05 level of variables in the final model with and without adjustment.	209
6.8	Unadjusted confidence interval for the final model.	210
6.9	HPoSI (simultaneous) for the final model.	210
6.10	HPoSI (marginal) for the final model.	210

LIST OF ILLUSTRATIONS

1.1	Attained level of a classical test when the hypothesis is selected after data exploration.	3
1.2	Tukey’s Ladder of Transformation. The Four Quadrant Approach. . .	4
1.3	Telomere Length	10
2.1	Illustration of the Unified Framework. Here $G_{j,M}$ is a standard Gaussian random variable and $(G_{j,M})_{j,M}$ is a Gaussian random vector with covariance matching that of the vector of averages.	27
2.2	The following chapter is based on Kuchibhotla et al. (2018).	29
3.1	The following chapter is based on Kuchibhotla et al. (2020)	70

4.1	Comparison of “UPoSI” with “PoSI” (Berk et al., 2013) and “selective Inference” (Tibshirani et al., 2016). Methods included are the “UPoSI” confidence regions $\hat{\mathcal{R}}_{n,M}^\dagger$ (4.12) and the projected Box regions: “UPoSIBox” regions. The first two plots provide comparisons with the “PoSI” regions (5.16) of Berk et al. (2013). The next four plots show comparisons with “selective Inference.” Rather than providing overall simultaneous coverage, we show simultaneous coverage for different model sizes separately: $1 \leq M \leq 15$ for comparison with “PoSI” and $1 \leq M \leq 5$ for comparison with “selective Inference.” Because the volume of a region in $ M $ dimensions scales like $C^{ M }$ for some constant C , we plot $\log(\text{Leb}(\hat{\mathcal{R}}_{n,M}^\dagger))/ M $, which allows comparison across different model sizes. Recall that in Setting C models fall into two groups: those that contain the last covariate, and those that don’t. This is the reason for showing two dots for each model size in Setting C. The size of dots indicates the proportion of models in each group. The dashed lines in the coverage plots show the nominal confidence level 0.95.	107
4.2	Comparison of coverage and volume of UPoSI with sample splitting. In all cases the volume of our confidence regions are at least as good as sample splitting. The latter is slightly more conservative in coverage in some cases, but not dramatically so.	108
4.3	The following chapter is partly based on Kuchibhotla et al. (2019).	139
5.1	Illustrating the dependence on k of the max- $ t $ statistic: Telomere length analysis. The most correlated covariate in this data is the interaction between the telomere lengths of the parents.	164

5.2	Distribution of max- t for one model that includes the last covariate and one that excludes that last covariate for $p = 20, 100$ under Setting (5.18).	166
5.3	Average ratio of widths PoSI vs HPoSI of 1,000 simulations with different maximal model size k , i.e., $ M \leq k, k = 1, 3, 5, 7, 9, 10$ for $p = 11, 12, \dots, 20$ under Setting (5.18).	176
5.4	Illustrating the dependence on k : Telomere length analysis. Comparison of PoSI and HPoSI.	177

Chapter 1

Motivating Examples

In 2005, the Stanford epidemiologist John Ioannidis made a dramatic claim: “most published research findings are false.” The claim is largely believed to be true. It gave rise to the term “reproducibility/replicability crisis.” Several factors have been identified as possible causes, first among them publication bias, that is, the fact that null findings tend not to get published. This is an institutional problem whose solution is the reform of publication policies. Another contributing factor, closer to home for us statisticians, is the breakdown of the classical statistical inference framework under the current practice of statistics. In the modern practice of statistics, data analysts tend to use many forms of data exploration before applying statistical inference, and this is a problem that requires our serious attention. This aspect of the replicability crisis will be the backdrop of the thesis. The main goal of the thesis is to provide a unified framework for resolving the problem of **V**alid **I**nference after **D**ata **E**xploration (**VIDE**).

Classical statistical inference framework is built to provide valid statistical conclusions when the hypotheses to test and the model to fit are decided without the involvement of the data at hand. The practice of statistics does not follow this se-

quence, as will be shown in this chapter. This deviation from the classical inference framework can drastically invalidate the conclusions. For an illustration of the drastic invalidity when the hypothesis to test is chosen based on the data, consider the following example:

1. Generate 500 observations from $(Y_1, X_1), \dots, (Y_n, X_n) \stackrel{iid}{\sim} N(0, I_{p+1})$, for some $p \geq 1$. In this distribution, Y_i 's are independent of X_i 's.
2. Select one covariate which is the most correlated with the response, that is,

$$\hat{j} := \arg \max_{1 \leq j \leq p} |\widehat{\text{corr}}(Y, X_j)|.$$

Here $\widehat{\text{corr}}$ is computed based on the 500 observations.

3. Compute the least squares estimator

$$(\hat{\alpha}_{\hat{j}}, \hat{\beta}_{\hat{j}}) := \arg \min_{(\theta_1, \theta_2)} \frac{1}{n} \sum_{i=1}^n (Y_i - \theta_1 - \theta_2 X_{i,\hat{j}})^2,$$

where $X_{i,j}$ represents the j -th coordinate of X_i .

4. Test the hypothesis $H_{0,\hat{j}}$ of insignificant coefficient based on the estimator $\hat{\beta}_{\hat{j}}$. The classical test of level 0.05 in this case is

$$\text{Reject } H_{0,\hat{j}} \text{ if } \left| \frac{n^{1/2} \hat{\beta}_{\hat{j}}}{\hat{\sigma}_{\hat{j}}} \right| \geq 1.96. \quad (1.1)$$

Here $\hat{\sigma}_{\hat{j}}/n^{1/2}$ is the classical estimator of the standard error of $\hat{\beta}_{\hat{j}}$ (disregarding the randomness of \hat{j}). This test is same as looking at the summary of $\text{lm}(Y \sim X_{\hat{j}})$ and taking the decision of reject if the p -value is less than 0.05.

Because the response is uncorrelated with all the covariates, one might naively expect

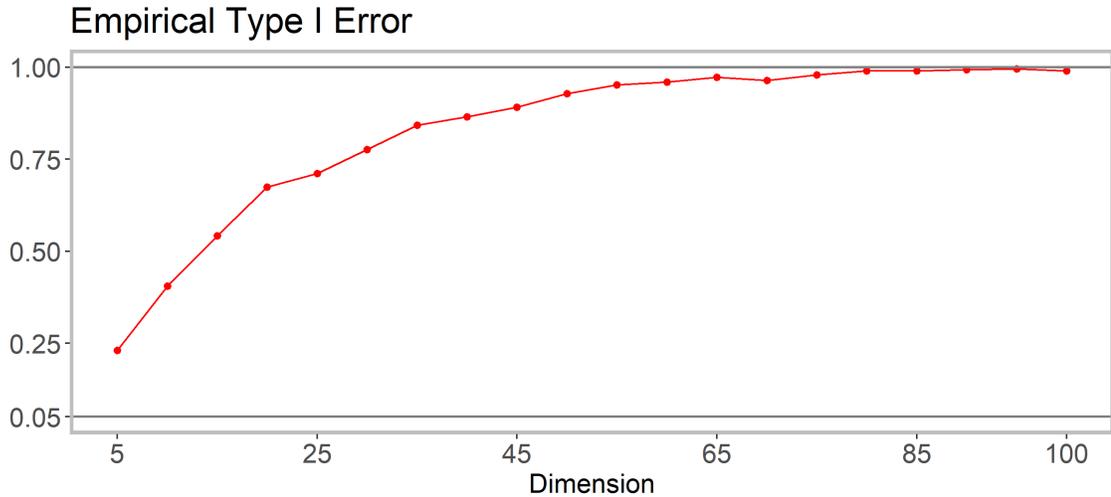


Figure 1.1: Attained level of a classical test when the hypothesis is selected after data exploration.

that the test (1.1) controls Type I error at 0.05. Figure 1.1 shows the attained true level of the test (1.1). This example shows that the classical statistical procedures do not solve the VIDE problem and requires a non-trivial adjustment.

There are many ways of mathematically formalizing the VIDE problem. In the following sections, I will provide a few examples from textbooks and published research. Then the VIDE problem will be formalized mathematically at the end of this chapter and will be solved in the forthcoming chapters.

1.1 Practice of Statistics in Textbooks/Education

In this section, I present several examples from textbooks and educational journals where the full procedure described involves testing hypotheses obtained after data exploration.

1.1.1 Case Study TE1: Moore and McCabe (1998)

Example 11.1 of Moore and McCabe (1998) introduces the GPA dataset for predicting the college GPA of students based on the high school scores in Math (HSM), Science

(HSS), and English (HSE). In the process of refining the basic linear model of GPA on HSM, HSS, HSE, the authors (page 724) write

Because the variable HSS has the largest P -value of the three explanatory variables and therefore appears to contribute the least to our explanation of GPA, we rerun the regression using only HSM and HSE as explanatory variables. The F statistic indicates that we reject the null hypothesis that the regression coefficients for the two explanatory variables are both zero. The P value is still 0.0001.

This is similar to what was done in the illustrative example. The results of the first linear model suggested the next hypothesis to test and hence is obtained as a result of data exploration. This invalidates the classical F-test. A side point to note here is that the second linear model (with HSM and HSE) might be misspecified, that is, may not satisfy all the assumptions of the classical linear model.

1.1.2 Case Study TE2: [Stine and Foster \(2013\)](#)

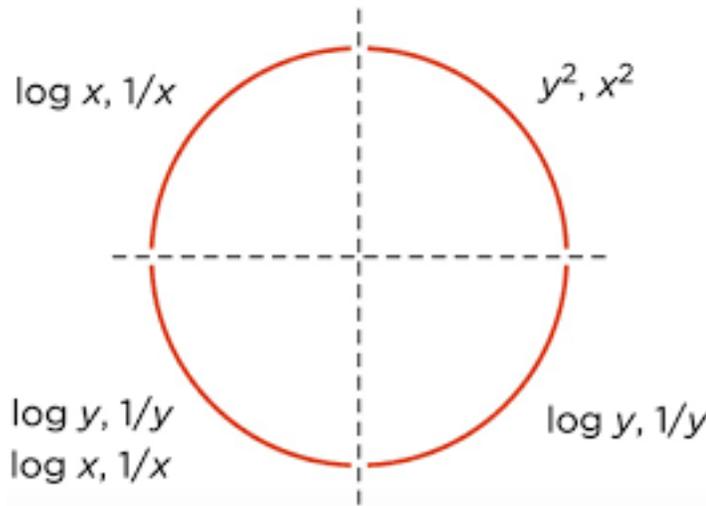


Figure 1.2: Tukey's Ladder of Transformation. The Four Quadrant Approach.

In the context of fitting a curve to bivariate data, [Stine and Foster \(2013, page 515\)](#) write

Deciding on a transformation requires several skills. First, think about the context of the problem: why should the association be linear? Then, once you see curvature in the scatterplot, compare the curvature to the bending patterns shown in [Figure 1.2](#). Among the choices offered, find the one that captures the curvature of the data and produces an interpretable equation. **Above all, don't be afraid to try several.**¹ Picking a transformation requires practice, and you may need to try several to find one that is interpretable and captures the pattern in the data.

Unlike the suggestion of Moore and McCabe in [Section 1.1.1](#), the suggestion here is more dangerous in the sense that it (actively) advises the data analyst to make subjective decisions on what transformations to try and use in the final model. Because this advice is based on visualization, it is not possible to mathematically analyze the selection method. (The suggestion in [1.1.1](#) is backward elimination which is analytically precise and is possibly amenable to mathematical analysis.)

1.1.3 Case Study TE3: [Pardoe \(2008\)](#)

This paper from the *Journal of Statistics Education* is written to address the challenges of teaching complicated aspects of linear regression modeling using Oregon realtor data. The paper, however, spells out the details of how linear regression modeling is usually taught in basic courses and this way invalidates the classical inference so badly that it may not be possible to adjust for it.

The paper models the price of a home in terms of 12 features of the home, including number of bedrooms and number of bathrooms. Section 3 of the paper fits, model 1,

¹Emphasis added here.

a linear regression for price on all 12 features. Then the author writes

However, the residuals of model 1 fail to satisfy the zero mean (linearity) assumption in a plot of residuals versus Age, displaying a relatively pronounced curved pattern. ... To attempt to correct this failing, we will add an Age² transformation to the model, which as discussed above was also suggested from the realtor's experience. The finding that the residual plot with Age has a curved pattern does not necessarily mean that an Age² transformation will correct this problem, but it is certainly worth trying.

This might look similar to the example of transformations in Section 1.1.2 but note that the decision to try transformations is done based on the data. Further, the decision of trying the square transformation is ad hoc and is not chosen from a starting family of transformations (as in Section 1.1.2). The modeling, in the paper, does not stop there and the paper proceeds

In addition, both Bath and Bed have relatively large individual *t*-test *p*-values in model 1, which appears to contradict the notion that home prices should increase with the number of bedrooms and bathrooms. ... The instructor can guide the students in seeing that to model such a relationship we need to add a Bath×Bed interaction term to the model.

This along with the square transformation of Age is called model 2. The author looks at the classical summary table as if no exploration has been done and writes

However, the model includes some terms with large individual *t*-test *p*-values, suggesting that perhaps it is more complicated that it needs and
...

and constructs a more refined model along with an interpretation.

This case study shows how fluid the modeling process is and how much it differs the classical mathematical framework of inference for linear regression. The classical framework requires fitting one model (decided a priori) and then infer.

1.2 Practice of Statistics in Literature

The previous section has shown various examples of how the practice of statistics differs from the mathematical inference framework in textbooks and education. To further illustrate the practice of data exploration, I will now present a few case studies from the literature.

1.2.1 Case Study L1: [Harrison Jr and Rubinfeld \(1978\)](#)

This is the paper that introduced the well-known Boston housing data². Although forgotten in the subsequent use of this data, the data was collected to measure the willingness to pay for clean air. Boston is divided into census tracts and in each tract, the median (MV) of the property value of homes among those in the tract. The concentration of nitrogen oxide (NOX) is used as an inverse proxy for clean air where the response/dependent variable is MV. The dataset includes 12 more confounders that are adjusted for in the linear models. In fitting a model for MV, the authors write (on page 86)

One of the major objectives in estimating the hedonic housing equation was to determine the best fitting functional form. Comparing models with either median value of owner-occupied homes (MV) or $\text{Log}(\text{MV})$ as the dependent variable, we found that the semilog version provided a slightly better fit. Using $\text{Log}(\text{MV})$ as the dependent variable, we concentrated

²The data in this paper seems to be wrongly coded in a few places. See [Gilley et al. \(1996\)](#) for details.

on estimating a nonlinear term in NOX, i.e., we included NOX^p in the equation, where p is an unknown parameter. . . .

The statistic fit in the equation was best when p was set equal to 2.0, i.e., when NOX^2 was in the equation. . . . The NOX variable has a negative sign and is highly significant.

The authors essentially explored the dataset to obtain the “right” transformations for the response and the covariate of interest. Further they ignore the exploration and report the statistical significance of coefficient of NOX without any adjustment.

1.2.2 Case Study L2: [Whittingham et al. \(2006\)](#)

The authors of this paper do not use classical inference after data exploration but show that this practice with stepwise form of data exploration is more prevalent in ecology and animal behavior. The authors describe the dangers of using stepwise selection methods for inference and surveys several papers from the literature to illustrate that this is common practice. On page 1184, the authors write

A second problem with stepwise multiple regression is more widely recognized and yet appears not to have deterred many ecologists from using the technique. . . . In particular, it is easy to overlook the fact that a single stepwise regression does not represent one hypothesis test but, rather, involves a large number of tests. This inevitably inflates the probability of Type I errors (false positive results). . . . Finally, owing to the selection of variables to include on the basis of the observed data, the distribution of the F-statistic is also affected, invalidating tests of the overall statistical significance of the final model.

For a similar paper with practical recommendations, see [Lydersen \(2014\)](#). For a

paper recommending variable selection without any indication of its consequences, see [Chowdhury and Turin \(2020\)](#).

1.2.3 Case Study L3: [Wiens et al. \(2015\)](#)

The authors of the paper develop models to predict post-discharge mortality which is defined as a binary random variable taking value 1 if the child dies within six months of discharge. The statistical analysis is based on 1307 enrolled participants. In the statistical analysis section of the paper, the authors write³

All variables were assessed using **univariate logistic regression to determine their level of association with the primary outcome**. Continuous variables were assessed for model fit using the Hosmer-Lemeshow test. Missing data was imputed by the method of multivariate imputation using chained equations. Following univariate analysis, candidate models were generated using a **stepwise selection procedure minimizing Akaike’s Information Criterion (AIC)**. This method is considered asymptotically equivalent to cross-validation and bootstrapping. All models generated in this sequence having **AIC values within 10% of the lowest value** were considered as reasonable candidates. The final selection of a model was judged on **model parsimony** (the simpler the better), **availability of the predictors** (with respect to minimal resources and cost), and the **attained sensitivity** (with at least 50% specificity).

It is easy to recognize that the the process described above is not reproducible (because of subjectivity and incomplete details) and hence all the inference for the models reported in Table 3 of the paper lack validity.

³Emphasis added here.

1.2.4 Case Study L4: Nersisyan et al. (2019)

The authors of this paper study the inheritance patterns of telomere length. Telomere is the end cap of a chromosome (as shown in Figure 1.3) and reduces in length as the cell divides. Hence, the length of the telomere acts a biological age of a person. The

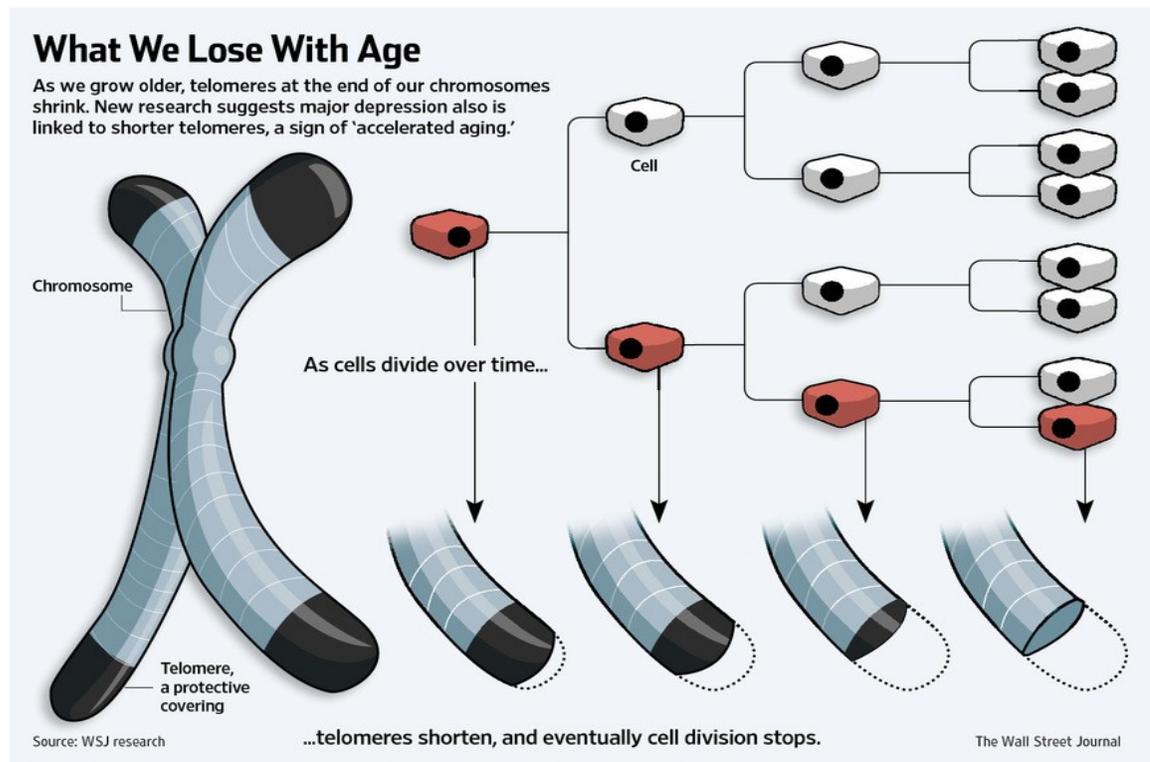


Figure 1.3: Telomere Length

first problem the authors study is associating an offspring's mean telomere length (MTL) to the age and sex of the offspring along with the telomere length of the parents (mother's mMTL and father's fMTL) and the age of the parents at conception (mother's MAC and father's PAC). In developing the models, the regression analysis section of the methods in the paper mentions⁴

Multivariate linear regression (MLR) analysis was performed to evaluate

⁴Emphasis added here.

the correlation between MTL and age and sex in the studied population. . . . Two of the families with missing data for the mother were removed, and two families with discordant age differences at the time of data collection and at conception were also discarded. Overall, MLR were done on 246 families. A set of pairwise regressions on the predictors were performed to estimate dependence between variables, and interaction terms were introduced for correlated predictors. The MLR models were tested by **sequential introduction of predictors and interaction terms**. The best model was chosen based on maximization of the adjusted R square term: ultimately, from the **three best models with similar adjusted R squared values the simplest one was chosen**.

1.2.5 Case Study L5: [Bolt et al. \(2016\)](#)

The authors of this paper examine the variables that are significantly associated with communication in every day activities, or communicative participation, in adult survivors of head and neck cancer (HNC). In the statistical analysis section (page 1148) of the paper, the authors write

The associations of the 17 variables with communicative participation were examined with multiple linear regression analysis in SPSS, version 18.0 (IBM). Communicative participation, age, time since diagnosis, and self-reported cognitive function were continuous variables; all others were categorical variables. Throughout the process of backward stepwise regression, model fit was analyzed with an overall regression F statistic. Individual variables with regression coefficients significant at $P < .05$ were retained in the model.

Because the final selected set of variables are obtained through data exploration, they

cannot be confirmed as significantly associated variables using the classical tests.

1.3 Some Observations and Formulation of the Problem

In all of the works reported above, the method of data analysis constitutes the following: Have a question of interest, get the dataset, explore the data to find a good model to fit or find the subset of covariates to be used in the model or find the transformations for variables to be used in the model, and then fit the model to draw inference or statistical conclusions. For example, in the context of fitting a linear regression with a treatment variable. The question of interest could be “is there a non-zero treatment effect?” In presence of confounders, one might select a subset of confounders to be used in the final model or one might select a transformation for the response/confounders. Then fit the model with selected set of confounders and transformations.

The illustrative example discussed in the beginning of the chapter (Figure 1.1) shows that in this practice classical tests or confidence regions cannot be used for reliable conclusions. A reasonable mathematical formulation of the problem (in case of linear regression) could be as follows: Suppose we have observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$.

1. For each $M \subseteq \{1, 2, \dots, p\}$ corresponding to indices of covariates, define the “target” of estimation by

$$\beta_M := \arg \min_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(Y_i - X_{i,M}^\top \theta)^2].$$

2. Based on the data, select a subset $\hat{M} \subseteq \{1, 2, \dots, p\}$ of covariates using whatever

method of practitioner’s choice. (This freedom in the selection method should be allowed to solve the problems in the above practical scenarios.)

3. Calculate the estimator

$$\hat{\beta}_{\hat{M}} := \arg \min_{\theta \in \mathbb{R}^{|\hat{M}|}} \frac{1}{n} \sum_{i=1}^n (Y_i - X_{i,\hat{M}}^\top \theta)^2.$$

This estimator “targets” $\beta_{\hat{M}}$ (the evaluation of the map $M \mapsto \beta_M$ at $M = \hat{M}$), that is, $\|\hat{\beta}_{\hat{M}} - \beta_{\hat{M}}\| = o_p(1)$. This fact will be shown in later chapters. In all of the case studies presented before, the practitioners use the estimator $\hat{\beta}_{\hat{M}}$ for inference or statistical conclusions.

4. Because $\hat{\beta}_{\hat{M}}$ targets $\beta_{\hat{M}}$, inference based on $\hat{\beta}_{\hat{M}}$ is inference for $\beta_{\hat{M}}$ and hence the VIDE problem in this case is to construct a valid confidence region $\hat{\mathcal{R}}_{\hat{M}}$ that satisfies

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\beta_{\hat{M}} \in \hat{\mathcal{R}}_{\hat{M}} \right) \geq 1 - \alpha, \quad (1.2)$$

irrespective of how $\hat{M} \subseteq \{1, 2, \dots, p\}$ is obtained based on the data.

Selection of variables is only one of many outcomes of data exploration. As described above, variable transformation can also be seen as an outcome of data exploration. For each transformation $g : \mathbb{R} \rightarrow \mathbb{R}$, define the “target”

$$\beta_g := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(g(Y_i) - X_i^\top \theta)^2].$$

Similarly, the estimator $\hat{\beta}_g$ is obtained as

$$\hat{\beta}_g := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (g(Y_i) - X_i^\top \theta)^2.$$

Based on the data, the practitioner chooses a transformation $\hat{g} \in \mathcal{G}$ from a class of transformations. The class of Box-Cox transformations is one such example: $\{y \mapsto (y^\lambda - 1)/\lambda : \lambda > 0\}$. The VIDE problem in this case is to construct a valid confidence region $\hat{\mathcal{R}}_{\hat{g}}$ for $\beta_{\hat{g}}$ in that it satisfies

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\beta_{\hat{g}} \in \hat{\mathcal{R}}_{\hat{g}} \right) \geq 1 - \alpha, \quad (1.3)$$

irrespective of how $\hat{g} \in \mathcal{G}$ is obtained based on the data.

The VIDE problems (1.2) and (1.3) represent the prototypical problems solved in this thesis. Extensions are possible to logistic, Poisson, and Cox regression models. An even more general VIDE problem can be described as follows. Suppose Z_1, \dots, Z_n are observations taking values in a set \mathcal{Z} . Consider a universe \mathcal{Q} of all possible selections and for every $q \in \mathcal{Q}$ define the estimator

$$\hat{\theta}_q := \arg \min_{\theta \in \Theta_q} \frac{1}{n} \sum_{i=1}^n \ell_q(\theta, Z_i),$$

for a loss function $\ell_q(\cdot, \cdot)$ and a “parameter” set Θ_q (that could possibly depend on q). The data analyst can now choose an element $\hat{q} \in \mathcal{Q}$ and the inference is to be based on the estimator $\hat{\theta}_{\hat{q}}$. The VIDE problem is to construct a confidence region $\hat{\mathcal{R}}_{\hat{q}}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\theta_{\hat{q}} \in \hat{\mathcal{R}}_{\hat{q}} \right) \geq 1 - \alpha, \quad (1.4)$$

irrespective of how $\hat{q} \in \mathcal{Q}$ is chosen based on the data. Here the “target” $\theta_{\hat{q}}$ is defined as the evaluation of the map $q \mapsto \theta_q$, at $q = \hat{q}$, given by

$$\theta_q := \arg \min_{\theta \in \Theta_q} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ell_q(\theta, Z_i)].$$

Both covariate selection and transformation selection are special cases as follows:

- For covariate selection, take $Z_i = (X_i, Y_i)$, $\mathcal{Q} = \{M : M \subseteq \{1, 2, \dots, p\}\}$, for $q = M \in \mathcal{Q}$, $\Theta_q = \mathbb{R}^{|M|}$, and $\hat{\theta}_q = \hat{\beta}_M$.
- For transformation selection, take $\mathcal{Q} = \{g : \mathbb{R} \rightarrow \mathbb{R} : g \in \mathcal{G}\}$, for $q = g \in \mathcal{G}$, $\Theta_q = \mathbb{R}^p$, and $\hat{\theta}_q = \hat{\beta}_g$.

The most important assumption of the VIDE framework is that the universe of estimators $\{\hat{\theta}_q : q \in \mathcal{Q}\}$ is prefixed and is not allowed to depend on the data; it must not be data dependent. For instance, one cannot choose $\ell_q(\cdot, \cdot)$, or Θ_q , or \mathcal{Q} based on the data.

Some Limitations of the Framework The formulation of the VIDE problem in (1.4) is very general but still has some limitations and does not cover certain types of exploration that would be considered reasonable/intuitive.

For instance, consider the following data exploration. Start with the observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$. Explore the data to find that modeling the response in terms of the covariates is not enough and pairwise interactions are needed to get a better model. Then choose a model $\hat{M} \subseteq \{1, 2, \dots, p, (1, 2), (1, 3), \dots, (p-1, p)\}$. Perform linear regression and draw some statistical conclusions. This does not fit into the problem formulation of (1.2) and (1.3), the reason being that the decision of adding more covariates (such as pairwise interactions) is based on the data and the analyst could have taken a decision of adding more transformed variables instead of interactions. It will be shown in Chapter 4 that the VIDE problem is impossible to solve for these more general data exploration procedures.

End of Chapter 1.

Chapter 2

Outline of the Thesis

In this chapter, we provide a basic introduction to the post-selection inference problem along with discussions of relevant literature. Section 2.1 provides generalities about the VIDE problem along with background from the literature. Section 2.2 describes how we approach the post-selection problem with ‘simultaneous inference.’ Section 2.3 presents an alternative approach based on ‘selective inference.’ Section 2.4 discusses briefly a solution based on ‘sample splitting.’ Section 2.5 gives previews of the remaining chapters.

2.1 General Introduction to VIDE and Some Review

In recent times, there has been a crisis in the sciences because too many published research results are found to lack reproducibility. Some of this crisis has been attributed to a failure of statistical methods to account for data-dependent exploration and modeling that precedes statistical inference. Data-dependent actions such as selection of subsets of cases, of covariates, of responses, of transformations and of model types has been aptly named “researcher degrees of freedom” (Simmons et al., 2011),

and these may well be significant contributing factors in the current crisis. Classical statistics does not account for them because it is built on a framework where all modeling decisions are made *independently of the data on which inference is based*. But if the data are in fact used to this end prior to statistical inference, then such inference loses its justifications and the ensuing validity conferred on it by classical theories. It is therefore critical that the theory of statistical inference be brought up to date to account for data-driven modeling. Updating the theory that justifies statistical inferences usually requires modifying the procedures of inference such as hypothesis tests and confidence intervals. As a consequence, the new procedures may lose some power relative to the previously stipulated but illusionary power derived from classical theories. This is a necessary price to be paid for better justification of statistical inference in the context of the pre-inferential liberties taken in today's data-analytic practice. While updating of statistical theories and inference procedures will not solve all problems underlying the current crisis, it is a necessary step as it may help mitigate at least some aspects of the crisis. In what follows we refer to all data-analytic decisions that are made using the data prior to inference as "data-driven modeling."

A second issue with theories of classical statistical inference is that many of them rely on the assumption that the data have been correctly modeled in a probabilistic sense. This means the theories tend to assume that the probability model used for the data correctly captures the observable features of the data generating process. Justifications of statistical inferences derived from such theories may therefore be invalid if the model is incorrect or (using the technical term) "misspecified." With the proliferation of data-analytic approaches in science and business, it is becoming ever more unrealistic to assume that all statistical models are correctly specified and inferences are made only after carefully vetting the model for correct specification, for example, using model diagnostics. Such vetting may never have been realistic in the first place,

and it should also be said that pre-inferential diagnostics should be counted among “researcher degrees of freedom” as they may result in data-driven modeling decisions. It is therefore a mandate of realism to look for so-called “model-robust” methods of statistical inference, and for statistical theory to provide their justifications. In matters of misspecification the situation is somewhat less dire than data-driven modeling as there exists a rich literature on the study of inference when models are misspecified. We will naturally draw on extant proposals for misspecification-robust or (using the technical term) “model-robust” inference and adapt them to our purposes.

To summarize, there exist at least two ways in which inference methods derived from classical mathematical statistics can be invalidated, namely,

(P1) data-driven modeling prior to statistical inference, and

(P2) model misspecification.

In light of the reproducibility crisis in the sciences, it is of considerable interest, even urgency, to develop methods of statistical inference and associated theoretical justifications that account for both **(P1)** and **(P2)**. Even though these problems are manifest in almost all statistical procedures used in practice, it is no simple task to provide methods of valid statistical inference that address these problems in greater generality. For this reason the present article puts forth specifically a method of valid inference for the case that the fitting procedure is ordinary least squares (OLS) linear regression. Here there exists a literature that documents the drastic effects of ignoring **(P1)** and **(P2)**; see, for example, [Buehler and Feddersen \(1963\)](#), [Olshen \(1973\)](#), [Rencher and Pun \(1980\)](#), and [Freedman \(1983\)](#). We will address one particular form of problem **(P1)**, namely, data-driven selection of regressor variables/covariates, and we will deal with several forms of problem **(P2)**.

Some of the earliest work that studies estimators under data-dependent modeling

(P1) include [Hjort and Claeskens \(2003\)](#) and [Claeskens and Carroll \(2007\)](#). Although these articles deal with a general class of statistical procedures, a major limitation, in view of the current article, is that the data-dependent modeling is restricted to a very narrow class of principled variable selection methods such as optimization of AIC or some other information criterion. The fact is, however, that few data analysts will confine themselves to a strict protocol of data-driven modeling. To address broader aspects of “researcher degrees of freedom” there have more recently emerged proposals that provide validity of statistical inference in the case of arbitrary data-driven selection of covariates. The first such proposal was by [Berk et al. \(2013\)](#) who solve the problem allowing misspecified response means but retaining the classical assumptions of homoskedastic and normally distributed errors. We refer to [Berk et al. \(2013\)](#) for many other prior works related to problem (P1) where data-driven modeling consists of selection of covariates. A more recent article that expands on [Berk et al. \(2013\)](#) is by [Bachoc et al. \(2016\)](#). An alternative approach is by [Lee et al. \(2016\)](#), [Tibshirani et al. \(2016\)](#), [Tian et al. \(2016\)](#) (for example). Similar to [Hjort and Claeskens \(2003\)](#), these proposals do not insure validity of inference against arbitrary covariate selection but against specific selection methods such as the lasso or stepwise forward selection. This type of post-selection inference is conditional on the selected model and dependent on distributional assumptions, thereby not addressing problem (P2).

Most of the above mentioned solutions related to problem (P1) have taken certain correct model or distributional assumptions for granted and can easily break down once such assumptions fail, which brings us to problem (P2). In modern terminology of mathematical statistics, one would need semi-parametric inference, not parametric inference when dealing with many of the classical inference procedures under misspecification. To expand on this, when considering a linear regression model, the classical

framework assumes that the conditional distribution is completely known except for the slope vector parameter and possibly the conditional homoscedastic variance parameter. This is a traditional parametric model interpretation of linear regression. Alternatively, linear regression can be interpreted as an algorithm that fits a linear function as (only) an approximation of the conditional expectation function with no other assumptions in the sense that the joint (and so conditional) distributions are left completely unspecified. From this viewpoint, the problem has a parametric component of interest and an infinite dimensional nuisance parameter of no or secondary interest, thus representing a semi-parametric problem. For more on this view, see [Kosorok \(2008\)](#) and Section 3.A of Chapter 3. The inference differs with the change in the view of linear regression. This was pointed out and expanded further to other problems with an indication of correct inference in [Berk et al. \(2014\)](#) and [Buja et al. \(2016\)](#). [Buja et al. \(2016\)](#) refers to this semi-parametric view as “assumption-lean linear regression”. Our setting is much more general than this framework as shown in Chapter 3. The only reference that provides a method of valid post-selection inference taking into account both **(P1)** and **(P2)** is [Bachoc et al. \(2016\)](#), although it remains with an interpretation of covariates as fixed.

As indicated above, there have been several approaches that attempt to provide solutions to VIDE. They can be characterized by the following terms, to be explained below:

1. Simultaneous Inference,
2. Selective Inference, and,
3. Sample Splitting.

We will discuss these approaches in the following sections in turn.

2.2 Simultaneous Inference Approach to VIDE

Simultaneous inference approach or the uniform inference approach is the one proposed by [Berk et al. \(2013\)](#) and extended by [Bachoc et al. \(2016\)](#). The basic idea behind this approach is to turn the problem of valid post-selection inference into a simultaneous inference problem. Suppose $\{\theta_q : q \in \mathcal{Q}\}$ are a set of real-valued parameters (or functionals) indexed by the elements of \mathcal{Q} . Based on the data, the analyst selects an element $\hat{q} \in \mathcal{Q}$ and uses $\hat{\theta}_{\hat{q}}$ as an estimator of $\theta_{\hat{q}}$. Without specific assumptions on the selection procedure, all one can say is that $\hat{\theta}_{\hat{q}}$ is estimating $\theta_{\hat{q}}$ and planning to use $\hat{\theta}_{\hat{q}}$ for inference is same as trying to infer about $\theta_{\hat{q}}$. To form a confidence region for $\theta_{\hat{q}}$, simultaneous inference approach constructs the set of confidence regions $\{\hat{\mathcal{R}}_q : q \in \mathcal{Q}\}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{q \in \mathcal{Q}} \{\theta_q \in \hat{\mathcal{R}}_q\} \right) \geq 1 - \alpha, \quad (2.1)$$

which would readily imply for any $\hat{q} \in \mathcal{Q}$ that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\theta_{\hat{q}} \in \hat{\mathcal{R}}_{\hat{q}} \right) \geq 1 - \alpha. \quad (2.2)$$

To construct the set of confidence regions satisfying (2.1), [Berk et al. \(2013\)](#) and [Bachoc et al. \(2016\)](#) construct the quantiles of the maximum of the test statistics used for testing hypothesis about θ_q . Precisely, let $T_q(\theta_q)$ be a test statistic used for testing hypothesis about θ_q . Without loss of generality, assume that $T_q(\theta_q)$ has an asymptotic mean zero and asymptotic variance 1. Now a procedure to solve the multiple hypothesis testing problem about $H_q : \theta_q = \theta_{0,q}$ for $q \in \mathcal{Q}$ can be based on

the test statistic

$$T_{\mathcal{Q}} := \max_{q \in \mathcal{Q}} |T_q(\theta_{0,q})|.$$

A natural test statistic $T_q(\theta_q)$ is given by

$$T_q(\theta_q) := \frac{\hat{\theta}_q - \theta_q}{\hat{\sigma}_q(\hat{\theta}_q)},$$

where $\hat{\sigma}_q(\hat{\theta}_q)$ is an estimator of the asymptotic standard deviation of $\hat{\theta}_q$. Note that there is no special reason for \mathcal{Q} to be finite or even countable in this setting, although, neither [Berk et al. \(2013\)](#) or [Bachoc et al. \(2016\)](#) discuss their approach in this setting. The set of confidence regions satisfying (2.1) can be constructed by inverting this test.

Simply stating the simultaneous approach to post-selection inference tests for all possible parameters or functionals before the analyst chooses one of them randomly. By doing this, this approach can account for all types of selection and does not require any special properties of \hat{q} .

It is often desirable to have an infimum over a set of possible distributions of the observations after \liminf in (2.2) so that a form of uniformity holds and the confidence guarantee holds even under slight deviations of the true distributions; see [Pötscher \(2002\)](#) and [Leeb and Pötscher \(2005\)](#) for reasons on why this is desired. The set of possible distributions usually includes all distributions satisfying certain moment restrictions. In the settings considered by [Berk et al. \(2013\)](#) and [Bachoc et al. \(2016\)](#), this uniformity holds by construction. This thesis is based on the simultaneous approach to post-selection inference. In the literature this approach is sometimes referred to as uniform inference (possibly because a maximum is taken in the test statistic). More details related to this approach can be found in Chapter 4.

2.3 Selective Inference Approach to VIDE

The setting of the problem is same as in the previous section. Selective inference approach to post-selection inference has originated from Stanford. In Selective inference too, the goal is to construct confidence region $\hat{\mathcal{R}}_{\hat{q}}$ for a randomly chosen \hat{q} based on the data to satisfy (2.2). Instead of turning to the simultaneous statement like in (2.1), selective inference constructs $\hat{\mathcal{R}}_{\hat{q}}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\theta_{\hat{q}} \in \hat{\mathcal{R}}_{\hat{q}} | \hat{q} \right) \geq 1 - \alpha. \quad (2.3)$$

For this reason, sometimes this approach is referred to as a conditional inference. This approach, as of now seen in the literature, is slightly restrictive in that an application of this approach requires a special structure on the estimator and the selection procedure. The construction of $\hat{\mathcal{R}}_{\hat{q}}$ proceeds by first approximating the conditional distribution of the data given $\hat{q} = q$ for any $q \in \mathcal{Q}$ and then using this conditional distribution to get the conditional distribution of $\hat{\theta}_{\hat{q}}$ leading to the confidence region. See [Lee et al. \(2016\)](#), [Tibshirani et al. \(2016\)](#), [Tian et al. \(2016\)](#), and [Tian and Taylor \(2017\)](#) for more details related to this approach.

Here too it is desirable to have an infimum over a set of all distributions after \liminf in (2.3) and [Tibshirani et al. \(2018\)](#) proved such uniformity results. Also, see [Leeb and Pötscher \(2006\)](#) for some complementary impossibility results.

2.4 Sample Splitting Approach to VIDE

A classical and possibly the oldest solution for post-selection problems is sample splitting. The basic idea is split the available sample into two parts: training and test data. These could be of different sizes but usually taken to be of almost equal

sizes. Use the training data to explore the data and select \hat{q} . Once the selection is done, ignore the training data and compute the estimator $\hat{\theta}_{\hat{q}}$ based on the test data with \hat{q} from the training data. Because \hat{q} is independent of the test data (when the sample consists of independent observations),

$$\mathbb{P}\left(\hat{\theta}_{\hat{q}} - \theta_{\hat{q}} \in A \mid \hat{q} = q\right) = \mathbb{P}\left(\hat{\theta}_q - \theta_q \in A\right) \quad \text{for all Borel sets } A.$$

This implies that the usual asymptotics work as expected on the test data as if no selection was performed. A detailed presentation of sample splitting as a solution of post-selection inference problem was given in [Zhang \(2012, Chapter 2\)](#). Sample splitting in light of increasing dimension is discussed thoroughly in [Rinaldo et al. \(2016\)](#). There are three main disadvantages of sample splitting in comparison to the contents of this thesis:

- Sample splitting only gives one shot at selection. In practice, it so often happens that a set of variables is selected using a particular method and the estimator obtained from this selection (based on the test data) does not satisfy the analyst’s “criterion.” In such cases a second shot at selection is required. With sample splitting, it is hard to solve this kind of sequential selection problem.
- Sample splitting is possibly not the best use of data ([Fithian et al., 2014](#)). In the context of current applications where big data is too often encountered, the analyst cannot spare any data. By splitting the data, only a part of data is being used for inference. Of course, the other half is being used for selection but the final confidence intervals or tests are based on only a part of the data. Even though sample splitting is a simple method to avoid invalid statistical conclusions, the analyst might be tempted to use the full data multiple times for selection and inference.

- Sample splitting is invalid for dependent data. Sample splitting inherently assumes independence of observations in the data. If the observations are dependent then sample splitting is invalid and no such simple alternative exists yet. As mentioned previously, dependent data are also in the realm of this thesis and we are looking for a unified solution that applies, in principle, to various settings. Recently, [Lunde \(2019\)](#) proved that sample splitting guarantees can be extended to weakly dependent data. The subject, however, is not mature enough to apply the results for the high-dimensional case.

There are few more minor issues with sample splitting. The effect of split sizes is not clear in many problems and the clear guideline for a choice is not present. The randomness in the split sample also causes trouble with interpretation since a change in the split sample, there can be a change in the selection and so the target of estimation. Note that this effect of randomness is different from that of the randomness in bootstrap or subsampling. In bootstrap or subsampling, the randomness disappears with the number of replications diverging but in sample splitting it does not. The quantity being estimated using test data changes with every split sample.

2.5 Outline of the Thesis

The remaining thesis is organized as follows. As mentioned in the abstract, misspecification is a natural outcome of selection based on data. For this reason, it is important to revisit the classical topics of linear regression in light of misspecified models and understand what are the main departures from the classical theory and how should one do inference in misspecified models. This is the main topic of Chapter 3. Here the discussion is restricted to linear regression since the estimator is explicitly known and many calculations can be done simply. Very similar results, however, continue to hold for general M -estimation problems as shown in the later chapters. The dis-

cussion here is very closely related to the contents of [Buja et al. \(2014\)](#) and [Buja et al. \(2016\)](#), but is more general than the discussion in those papers. We also discuss variance and distribution estimation of the least squares linear regression estimator using two different types of bootstrap in Chapter 3.

A general formulation of valid post-selection inference problem in linear regression is provided in Chapter 4. Chapter 4 provides a method of valid post-selection inference using the idea of simultaneous confidence regions in a way different from the discussion above and that of [Berk et al. \(2013\)](#), [Bachoc et al. \(2016\)](#). This is a special construction for linear regression and is difficult to extend to other M -estimation problems. The main advantage of these confidence regions is that the computation is not NP-hard as is the case with the construction in Section 2.2. The implementation of these regions is based on the high-dimensional central limit theorem and multiplier bootstrap.

The method of post-selection inference discussed in Section 4.2 is the most general construction of valid post-selection confidence regions. A unified framework of proving validity of these confidence regions is the main focus of Chapter 5. This unified framework encompasses the settings of [Berk et al. \(2013\)](#) and [Bachoc et al. \(2016\)](#). One of the main assumptions of this unified framework is an asymptotic uniform linear representation of the estimators around the target. This means that there exists estimators $\{\hat{\theta}_q : q \in \mathcal{Q}\}$ and functions $\{\psi_q^{(i)}(\cdot) : q \in \mathcal{Q}, 1 \leq i \leq n\}$ such that,

$$\max_{q \in \mathcal{Q}} \left| \Psi_{n,q}^{-1/2} \left(\hat{\theta}_q - \theta_q - \frac{1}{n} \sum_{i=1}^n \psi_q^{(i)}(Z_i) \right) \right| = o_p(n^{-1/2}), \quad (2.4)$$

where

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} [\psi_q^{(i)}(Z_i)] = 0 \quad \text{for all } q \in \mathcal{Q}, \quad \text{and} \quad \Psi_{n,q} := \frac{1}{n} \sum_{i=1}^n \text{Var} (\psi_q^{(i)}(Z_i)).$$

This assumption (with \mathcal{Q} a singleton) is an integral part of the usual asymptotic normality proofs for M -estimation problems. An application of this framework for linear regression is also given in Chapter 5 where the assumptions of the unified framework are verified under some tail assumption on the observations. Succinctly, the “three line proof” of how this unified framework for the case of covariate selection works as illustrated in Figure 2.1.

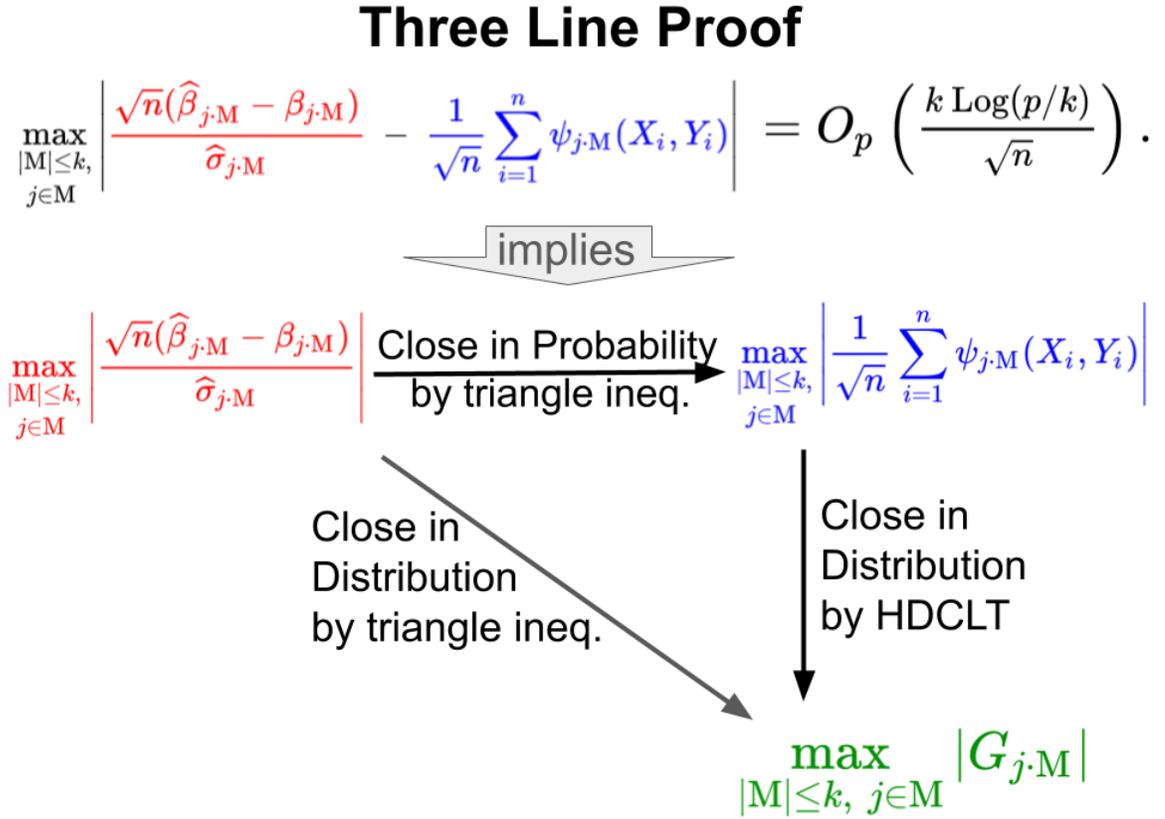


Figure 2.1: Illustration of the Unified Framework. Here $G_{j \cdot \mathbf{M}}$ is a standard Gaussian random variable and $(G_{j \cdot \mathbf{M}})_{j \cdot \mathbf{M}}$ is a Gaussian random vector with covariance matching that of the vector of averages.

In Chapter 5, we verify all the assumptions of this unified framework for the case of linear regression and we provide references where the assumptions are verified for a class of general M -estimation problems with a convex loss function. The general

class includes all the generalized linear models with both canonical and non-canonical link functions, in particular, logistic, Poisson, negative binomial regression models.

All the technical results here are written with the focus of valid post-selection inference. There is a “spin-off” of this framework that can be written with the focus of proving asymptotics for post-selection estimators. Suppose \hat{q} is a random choice based on the data satisfying $\mathfrak{D}(\hat{q}, q_0) = o_p(1)$ for some distance $\mathfrak{D}(\cdot, \cdot)$ on \mathcal{Q} , then under assumption (2.4), it follows that

$$\sqrt{n} \left(\hat{\theta}_{\hat{q}} - \theta_{\hat{q}} \right) \xrightarrow{\mathcal{L}} N(0, \Psi_{n, q_0}),$$

as long as $\Psi_{n, q} - \Psi_{n, q_0} = o(1)$ for $\mathfrak{D}(q, q_0) = o(1)$. Note that the centering for $\hat{\theta}_{\hat{q}}$ is $\theta_{\hat{q}}$ not θ_{q_0} . In general, this cannot be improved in the sense that the result does not hold with θ_{q_0} as centering. In case of variable selection, $\theta_q - \theta_{q_0} = 0$ for $\mathfrak{D}(q, q_0) = o(1)$ since $\hat{q} = q_0$ eventually. In case of continuous q , the centering can be replaced by proving a rate of convergence of $\mathfrak{D}(\hat{q}, q_0)$ to zero and then a Hölder-type continuity of $q \mapsto \theta_q$.

Till this part of the thesis, all the chapters are mostly technical in nature and in Chapter 6, real data examples are provided demonstrating the application of all the post-selection confidence regions described. Even though the approach presented in Chapter 4 is computationally simple, it is NOT the case with the unified framework where the computation is NP-hard.

Finally, the thesis ends with some concluding remarks in Chapter 7.

End of Chapter 2.

Model-free Study of Ordinary Least Squares Linear Regression

Arun K. Kuchibhotla, Lawrence D. Brown, and Andreas Buja

University of Pennsylvania
e-mail: arunku@wharton.upenn.edu

Abstract: Ordinary least squares (OLS) linear regression is one of the most basic statistical techniques for data analysis. In the main stream literature and the statistical education, the study of linear regression is typically restricted to the case where the covariates are fixed, errors are mean zero Gaussians with variance independent of the (fixed) covariates. Even though OLS has been studied under misspecification from as early as the 1960's, the implications have not yet caught up with the main stream literature and applied sciences. The present article is an attempt at a unified viewpoint that makes the various implications of misspecification stand out.

1. Introduction and Motivation

The aim of this article is to provide what we call an “upside down analysis” for linear regression. While traditional linear regression analysis starts with assumptions such as fixed covariates as well as linearity and Gaussian errors, upside down analysis starts with a given estimator – OLS in this case – and finds the most general conditions under which the estimator “works” in the sense that it has a well-defined target and permits inference. In our upside down analysis, essentially all we need is a form of law of large numbers (LLN) and a central limit theorem (CLT) for second moments of the response and the covariates. Such LLNs and CLTs are satisfied in numerous situations, including strong mixing random variables, martingales, Markov chains, time series processes, . . . (see, e.g., chapters 3 and 5 of [White \(2001\)](#)). LLNs and CLTs can accommodate non-identical distributions of random vectors, a fact that turns out to be a particularly useful feature of the proposed analysis: It allows a treatment of fixed and random covariates in a unified way by thinking of fixed values of covariates as degenerate point mass distributions.

It should be mentioned here that most of the results presented in this article are known in the literature but are scattered. A unified treatment as given in

Figure 2.2: The following chapter is based on [Kuchibhotla et al. \(2018\)](#).

Chapter 3

Assumption-lean Framework for Linear Regression

The aim of this chapter is to provide what we call an “upside down analysis” for linear regression. While traditional linear regression analysis starts with assumptions such as fixed covariates as well as linearity and Gaussian errors, the upside down analysis starts with a given estimator – OLS in this case – and finds the most general conditions under which the estimator “works” in the sense that it has a well-defined target and permits inference. In our upside down analysis, essentially all we need is a form of the law of large numbers (LLN) and a central limit theorem (CLT) for second moments of the response and the covariates. Such LLNs and CLTs are satisfied in numerous situations, including strong mixing random variables, martingales, Markov chains, time series processes, . . . (see, e.g., chapters 3 and 5 of [White \(2001\)](#)). LLNs and CLTs can accommodate non-identical distributions of random vectors, a fact that turns out to be a particularly useful feature of the proposed analysis: It allows a unified treatment of fixed and random covariates by thinking of fixed values of covariates as degenerate point mass distributions.

It should be mentioned here that most of the results presented in this chapter are known in the literature but are scattered. A unified treatment as given in this chapter appears to be non-existent. Somewhat close in spirit but executing a traditional “upside up” analysis is by [White \(1980\)](#) who studies linear regression under the assumption of independence allowing for non-identical distributions. Our analysis sidesteps his assumption of absent correlation between covariates and errors. [Gallant and White \(1988\)](#) and [White \(2001\)](#) extend the analysis of [White \(1980\)](#) to certain dependence structures but remain traditional in that they define targets of estimation in terms of asymptotic limits for a fixed number of covariates, whereas we define sample size-dependent targets and allow the number of covariates to grow.

An essential difference of our “upside down” approach to these traditional treatments is that the latter assume the existence of a single target such that certain conditions are satisfied. For example, the traditional linear model assumes there exists a β_0 such that $Y_i = X_i'\beta_0 + \varepsilon_i$ and $\varepsilon_i \sim N(0, \sigma^2)$ iid, or, as in [White \(1980\)](#), $\mathbb{E}[X_i\varepsilon_i] = 0$ for all $i = 1 \dots n$. In contrast, we make no such assumptions; rather, we construct *sequences of targets* for the OLS procedure that are *intrinsic to OLS* without postulating a single target that is extraneous to the procedure. This is crucially possible by postulating LLNs and CLTs for the components of the normal equations (estimating equations).

This chapter is organized as follows. Section [3.1](#) describes the concept of “target of estimation” and provides the minimal assumptions under which the least squares estimator “works”. Even though the definition of the target can be done under very minimal assumptions, it is hard to proceed further to inference under such minimal assumptions. For this reason, we add an assumption on independence of observations to proceed. In Section [3.2](#), the problem is studied under the only assumption of fixed covariates and none of the other classical assumptions as mentioned above. In Section

3.3, the problem is studied under the assumption that the observations are independent and identically distributed random vectors. After a preliminary understanding of the problem in both fixed and random covariates, a unified framework is developed for the problem in Section 3.4 along with a normal approximation. To do inference (or more specifically confidence intervals), a “good” variance estimator is needed. Section 3.5 provides theory about “asymptotic” variance estimation and also bootstrap based variance estimation. In Section 3.6, the problem of testing hypothesis about the target of estimation is considered. We end this chapter with some concluding remarks in Section 3.7. A “non-technical” discussion of efficiency of estimators is included in appendix 3.A.

In what follows the random variables and their realized values are both denoted by capital letters such as X and Y . For any vector $v \in \mathbb{R}^q$, let $v(j)$ denote the j -th coordinate of v for $1 \leq j \leq q$. For any real-valued function $f(\cdot)$, $\arg \min_x f(x)$ denotes the set of all (global) minimizers of $f(\cdot)$ and the statement

$$x^* := \arg \min_x f(x),$$

should be understood as stating x^* is any element of the set of all minimizers of $f(\cdot)$. Throughout this chapter, the symbol C is used to denote a universal constant that can be different at different contexts.

3.1 Target of Estimation

Suppose $(X_i^\top, Y_i)^\top \in \mathbb{R}^p \times \mathbb{R}, 1 \leq i \leq n$ are random vectors obtained from n cases under study. A linear regression is performed on this data and assuming invertibility

of the matrix involved, the estimator of the “slope” $\hat{\beta}_n$ is given by

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i Y_i \right).$$

It is readily seen that $\hat{\beta}_n$ is a function of two averages: one is a matrix average and the other is a vector average. For notational convenience, let

$$\begin{aligned} \hat{\Sigma}_n &:= \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \in \mathbb{R}^p \times \mathbb{R}^p; \quad \text{and} \\ \hat{\Gamma}_n &:= \frac{1}{n} \sum_{i=1}^n X_i Y_i \in \mathbb{R}^p. \end{aligned}$$

In the classical linear regression theory one includes the linearity assumption $Y_i = X_i^\top \beta_0 + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i | X_i] = 0$. Under this assumption, it is easy to see that

$$\mathbb{E}[\hat{\beta}_n | X_1, \dots, X_n] = \beta_0 \quad \Rightarrow \quad \mathbb{E}[\hat{\beta}_n] = \beta_0.$$

Observe that independence of the observations is not required in this calculation. Since $\hat{\beta}_n$ is unbiased for β_0 , the estimator $\hat{\beta}_n$ can be thought of as estimating β_0 . The main question of this chapter (and also of the thesis) is “what is it estimating if the linearity assumption is not true?”.

As mentioned $\hat{\beta}_n$ is a function of two averages $\hat{\Sigma}_n$ and $\hat{\Gamma}_n$. If there exist a (non-random) matrix Σ_n and a (non-random) vector Γ_n such that as $n \rightarrow \infty$

$$\hat{\Sigma}_n - \Sigma_n = o_p(1) \quad \text{and} \quad \hat{\Gamma}_n - \Gamma_n = o_p(1), \tag{3.1}$$

then it is not unreasonable to expect that $\hat{\beta}_n$ is getting close to

$$\beta_n := \Sigma_n^{-1} \Gamma_n \quad (\text{assuming invertibility of } \Sigma_n),$$

in the sense that $\hat{\beta}_n - \beta_n = o_p(1)$. Indeed, this can be easily proven by Slutsky's theorem. There are many cases where assumption (3.1) holds true and some of these are listed below.

- If, for all $1 \leq i \neq j \leq n$ and $1 \leq l, m \leq p$, the random vectors satisfy

$$\begin{aligned} \text{Var}(X_i(l)X_i(m)) &< \infty \quad \text{and} \quad \text{Var}(X_i(l)Y_i) < \infty; \\ \text{Cov}(X_i(l)X_i(m), X_j(l)X_j(m)) &\leq 0 \quad \text{and} \quad \text{Cov}(X_i(l)Y_i, X_j(l)Y_j) \leq 0, \end{aligned} \quad (3.2)$$

and p is fixed (not changing with n) then the random vectors satisfy assumption (3.1) with

$$\Sigma_n := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i X_i^\top] \quad \text{and} \quad \Gamma_n := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i Y_i].$$

[A special case in (3.2) is when the observations are independent of each other. See Shao (2000) for a generalization of this condition.] The proof follows by proving that $\hat{\Sigma}_n - \Sigma_n$ and $\hat{\Gamma}_n - \Gamma_n$ converge coordinate-wise in probability to zero. Since p is fixed and does not change with n , it follows that they also converge to zero in any norm. The coordinate-wise convergence in probability can be shown by directly calculating the variance and proving that it converges to zero. See Theorem 2.2.1 of Durrett (2010). Assumption (3.2) essentially declares that the observations are “negatively associated”.

- If the random vectors $(X_i^\top, Y_i)^\top$ are independent of each other and satisfy

$$\max_{1 \leq i \leq n} \mathbb{E} \left[|X_i(l)X_i(m)|^{1+\delta} \right] < \infty \quad \text{for all} \quad 1 \leq l, m \leq p,$$

with p fixed, then the random vectors satisfy assumption (3.1) with $\Sigma_n = \mathbb{E}[\hat{\Sigma}_n]$

and $\Gamma_n = \mathbb{E}[\widehat{\Gamma}_n]$. The proof follows by using Theorem 3.7 and Corollary 3.9 of [White \(2001\)](#). The point of this example is that, under the independence assumption, boundedness of the fourth moment of $X_i(l)$ is not required; it can be reduced to $(1 + \delta)$ -th moment of $X_i^2(l)$.

Based on these calculations and consistency, we define the target of estimation as follows.

Definition 1. *If the random vectors are distributed in such a way that $\widehat{\beta}_n$ satisfies*

$$\left\| \widehat{\beta}_n - \beta_n \right\|_2 = o_p(1) \quad \text{as } n \rightarrow \infty,$$

*for some vector β_n , then we say $\widehat{\beta}_n$ is estimating β_n and the vector β_n is called the **target of estimation**.*

Remark 3.1.1 In classical mathematical statistics, one has a target of inference (or a parameter of interest) in mind and the goal is to estimate that parameter. In contrast, we start here with the estimator and analyze what it is estimating – which is then assigned as the target of estimation. This process is what we call an “upside down analysis”. This approach is also similar in spirit to the thinking in machine learning where a method of computation is introduced first rather than a model. A similar treatment similar can be found in Chapter 3 of [Pötscher and Prucha \(1997\)](#).

◇

Remark 3.1.2 The target of estimation β_n is allowed to depend on n, p and so can change when n (or p) is increased. Because of this feature, β_n might sometimes be referred to as a “moving target”. Just from the definition above, β_n is not unique in that one can always add a small constant (converging to zero) and that vector can still be called the target of estimation. In all the cases to be dealt with in the

subsequent chapters, the choice of the target of estimation will be clear and taking any of the equivalent ones does not change the story. Also, it is not required that $\{\beta_n\}$ as a sequence of non-random vectors converge to some (non-random) vector. \diamond

Remark 3.1.3 The choice of the Euclidean norm in the above definition is only for concreteness and can be replaced by any other norm depending on the context. The choice of norm only matters in so far as consistency in the sense of Definition 1 can be proven for some norms and not for others. This may be an issue when one allows p to grow at certain rates as a function of n . \diamond

The example settings and the calculations above have shown that the target of estimation is well-defined for linear regression in many cases. There is, however, nothing special about linear regression and the target of estimation can be easily derived for a large class of estimators (possibly inspired by a very different distributional model for the response). Note that the least squares estimator can also be defined as

$$\hat{\beta} = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \theta)^2.$$

The target of estimation in our example setting can be written as

$$\beta_n = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(Y_i - X_i^\top \theta)^2].$$

What is noteworthy in this representation is that the empirical objective function (based on the observations) got replaced by its expected value (or more generally the limit of the empirical objective). This is a pattern that holds in general problems. To elaborate, suppose $Z_i \in \mathbb{R}^q, 1 \leq i \leq n$ are random vectors obtained from n cases

under study and the estimator $\hat{\theta}$ obtained by solving the minimization problem

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \rho(Z_i, \theta),$$

is considered for some (loss) function $\rho : \mathbb{R}^q \times \mathbb{R}^k \rightarrow \mathbb{R}$. Then under mild conditions it can be proved that the target of estimation for $\hat{\theta}_n$ is θ_n given by the minimization problem

$$\theta_n := \arg \min_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\rho(Z_i, \theta)].$$

These kind of optimization is called an M -estimation problem. [Kuchibhotla \(2018\)](#) provides a proof of consistency for a specialized form of $\rho(\cdot, \cdot)$, but the techniques there extend to any loss function. For the rest of this chapter and the next, we continue with linear regression since it has a simple estimator that is known in closed form and hence many properties are easier to analyze. It should, however, be understood that most of the techniques here do generalize to arbitrary M -estimation problems.

In the two sections to follow the problem of linear regression is considered under two settings:

1. independent random vectors with fixed covariates; and
2. independent and identically distributed random vectors.

We provide only a preliminary analysis, and a more complete study is considered in the unified framework of Section [3.4](#) which includes both these settings as special cases. One of the main ingredients in this analysis is the multidimensional Berry-Esseen bound from [Bentkus \(2004\)](#).

Theorem 1 (Berry-Esseen Bound; Theorem 1.1 of [Bentkus \(2004\)](#)). *Suppose W_1, \dots, W_n are independent mean zero random vectors in \mathbb{R}^d . Then there exists a universal con-*

stant $C > 0$ such that

$$\sup_{A \in \mathcal{C}_d} \left| \mathbb{P} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \in A \right) - \mathbb{P} (N(0, \Upsilon_n) \in A) \right| \leq C \frac{d^{1/4}}{n^{1/2}} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|\Upsilon_n^{-1/2} W_k\|_2^3 \right] \right),$$

where \mathcal{C}_d denotes the set of all convex subsets of \mathbb{R}^d and

$$\Upsilon_n := \frac{1}{n} \sum_{i=1}^n \text{Var}(W_i).$$

3.2 Linear Regression with Fixed Covariates

In this section, we consider the problem of linear regression under the assumption that the covariates are fixed (non-random) constants. As mentioned before, this is one of the classical assumptions related to linear and generalized linear models. For simplicity, let the observations be denoted by $(x_i^\top, Y_i)^\top \in \mathbb{R}^p \times \mathbb{R}$, $1 \leq i \leq n$. The covariates are written in lower case to emphasize that they are fixed. And we assume the Y_i 's are independent random variables. The least squares estimator is given by

$$\hat{\beta}_n := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \{Y_i - x_i^\top \theta\}^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i Y_i \right).$$

The target of estimation is then given by

$$\beta_n := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(Y_i - x_i^\top \theta)^2] = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i \mathbb{E} [Y_i] \right), \quad (3.3)$$

where the expectation is taken with respect to the measure of Y_i . For simplicity let

$$\mu_i := \mathbb{E} [Y_i] \quad \text{and} \quad \Sigma_n := \frac{1}{n} \sum_{i=1}^n x_i x_i^\top. \quad (3.4)$$

Note that because of fixed covariates we have $\widehat{\Sigma}_n = \Sigma_n$ for all n . It follows that

$$\sqrt{n} \left(\widehat{\beta}_n - \beta_n \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma_n^{-1} x_i [Y_i - \mu_i] \quad (3.5)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma_n^{-1} x_i [Y_i - x_i^\top \beta_n]. \quad (3.6)$$

The first equation (3.5) is specific to the fixed covariate setting, while the second equation (3.6) is valid irrespective of whether x_i 's are fixed or random. It is also important to note that the summands in (3.5) are mean zero while the ones in (3.6) are not. This follows from the population normal equations obtained by differentiating the objective function (3.3) defining the target β_n :

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} [x_i (Y_i - x_i^\top \beta_n)] = 0.$$

(Without any further assumptions, there is no reason for the individual summand expectations to be zero.)

Since the covariates are fixed, it is clear that $\sqrt{n} \left(\widehat{\beta}_n - \beta_n \right)$ is a scaled average of independent random vectors and the expectation of the average is zero. Therefore, by the multidimensional Berry-Esseen bound (Theorem 1), we obtain

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(\sqrt{n} \left(\widehat{\beta}_n - \beta_n \right) \in A \right) - \mathbb{P} (Z \in A) \right| \leq C \frac{p^{1/4}}{n^{1/2}} \gamma_n,$$

where \mathcal{C}_p is the set of all convex subsets of \mathbb{R}^p , Z is a Gaussian random vector with mean zero, and the variance Ψ_n given by

$$\Psi_n := n \text{Var}(\widehat{\beta}_n - \beta_n) = \Sigma_n^{-1} K_n \Sigma_n^{-1}, \quad K_n := \frac{1}{n} \sum_{i=1}^n x_i x_i^\top \text{Var}(Y_i).$$

Here, γ_n is defined as

$$\gamma_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [|Y_i - \mu_i|^3] \|\Psi_n^{-1} \Sigma_n^{-1} x_i\|_2^3 = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [|Y_i - \mu_i|^3] \|\Sigma_n K_n^{-1} x_i\|_2^3.$$

Under certain (rate) assumptions on p , this implies that

$$\sqrt{n} (\hat{\beta}_n - \beta_n) \stackrel{\mathcal{L}}{\approx} N(0, \Sigma_n^{-1} K_n \Sigma_n^{-1}).$$

We used the notation $\stackrel{\mathcal{L}}{\approx}$ to denote approximation in law (or distribution). To summarize, all we need to assume for this asymptotic convergence result is the finiteness of the third central moment of Y_i and non-singularity of some matrices. By comparison, classical linear regression analysis based on fixed covariates and homoscedastic Gaussian errors requires the assumption of linearity of the mean response in order to be valid. In particular, Σ_n^{-1}/n defined in (3.4) is *not* the variance of $\hat{\beta}_n$.

In order to do inference using the estimator $\hat{\beta}_n$, one should be able to estimate the asymptotic variance Ψ_n . Note that the Σ_n factors of Ψ_n are known and need not be estimated. All we need to estimate is K_n , the variance of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i Y_i.$$

By the multidimensional Berry-Esseen bound, we get

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i [Y_i - \mu_i] \in A \right) - \mathbb{P} (N(0, K_n) \in A) \right| \leq C \frac{p^{1/4}}{n^{1/2}} \alpha_n,$$

where

$$\alpha_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [|Y_i - \mu_i|^3] \|K_n^{-1} x_i\|_2^3.$$

In general, the summands $x_i Y_i$ are non-identically distributed, even if Y_i 's are iden-

tically distributed. Since we do not know their true expectations, it is *impossible* to estimate the variance K_n . (See Section 3.5 for details, and Liu and Singh (1995) for a related problem.) It *is*, however, possible to construct a conservative estimator of K_n . This construction will be described in Section 3.4 (see Fahrmeir (1990, page 492), and also Bachoc et al. (2016) for an alternative proposal).

Remark 3.2.1 The comment about the impossibility of estimation of “asymptotic” variance should be understood carefully. The impossibility mentioned here is in the general context of fixed covariates with no more model assumptions than independence of observations. In fact, if it is additionally assumed that $\text{Var}(Y_i) = \sigma^2(x_i)$ for some continuous function $\sigma(\cdot)$, then the matrix K_n can be estimated consistently by non-parametrically estimating the function $\sigma(\cdot)$ (see, e.g., Abadie et al. (2014)). \diamond

3.3 Linear Regression with Random Covariates

Suppose we have n subjects producing observations $(X_i^\top, Y_i)^\top \in \mathbb{R}^p \times \mathbb{R}$, $1 \leq i \leq n$ and we apply linear regression on this data. In this section, we assume that these observations are random vectors that are not only independent but also identically distributed. Let $(X^\top, Y)^\top$ be a generic random vector that is identically distributed with the observations. The least squares estimator is still given by

$$\hat{\beta}_n := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \{Y_i - X_i^\top \theta\}^2 = \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i Y_i \right). \quad (3.7)$$

In this case, the target of estimation becomes

$$\beta_n := \arg \min_{\theta \in \mathbb{R}^p} \mathbb{E} \left[\{Y - X^\top \theta\}^2 \right] = (\mathbb{E} [X X^\top])^{-1} (\mathbb{E} [X Y]). \quad (3.8)$$

Note that the target β_n does not overtly depend on n because of identical distribution of the random vectors. We still index the target by n to have a consistent notation. Furthermore, in theory that follows the dimension p of β_n may be allowed to depend on n , which introduces an indirect dependence of β_n on n . For this reason all further population quantities will also be indexed by n . – From definitions (3.7) and (3.8), we have

$$\sqrt{n} \widehat{\Sigma}_n (\widehat{\beta}_n - \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n).$$

In this case of iid random vectors it follows that the terms $X_i(Y_i - X_i^\top \beta_n)$ are independent and identically distributed random vectors with mean zero. Therefore, by the multidimensional Berry-Esseen bound (Theorem 1), it follows that

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i [Y_i - X_i^\top \beta_n] \in A \right) - \mathbb{P} (N(0, K_n) \in A) \right| \leq C \frac{p^{1/4}}{n^{1/2}} \alpha_n,$$

where

$$K_n := \mathbb{E} [X X^\top (Y - X^\top \beta_n)^2] \quad \text{and} \quad \alpha_n := \mathbb{E} \left[\|K_n^{-1/2} X_i (Y_i - X_i^\top \beta_n)\|_2^3 \right].$$

Therefore, under certain rate constraints on p , $\sqrt{n} \widehat{\Sigma}_n (\widehat{\beta}_n - \beta_n)$ is approximately normally distributed with mean zero and variance matrix K_n . Since the random vectors are assumed to be iid, under finite fourth moment assumptions on the covariates, it follows that

$$\left\| \widehat{\Sigma}_n - \Sigma_n \right\|_{op} = o_p(1), \quad \text{where} \quad \Sigma_n := \mathbb{E} [X X^\top].$$

See [Vershynin \(2012\)](#) for more details related to the exact rate of this convergence when $p/n = o(1)$. Some related results are presented in Chapter 5. Thus, by Slutsky's

theorem it follows that

$$\sqrt{n}(\hat{\beta}_n - \beta_n) \stackrel{\mathcal{L}}{\approx} N(0, \Sigma_n^{-1} K_n \Sigma_n^{-1}),$$

where we used the notation $\stackrel{\mathcal{L}}{\approx}$ for approximation in law (or distribution) as in the previous section. Again, for inference about β_n using the estimator $\hat{\beta}_n$, one needs to estimate Σ_n and K_n . The matrix Σ_n can be estimated readily by $\hat{\Sigma}_n$, but, to estimate K_n , recall that one needs the variance of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n).$$

Because this is just a scaled average of n independent identically distributed random vectors with mean zero, K_n can be consistently estimated by

$$\hat{K}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \hat{\beta}_n)^2.$$

To show that \hat{K}_n is consistent for K_n , one can use the fact that $\hat{\beta}_n$ is consistent for β_n (see Section 3.5 for more details). Thus, a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta}_n - \beta_n)$ is given by

$$\left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \hat{\beta}_n)^2 \right) \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1}. \quad (3.9)$$

This is often referred to as the sandwich estimator of the asymptotic variance (see Section 3.4 for more details). It is noteworthy that consistent estimation of the “asymptotic” variance of $\hat{\beta}_n$ is possible under iid random vectors and is not possible under fixed covariates without further assumptions.

3.4 Unified Framework for Linear Regression

Before proceeding to unify both the settings of fixed and random covariates, let us recall the main similarities and differences in the analysis presented in the previous sections. First the similarities:

1. In both cases, the least squares estimator $\hat{\beta}_n$ has an “asymptotic” normal distribution with mean β_n , the “moving target” of estimation, and “moving” variance approximation

$$\frac{1}{n}\Sigma_n^{-1}K_n\Sigma_n^{-1}, \quad \text{where} \quad K_n := \frac{1}{n}\sum_{i=1}^n \text{Var}(X_i(Y_i - X_i^\top\beta_n)).$$

Note that the target of estimation β_n is different in the fixed and random covariate cases.

2. The “asymptotic” normality result does not require any more assumptions than independence of observations and certain moment restrictions such as invertibility of the second moment matrix of covariates and finite fourth moments of covariates. In particular, the classical assumptions of linearity and homoscedastic Gaussian errors are not required.

Now the differences:

1. The score vectors $X_i(Y_i - X_i^\top\beta_n)$ are independent in both settings but are mean zero only in the random covariate setting.
2. The “asymptotic” variance can be consistently estimated only in the random covariate setting and is impossible to estimate in the fixed covariate setting without further assumptions.

From this discussion, it is clear that the similarities hold because of the independence assumption and the differences arise from the additional assumption of identical distributions. The differences do not derive from the stochastic properties of the covariates. To provide a unified analysis of linear regression that covers both settings, we propose a framework where the random vectors $(X_i^\top, Y_i)^\top$ are independent but are allowed to be non-identically distributed.

Formally, the observations $(X_i^\top, Y_i)^\top \in \mathbb{R}^{p+1}, 1 \leq i \leq n$ are independent with possibly non-identical distributions. This framework is much more general than either of the two settings – fixed or random covariates. It allows for some random and some fixed covariates as well. The least squares linear regression estimator is still given by

$$\hat{\beta}_n = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \theta)^2. \quad (3.10)$$

The target of estimation in this framework can be defined as

$$\beta_n := \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[(Y_i - X_i^\top \theta)^2 \right]. \quad (3.11)$$

Recall the following notations:

$$\begin{aligned} \hat{\Sigma}_n &:= \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \hat{\Gamma}_n := \frac{1}{n} \sum_{i=1}^n X_i Y_i \in \mathbb{R}^p, \\ \Sigma_n &:= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i X_i^\top] \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \Gamma_n := \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i Y_i] \in \mathbb{R}^p. \end{aligned}$$

Using these matrices and vectors, the estimator and the target defined in (3.10) and

(3.11) can be rewritten as

$$\begin{aligned}\hat{\beta}_n &= \arg \min_{\theta \in \mathbb{R}^p} \theta^\top \hat{\Sigma}_n \theta - \theta^\top \hat{\Gamma}_n, \\ \beta_n &= \arg \min_{\theta \in \mathbb{R}^p} \theta^\top \Sigma_n \theta - \theta^\top \Gamma_n.\end{aligned}\tag{3.12}$$

Since these two objective functions are convex quadratic functions, the minimizers can be obtained as zeros of the derivative, proving that the estimator $\hat{\beta}_n$ satisfies

$$\hat{\Sigma}_n \hat{\beta}_n - \hat{\Gamma}_n = 0,\tag{3.13}$$

and the target β_n satisfies

$$\Sigma_n \beta_n - \Gamma_n = 0.\tag{3.14}$$

Adding and subtracting β_n from $\hat{\beta}_n$ in Equation (3.13) implies

$$\hat{\Sigma}_n (\hat{\beta}_n - \beta_n) = \hat{\Gamma}_n - \hat{\Sigma}_n \beta_n,\tag{3.15}$$

where the right hand side has zero expectation because of (3.14). Expanding the terms shows that

$$\sqrt{n} \hat{\Sigma}_n (\hat{\beta}_n - \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i,$$

where S_i denotes the score given by

$$S_i := X_i (Y_i - X_i^\top \beta_n) - \mathbb{E} [X_i (Y_i - X_i^\top \beta_n)].\tag{3.16}$$

By the multivariate Berry-Esseen bound (Theorem 1), it follows that

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(\sqrt{n} \hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) \in A \right) - \mathbb{P} \left(N(0, K_n) \in A \right) \right| \leq C \frac{p^{1/4}}{n^{1/2}} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| K_n^{-1/2} S_i \right\|_2^3 \right],$$

where

$$K_n = \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} [S_i S_i^\top].$$

This Berry-Esseen bound proves that

$$\sqrt{n} \hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) \stackrel{\mathcal{L}}{\approx} N(0, K_n).$$

And since $\left\| \hat{\Sigma}_n - \Sigma_n \right\|_{op} = o_p(1)$ as $n \rightarrow \infty$, it follows that

$$\sqrt{n} \left(\hat{\beta}_n - \beta_n \right) \stackrel{\mathcal{L}}{\approx} N \left(0, \Sigma_n^{-1} K_n \Sigma_n^{-1} \right).$$

Formally, we have proved the following theorem.

Theorem 2. *If p is fixed (not depending on n), $\mathbb{E} [\|X_i\|_2^6 + Y_i^6] \leq B < \infty$ for all $i \geq 1$, and K_n is invertible, then*

$$\sqrt{n} K_n^{-1/2} \hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) \xrightarrow{\mathcal{L}} N(0, I_p).$$

Here I_p denotes the identity matrix of dimension p .

The moment assumptions in this theorem are generous since the main goal of getting finite sample results for post-selection inference requires much stronger moment assumptions in the chapters to follow.

This completes the “asymptotic” study of linear regression estimator $\hat{\beta}_n$ in the unified framework. We write “asymptotic” because the normal approximations are

actually non-asymptotic.

Remark 3.4.1 (designation of covariates and response) It should be clear from the discussion throughout that singling out a response variable Y_i is arbitrary in principle and context-dependent in practice. It is up to the analyst to decide which variables should be treated as covariates/regressors and which is to be treated as the response. \diamond

3.5 Variance Estimation and Bootstrap in Unified Framework

3.5.1 Sandwich Variance Estimation

The “moving asymptotic” variance of $\sqrt{n}(\hat{\beta}_n - \beta_n)$, as shown in Theorem 2, is given by $\Sigma_n^{-1}K_n\Sigma_n^{-1}$. The Σ_n -part can be readily estimated by $\hat{\Sigma}_n$ and the only part still in need of estimation is K_n . Recall that

$$\begin{aligned} K_n &= \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[X_i X_i^\top (Y_i - X_i^\top \beta_n)^2 \right] \\ &\quad - \frac{1}{n} \sum_{i=1}^n (\mathbb{E} [X_i (Y_i - X_i^\top \beta_n)]) (\mathbb{E} [X_i (Y_i - X_i^\top \beta_n)])^\top. \end{aligned}$$

So, K_n is the variance of a scaled average of non-identically distributed independent random vectors. We prove in Lemma 1 that such a variance cannot be estimated consistently without further assumptions. Accepting this for the moment, note that

$$K_n \leq K_n^*, \quad \text{where} \quad K_n^* := \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[X_i X_i^\top (Y_i - X_i^\top \beta_n)^2 \right],$$

and the matrix K_n^* can be consistently estimated by

$$\check{K}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \left(Y_i - X_i^\top \hat{\beta}_n \right)^2.$$

Hence a conservative estimator of K_n does exist and one such is given by \check{K}_n . (The notation $\check{\cdot}$ is used instead of $\hat{\cdot}$ to emphasize that this is a conservative estimator and not a consistent one.) This provides a conservative estimator of the asymptotic variance as

$$\left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \hat{\beta}_n)^2 \right) \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1}. \quad (3.17)$$

This is the same as the sandwich estimator (3.9) introduced for linear regression with iid random vectors. However, it is important to realize that in the setting of iid random observations this is a consistent estimator, whereas in the unified framework it is only a conservative estimator.

In the following we prove consistency of \check{K}_n for K_n^* . For this, define an intermediate (unattainable) estimator

$$\bar{K}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \beta_n)^2.$$

This is an average of independent random matrices that is unbiased for K_n^* . Hence, under the assumptions of Theorem 2, by the results of Vershynin (2012),

$$\|\bar{K}_n - K_n^*\|_{op} = o_p(1).$$

It now suffices to show that $\check{K}_n - \bar{K}_n$ converges to zero in terms of the operator norm

in probability. Observe that

$$\begin{aligned}\check{K}_n - \bar{K}_n &= \frac{2}{n} \sum_{i=1}^n X_i X_i^\top \left[X_i^\top \hat{\beta}_n - X_i^\top \beta_n \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \left[X_i^\top \hat{\beta}_n - X_i^\top \beta_n \right]^2.\end{aligned}$$

Taking operator norm on both sides, we get

$$\begin{aligned}\|\check{K}_n - \bar{K}_n\|_{op} &\leq \frac{2}{n} \sum_{i=1}^n \|X_i\|_2^2 |Y_i - X_i^\top \beta_n| \left| X_i^\top (\hat{\beta}_n - \beta_n) \right| \\ &\quad + \frac{1}{n} \sum_{i=1}^n \|X_i\|_2^2 \left| X_i^\top (\hat{\beta}_n - \beta_n) \right|^2 \\ &\leq \left(1 + 2 \left(\frac{1}{n} \sum_{i=1}^n \|X_i\|_2^2 |Y_i - X_i^\top \beta_n|^2 \right)^{1/2} \right) \left(\frac{1}{n} \sum_{i=1}^n \|X_i\|_2^2 \left| X_i^\top (\hat{\beta}_n - \beta_n) \right|^2 \right).\end{aligned}$$

Under the assumptions of Theorem 2, the first term above is $O_p(1)$ and the second term is converging to zero. Therefore, $\check{K}_n - K_n^*$ converges in probability to zero in terms of the operator norm.

Remark 3.5.1 (Best Conservative Estimator) We have exhibited one conservative estimator for the “moving asymptotic” variance of $\hat{\beta}_n$, but many other conservative estimators exist, an example being the (delete-one) jackknife; see Long and Ervin (2000) for more details. It would be interesting to study the question of what comes closest to the true “asymptotic” variance, but we do not know of an answer at present. An interesting feature of the conservative estimator (3.17) is that it is consistent in the case of iid observations, but the jackknife estimator is known to be (asymptotically) conservative. \diamond

The following lemma proves that there does not exist a consistent estimator for the variance of an average of non-identically distributed independent random vectors.

The lemma is stated for real-valued random variables which implies the result for random vectors by taking projections. Thanks are due to Shaokun Li for the proof of this result. See Proposition 3.5 of [Bachoc et al. \(2016\)](#) for a related result.

Lemma 1. *Suppose W_1, \dots, W_n are independent random variables with $\mathbb{E}[W_i] = \mu_i$ and $\text{Var}(W_i) = \sigma_i^2$. Then there does not exist a consistent estimator for η_n^2 , where*

$$\eta_n^2 := \frac{1}{n} \sum_{i=1}^n \sigma_i^2.$$

Proof. We need to prove that there does not exist a sequence of measurable functions $\{f_n(W_1, W_2, \dots, W_n)\}$ such that as $n \rightarrow \infty$,

$$f_n(W_1, W_2, \dots, W_n) - \eta_n^2 \xrightarrow{P} 0,$$

for arbitrary $\{(\mu_i, \sigma_i^2) : 1 \leq i \leq n\}$. Assuming that such a sequence exists, we obtain from consistency in the special case $\sigma_i^2 = 0$ for $i \geq 1$ that

$$f_n(\mu_1, \mu_2, \dots, \mu_n) \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty, \quad (3.18)$$

for any fixed sequence $(\mu_i)_{i \geq 1}$. Now, fix $\varepsilon > 0$ and define the sequence of (measurable) sets

$$A_n = \{|f_n(W_1, W_2, \dots, W_n)| \leq \varepsilon\}.$$

Using (3.18), we have that for any sequence $(w_i)_{i \geq 1}$ as $n \rightarrow \infty$

$$\mathbb{P}(A_n | W_1 = w_1, \dots, W_n = w_n) = \mathbb{1}\{|f_n(w_1, w_2, \dots, w_n)| \leq \varepsilon\} \rightarrow 1.$$

Thus, $\mathbb{P}(A_n) \rightarrow 1$ as $n \rightarrow \infty$. This implies that as $n \rightarrow \infty$,

$$f_n(W_1, W_2, \dots, W_n) \xrightarrow{P} 0,$$

irrespective of what the true η_n^2 is. This contradicts the existence of a sequence consistent for η_n^2 . \square

Remark 3.5.2 The proof also implies that there is no other option than to overestimate the variance, if at all possible. \diamond

3.5.2 From Sandwich to Bootstrap Estimators

The sandwich estimator presented in (3.17) is a direct or closed-form estimator of standard error (squared). It would be of interest to understand how various versions of bootstrap work for the purpose of variance estimation or distributional approximation. In what follows we consider two different bootstrap approaches in the unified framework. These are different from the residual bootstrap and the nonparametric pairs bootstrap considered in the literature on linear regression. See [Freedman \(1981\)](#) and [Buja et al. \(2014\)](#) for more details. There are two reasons for this different approach we take. Firstly, the residual bootstrap isn't applicable because it assumes linearity and iid errors. Secondly, the pairs or x - y bootstrap can lead to singular linear systems in simulations. The bootstrap approaches provided here are applicable in the unified framework and bypass the problem of singular linear systems. We call this bootstrap methodology the “score bootstrap” since it is based on resampling scores. This idea was introduced and studied under classical model assumptions in [Hu and Kalbfleisch \(2000\)](#).

3.5.3 Multiplier Score Bootstrap

Let W_1, W_2, \dots, W_n be independent random variables that are in turn independent of (X_i, Y_i) and satisfy

$$\mathbb{E}[W_i] = 0, \quad \mathbb{E}[W_i^2] = 1, \quad \text{and} \quad \mathbb{E}[|W_i|^3] < \infty.$$

These variables need not be identically distributed but there is no special reason for them to be non-identically distributed except for allowing generality. Recall that

$$\sqrt{n} \hat{\Sigma}_n (\hat{\beta}_n - \beta_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n).$$

Define the estimated score vectors

$$\hat{S}_i = X_i (Y_i - X_i^\top \hat{\beta}_n),$$

and observe that $\frac{1}{n} \sum_{i=1}^n \hat{S}_i = 0$, which is just the normal equations satisfied by $\hat{\beta}_n$.

Set

$$T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \quad \text{and} \quad T_n^* := \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \hat{S}_i, \quad (3.19)$$

where S_i are the true scores defined in (3.16). Conditional on $\mathcal{Z}_n := \{(X_i, Y_i), 1 \leq i \leq n\}$, T_n^* is approximately normally distributed with mean zero and variance \check{K}_n and more precisely,

$$\sup_{A \in \mathcal{C}_p} |\mathbb{P}(T_n^* \in A | \mathcal{Z}_n) - \mathbb{P}(N(0, \check{K}_n) \in A | \mathcal{Z}_n)| \leq C \frac{p^{1/4}}{n^{1/2}} \frac{1}{n} \sum_{i=1}^n \left\| \check{K}_n^{-1/2} \hat{S}_i \right\|_2^3 \mathbb{E}[|W_i|^3], \quad (3.20)$$

As shown before $\|\check{K}_n - K_n^*\|_{op} = o_p(1)$, and so, as $n \rightarrow \infty$,

$$\begin{aligned} \sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(N(0, \check{K}_n) \in A \mid \mathcal{Z}_n \right) - \mathbb{P} \left(N(0, K_n^*) \in A \right) \right| &\leq p^{1/2} \left\| (K_n^*)^{-1} \check{K}_n - I_p \right\|_{op}^{1/2} \\ &= o_p(1). \end{aligned} \quad (3.21)$$

See Chapter 2, Example 2.3 of [DasGupta \(2008\)](#) for the inequality above. Recall the Berry-Esseen bound for linear regression as

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P} \left(\sqrt{n} \hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) \in A \right) - \mathbb{P} \left(N(0, K_n) \in A \right) \right| \leq C \frac{p^{1/4}}{n^{1/2}} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left\| K_n^{-1/2} S_i \right\|_2^3 \right]. \quad (3.22)$$

To show that the multiplier score bootstrap works, we need Anderson's Lemma.

Lemma 2 (Corollary 3, [Anderson \(1955\)](#)). *If $\xi \sim N(0, \Sigma)$ and A is any centrally symmetric convex set (that is, $x \in A$ implies $-x \in A$ and A convex), then for any y ,*

$$\mathbb{P}(\xi + y \in A) \leq \mathbb{P}(\xi \in A).$$

By Anderson's Lemma, for any centrally convex set A ,

$$\mathbb{P}(N(0, K_n^*) \in A) \leq \mathbb{P}(N(0, K_n) \in A),$$

and using bounds [\(3.20\)](#), [\(3.21\)](#) and [\(3.22\)](#), we get

$$\begin{aligned} \mathbb{P}(T_n \in A) &= \mathbb{P} \left(\sqrt{n} \hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) \in A \right) = \mathbb{P}(N(0, K_n) \in A) + o(1) \\ &\geq \mathbb{P}(N(0, K_n^*) \in A) + o(1) \\ &= \mathbb{P}(T_n^* \in A \mid \mathcal{Z}_n) + o_p(1), \end{aligned} \quad (3.23)$$

for all centrally symmetric convex sets in \mathbb{R}^p . Recall the definitions of T_n and T_n^*

from (3.19). For further use rewrite inequalities (3.23) as

$$\inf_{A \in \bar{\mathcal{C}}_p} \left(\int_A dP_{T_n} - \int_A dP_{T_n^* | \mathcal{Z}_n} \right) \geq o_p(1). \quad (3.24)$$

Here P_{T_n} and $P_{T_n^* | \mathcal{Z}_n}$ represent the probability measure of T_n and that of T_n^* conditional on \mathcal{Z}_n , respectively. The $o_p(1)$ on the right hand side is with respect to the distribution of \mathcal{Z}_n .

These inequalities can be used for an asymptotic justification of the simulation-based multiplier bootstrap: Suppose we generate B_n draws $(W_1^{*b}, \dots, W_n^{*b})$ ($b = 1, \dots, B_n$), calculate the associated bootstrap statistics T_n^{*b} , and construct the bootstrap empirical measure defined by

$$\hat{\mu}_n(A) = \frac{1}{B_n} \sum_{b=1}^{B_n} \mathbb{1}\{T_n^{*b} \in A\}, \quad \text{for any Borel set } A \subseteq \mathbb{R}^p.$$

The measure $\hat{\mu}_n(\cdot)$ is random due to randomness in \mathcal{Z}_n and in $(W_1^{*b}, \dots, W_n^{*b})$. Note that T_n^{*b} are iid random vectors conditional on \mathcal{Z}_n . For any Borel set A we have

$$\mathbb{E}[\hat{\mu}_n(A) | \mathcal{Z}_n] = \mathbb{P}(T_n^* \in A | \mathcal{Z}_n).$$

Hence for various classes of sets $\mathcal{C}^* \subseteq \mathcal{C}_p$, conditional on \mathcal{Z}_n , as $B_n \rightarrow \infty$, we have

$$\sup_{A \in \mathcal{C}^*} \left| \hat{\mu}_n(A) - \int_A dP_{T_n^* | \mathcal{Z}_n} \right| = o_p(1), \quad (3.25)$$

where $o_p(1)$ on the right hand side is with respect to the distribution of bootstrap samples. The class \mathcal{C}^* of sets that satisfy (3.25) are called Glivenko-Cantelli (GC) classes. The classes of all rectangles and ellipsoids have been shown to be GC classes. See [Elker et al. \(1979\)](#), [Devroye \(1982, Page 75\)](#) and [Pollard \(1984, Chapter II\)](#) for

more precise results.

Combining results (3.24) and (3.25), we obtain

$$\inf_{A \in \mathcal{C}^*} \left(\int_A dP_{T_n} - \frac{1}{B_n} \sum_{b=1}^{B_n} \mathbb{1}\{T_n^* \in A\} \right) \geq o_p(1), \quad (3.26)$$

where $o_p(1)$ refers to both the randomness of the data \mathcal{Z}_n and the randomness of the bootstrap samples. Suppose now we construct a set $\hat{\mathcal{R}}_n(\alpha) \in \mathcal{C}^*$ for some $\alpha \in [0, 1]$ such that

$$\frac{1}{B_n} \sum_{b=1}^{B_n} \mathbb{1}\{T_n^* \in \hat{\mathcal{R}}_n(\alpha)\} = 1 - \alpha.$$

Then from inequality (3.26), it follows that as $n \rightarrow \infty$,

$$\inf_{\alpha \in [0, 1]} \left(\int_{\hat{\mathcal{R}}_n(\alpha)} dP_{T_n} - (1 - \alpha) \right) \geq o_p(1),$$

where the $o_p(1)$ is exactly the one from (3.26). Since $\hat{\mathcal{R}}_n(\alpha)$ is random, the integral on the left hand side is a random quantity. Recall that $T_n = \sqrt{n} \hat{\Sigma}_n(\hat{\beta}_n - \beta_n)$ and so, the above inequality implies that the confidence region $\hat{\mathcal{R}}_n(\alpha)$ provides an asymptotically conservative confidence region for β_n . Note here that α can be chosen based on the data and validity still holds.

It is clear from this analysis that the multiplier score bootstrap ends up providing inference based on the same conservative variance estimator as the direct sandwich estimator constructed before. We observe that the main decision was to apply the bootstrap at the level of scores as opposed to the original data (and OLS applied to them). The resampling bootstrap at the level of scores would allow a similar analysis as given above for the multiplier bootstrap, and this will be outlined in the following subsection.

3.5.4 Resampling Score Bootstrap

We consider briefly the m -of- n resampling bootstrap applied to the score vectors. The associated resampling bootstrap statistic is

$$T_m^* := \frac{1}{\sqrt{m}} \sum_{j=1}^m \hat{S}_{I_j},$$

where $I_j, 1 \leq j \leq m$ represents a sample of m iid uniform random variables drawn from $\{1, 2, \dots, n\}$ (i.e., sampling with replacement). Applying the multidimensional Berry-Esseen bound conditional on the data \mathcal{Z}_n , we obtain

$$\sup_{A \in \mathcal{C}_p} \left| \mathbb{P}(T_m^* \in A | \mathcal{Z}_n) - \mathbb{P}(N(0, \check{K}_n) \in A) \right| \leq C \frac{p^{1/4}}{m^{1/2}} \frac{1}{n} \sum_{i=1}^n \left\| \check{K}_n^{-1/2} \hat{S}_i \right\|_2^3. \quad (3.27)$$

Now, retracing the steps of the previous subsection, we conclude that the resampling score bootstrap also produces asymptotically conservative inference based on the same conservative variance estimator as the sandwich.

Note that for fixed p one requires a large resampling size m for the normal approximation to be good. If m does not grow as fast as n , then the bound in (3.27) dominates the error in the coverage of the bootstrap confidence region.

3.6 Hypothesis Testing in the Unified Framework

In the previous sections, we considered inference based on confidence regions. In this section we consider inference based on hypothesis testing. Testing will play an important role in solving the post-selection inference problem as indicated in Section 2.2 of Chapter 2. Consider now the test of the hypothesis

$$H_0 : \beta_n(j) = \beta_{n,0} \quad \text{versus} \quad H_1 : \beta_n(j) \neq \beta_{n,0},$$

for a fixed $j \in \{1, 2, \dots, p\}$ and some fixed $\beta_{n,0} \in \mathbb{R}$. If $\beta_{n,0} = 0$, then this is the problem of establishing statistical significance of the (linear) effect as measured by the coefficient $\beta_n(j)$ of the j -th covariate on the response Y . The only estimator for β_n we considered was $\hat{\beta}_n$, and so a reasonable test can be based on $\hat{\beta}_n(j)$. Recall that

$$\hat{\Sigma}_n \left(\hat{\beta}_n - \beta_n \right) = \frac{1}{n} \sum_{i=1}^n X_i (Y_i - X_i^\top \beta_n),$$

where the right hand side has mean zero with ingredient vectors possibly of non-zero mean. Since $\hat{\Sigma}_n$ is a consistent estimator for Σ_n in the sense that $\left\| \hat{\Sigma}_n - \Sigma_n \right\| = o_p(1)$ as $n \rightarrow \infty$,

$$\sqrt{n} \left(\hat{\beta}_n - \beta_n \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\Sigma_n^{-1} X_i] (Y_i - X_i^\top \beta_n) + o_p(1). \quad (3.28)$$

The right hand side, by the multivariate Berry-Esseen bound, has an approximate normal distribution with mean zero and variance matrix

$$AV_n := \Sigma_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n \text{Var} (X_i (Y_i - X_i^\top \beta_n)) \right) \Sigma_n^{-1}.$$

This asymptotic normal approximation implies that for any fixed $1 \leq j \leq p$,

$$\frac{\sqrt{n} \left(\hat{\beta}_n(j) - \beta_n(j) \right)}{\sqrt{AV_n(j, j)}} \xrightarrow{\mathcal{L}} N(0, 1).$$

Here the notation $\xrightarrow{\mathcal{L}}$ is used to denote convergence in law (or distribution). As proved in previous sections, there does not exist a consistent estimator for AV_n (in this general framework) but there exists a (asymptotically) conservative estimator given by

$$\tilde{AV}_n := \hat{\Sigma}_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \hat{\beta}_n)^2 \right) \hat{\Sigma}_n^{-1}.$$

This is consistent for

$$AV_n^* := \Sigma_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i X_i^\top (Y_i - X_i^\top \beta_n)^2] \right) \Sigma_n^{-1}.$$

Thus, by Slutsky's theorem,

$$\frac{\sqrt{n}(\hat{\beta}_n(j) - \beta_n(j))}{\sqrt{\widetilde{AV}_n(j, j)}} \underset{\mathcal{L}}{\approx} N \left(0, \frac{AV_n(j, j)}{AV_n^*(j, j)} \right).$$

Here the variance of the normal distribution on the right is at most 1. Since this ratio cannot be estimated consistently, one solution is to conservatively use $N(0, 1)$ instead. To perform the test replace $\beta_n(j)$ by $\beta_{n,0}$ and use this normal distribution. So, the test is based on the statistic

$$t_j := \frac{\sqrt{n}(\hat{\beta}_n(j) - \beta_{n,0})}{\sqrt{\widetilde{AV}_n(j, j)}}.$$

In the classical linear regression model, the denominator for the same hypothesis testing problem is given by the classical estimator of the variance obtained under the assumption of correct specification. The test statistic t_j has then a t -distribution. That denominator is not valid in the unified framework which permits misspecification. The present statistic t_j hence cannot be assumed to have a t -distribution. Note that the test based on t_j leads to a conservative test, meaning the type-I error, in this general framework, would be strictly smaller than α (asymptotically). One subtle point here is that this conservativeness does not arise from AV_n^* but from the use of $N(0, 1)$ instead of the correct but unattainable normal distribution. Because t -distributions have heavier tails than $N(0, 1)$, their use would result in additional conservativeness. Such could be considered desirable by those who wish to account

for estimated degrees of freedom.

Suppose now we want to simultaneously test over all $1 \leq j \leq p$ instead of just one of them, that is,

$$H_0 : \beta_n = \beta_{n,0} \quad \text{versus} \quad H_1 : \beta_n \neq \beta_{n,0}$$

for some vector $\beta_{n,0} \in \mathbb{R}^p$. This testing problem is usually addressed by an F -test, but an intuitive alternative can be based on the “max- $|t|$ ” statistic defined by

$$T_{n,p} := \max_{1 \leq j \leq p} |t_j| = \max_{1 \leq j \leq p} \left| \frac{\sqrt{n}(\hat{\beta}_n(j) - \beta_0(j))}{\sqrt{\widetilde{AV}_n(j,j)}} \right|.$$

The name “max- $|t|$ ” derives from classical linear regression theory, but in the current context of a unified framework this is strictly speaking a misnomer.

We end this section with one last point: Even though all the above tests are asymptotically conservative, they may not be conservative for inference in finite samples because of asymptotic approximation error.

3.7 Conclusions on Assumptions for Linear Regression

What we find from the (essentially finite-sample) analysis in previous sections is that we do not need any of the usual model assumptions including linearity, normality and homoscedasticity. Only under independence assumptions on observations (along with some moment assumptions), we have asymptotic normality of the LSE around its corresponding target (properly scaled), and

$$\left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top (Y_i - X_i^\top \hat{\beta}_n)^2 \right) \left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right)^{-1},$$

is an asymptotically valid estimator of the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$. This should be understood in the sense that when observations are identically distributed this estimator is consistent, and when observations are non-identically distributed this estimator is asymptotically conservative (no consistent estimator exists in this case). The conservativeness in the broader context of generalized linear models was discussed in [Fahrmeir \(1990\)](#). See page 492 in [Fahrmeir \(1990\)](#).

In passing let us now make a comment on the assumption of independence of observations. When discussing and defining the target of estimation, it was shown that even the independence of observations is not needed. To make the rates and the asymptotic distribution concrete, the assumption of independence was introduced. Recollecting the technical tools that went into the derivation of [Theorem 2](#), it can be seen that the linear representation [\(3.15\)](#) (that holds without any assumptions on the random vectors) and the multivariate Berry-Esseen bound ([Theorem 1](#)) for mean zero independent random vectors are used. So, as long as a version of a Berry-Esseen bound or a multivariate central limit theorem exists, the assumption of independence can be replaced by a “weak” dependence assumption. See [Hörmann \(2009\)](#) for Berry-Esseen bounds for averages of mean zero random vectors under various dependence settings based on an approximation with m -dependent sequences. Also, see [Chapter 10 of Pötscher and Prucha \(1997\)](#). In the chapters to follow, for easy understanding, we focus on independent observations first and then discuss extensions to dependent structures.

APPENDIX

3.A Semiparametric Efficiency

The discussion in this chapter was restricted to the discussion of the “asymptotic” properties of the estimator that the data analyst started with and was not related to how well one can estimate the target of estimation. As mentioned before, the traditional mathematical statistics was designed under correctly specified parametric models and the goal is to efficiently estimate the true parameter that determines the distribution. From the point of view of previous sections and the current thesis, this question as it is does not make sense; the analyst wants to use the (least squares linear regression) estimator he/she chose irrespective of what the true model is. Alternatively, one might ask “having chosen an estimator that leads a particular target of estimation, is there an efficient way to estimate the target of estimation?”. For example, in case the analyst has chosen to use least squares linear regression estimator, the target of estimation becomes

$$\beta_n = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(Y_i - X_i^\top \theta)^2].$$

What is an efficient estimator of β_n ? is $\hat{\beta}_n$ an efficient estimator for β_n ? what does efficiency mean here? This question naturally leads to the area of semiparametric inference and the answer exists at least in the case of iid random vectors since [Levit \(1976\)](#). See example 5 on page 725 of [Levit \(1976\)](#). In this appendix, we provide a heuristic argument for how should an efficient estimator look like for the case of independent observations (without identical distributions assumption). See [Bolthausen et al. \(2002, Lectures 1-4, pages 336–382\)](#) and [McNeney \(1998\)](#) for ways to formalizing the result. The setting for semiparametric inference is as follows: suppose

Z_1, Z_2, \dots, Z_n are n independent random vectors with $Z_i \sim P_i$ for some probability distribution P_i and the target of estimation is $\psi(P^{\otimes n})$ for some functional ψ defined on a class of distributions \mathcal{P}_n (chosen also by the analyst). Here

$$P^{\otimes n} = \bigotimes_{i=1}^n P_i,$$

represents the joint distribution of (Z_1, Z_2, \dots, Z_n) and \mathcal{P}_n contains distributions of this type where each P_i varies over some set of probability distributions. Some example might clarify the problem:

1. Suppose Z_1, \dots, Z_n are n independent real-valued random variables and we want to estimate

$$\psi(P^{\otimes n}) := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i] = \int \left(\frac{1}{n} \sum_{i=1}^n z_i \right) dP_1(z_1) \dots dP_n(z_n).$$

Here \mathcal{P}_n can be taken to be the set of all joint distributions of Z_1, \dots, Z_n such that the marginal variances are all uniformly bounded. One can consider the same functional with random vectors too.

2. Suppose $Z_i \in \mathbb{R}^q, 1 \leq i \leq n$ are independent random vectors and $\rho : \mathbb{R}^q \times \mathbb{R}^k$ is some “loss” function. The functional to be estimated is

$$\psi(P^{\otimes n}) := \arg \min_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\rho(Z_i, \theta)].$$

Here too the class of joint distributions \mathcal{P}_n can be taken to be completely non-parametric as in the previous example except for some more moment restrictions to let the functional well-defined. Note that unlike the previous example, it may not be possible to explicitly write the functional in terms of $P^{\otimes n}$.

These are called semiparametric problems since the class of all distributions is mostly nonparametric (unrestricted) and the functional of interest is Euclidean (or parametric) in nature.

The basic idea of semiparametric efficiency is as stated by [Newey \(1990, Section 2\)](#) and [Bolthausen et al. \(2002, Section 1.2\)](#):

The semiparametric problem is at least as hard as any of the parametric problems that it encompasses.

To understand this idea, briefly consider the simpler case of identical distributions so that \mathcal{P}_n is a subset of the class of all joint distributions with marginal distributions restricted to be the same. Let the true distribution of observations be

$$\bigotimes_{i=1}^n P.$$

As a thought experiment, think of \mathcal{P}_n as constituted by joint distributions of the form

$$P_t^{g, \otimes n} := \bigotimes_{i=1}^n P_t^{(g)}, \tag{3.29}$$

for $t \in \mathbb{R}$ and g varying over some class of functions, \mathcal{G} with $P_{t=0}^{(g)} = P$ for any $g \in \mathcal{G}$. So, the nature can be thought of as picking a function $g \in \mathcal{G}$ and then producing observations from $P_t^{g, \otimes n}$. If the function g is known to the statistician, he/she could perform maximum likelihood estimation on the parametric (sub-)model:

$$\mathcal{P}_n^{(g)} := \left\{ \bigotimes_{i=1}^n P_t^{(g)} : t \in \mathbb{R} \right\},$$

to obtain \hat{t} , an estimator of t and then estimate the functional ψ by

$$\hat{\psi}_n^{(g)} := \psi \left(P_{\hat{t}}^{g, \otimes n} \right). \quad (3.30)$$

Under certain regularity conditions, this estimator would achieve the “smallest” variance asymptotically, if g were known to the statistician. However, g and \mathcal{G} are both unknown. Hence, the statistician cannot perform better than the largest variance of $\hat{\psi}_n^{(g)}$ over $g \in \mathcal{G}$. The parametric sub-model that leads to this largest variance is called the least favorable sub-model. To use this idea, one would usually take parametric sub-models of the form (3.29) that are contained in \mathcal{P}_n and take the largest efficient variance over $g \in \mathcal{G}$ as the best possible variance in the semiparametric setting.

To see this idea in action, note first that the variance of $\hat{\psi}_n^{(g)}$ (in (3.30)) asymptotically should be given by the Cramer-Rao lower bound, under regularity conditions. We recall the Cramer-Rao lower bound here with proof for completeness.

Lemma 3 (Cramer-Rao Lower Bound). *If $h(X)$ is an unbiased estimator of $\psi(P) \in \mathbb{R}$ for $P \in \{P_\theta : \theta \in \Theta\}$ absolutely continuous with respect to the Lebesgue measure for some open subset $\Theta \subseteq \mathbb{R}^k$, then under conditions allowing interchange of derivative and integral*

$$\text{Var}_{P_{\theta_0}}(h(X)) \geq \frac{[\psi'(P_{\theta_0})]^2}{\mathbb{E}[\dot{\ell}_{\theta_0}^2(X)]}.$$

Here

$$\psi'(P_{\theta_0}) = \frac{d\psi(P_\theta)}{d\theta} \Big|_{\theta=\theta_0} \quad \text{and} \quad \dot{\ell}_{\theta_0}(x) = \frac{d}{d\theta} \log dP_\theta(x) \Big|_{\theta=\theta_0}.$$

The function $\dot{\ell}_{\theta_0}(x)$ is called the likelihood score.

Proof. From the hypothesis of unbiasedness, we obtain

$$\int h(x) dP_\theta(x) = \psi(P_\theta) \quad \text{for all } \theta \in \Theta.$$

Now differentiating with respect to θ , it follows that

$$\int h(x) \dot{\ell}_{\theta_0}(x) dP_{\theta_0}(x) = \psi'(P_{\theta_0}).$$

Equivalently, this can be written as

$$\text{Cov}_{P_{\theta_0}} \left(h(X), \dot{\ell}_{\theta_0}(X) \right) = \psi'(P_{\theta_0}).$$

Using Cauchy-Schwarz inequality, we get

$$\text{Var}_{P_{\theta_0}} (h(X)) \geq \frac{[\psi'(P_{\theta_0})]^2}{\mathbb{E} \left[\dot{\ell}_{\theta_0}^2(X) \right]},$$

proving the result. □

Now getting back to the semiparametric problem with independent but possibly non-identically distributed observations, consider the parametric sub-model,

$$dP_t^{(g_1, \dots, g_n)}(z_1, \dots, z_n) := \prod_{i=1}^n c(t, g_i) K(tg_i(z_i)) dP_i(z_i), \quad t \in \mathbb{R},$$

where $g_i(\cdot)$, $1 \leq i \leq n$ represent any set of n functions satisfying $\int g_i(z) dP_i(z) = 0$ and $\int g_i^2(z) dP_i(z) < \infty$. Here the function $K(\cdot)$ is given by

$$K(u) = 2(1 + \exp(-2u))^{-1},$$

and $c(t, g_i)$ is a positive normalizing constant. Since

$$c(t, g_i) \int K(tg_i(z)) dP_i(z) = 1 \quad \text{for all } t \in \mathbb{R},$$

it follows that $c(0, g_i) = 1$. See example 1.12 on page 346 of [Bolthausen et al. \(2002\)](#). Differentiating with respect to t and taking $t = 0$ proves

$$c'(0, g_i) = \left. \frac{d}{dt} c(t, g_i) \right|_{t=0} = 0.$$

Therefore, the “likelihood score” term is given by

$$\left. \frac{d}{dt} \log \frac{dP_t^{(g_1, \dots, g_n)}(z_1, \dots, z_n)}{dP^{\otimes n}(z_1, \dots, z_n)} \right|_{t=0} = \sum_{i=1}^n g_i(z_i),$$

and

$$\sum_{i=1}^n \int g_i(z_i) dP_i(z_i) = 0.$$

By independence of observations,

$$\mathbb{E} \left[\left(\sum_{i=1}^n g_i(Z_i) \right)^2 \right] = \sum_{i=1}^n \mathbb{E} [g_i^2(Z_i)]. \quad (3.31)$$

To find the semiparametric lower bound, all we need to find is the “derivative” of the functional. For our purposes, all the functionals we work with are of the form given in example 2 above, that is

$$\psi(P^{\otimes n}) := \arg \min_{\theta \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\rho(Z_i, \theta)].$$

We deal with the case $k = 1$ and the general case follows by taking linear combinations of the functional. Assume that $\rho(\cdot, \cdot)$ is twice differentiable with respect to the second argument and let

$$\Psi(z, \theta) := \frac{d}{d\theta} \rho(z, \theta) \quad \text{and} \quad \dot{\Psi}(z, \theta) := \frac{d}{d\theta} \Psi(z, \theta).$$

Using this differentiability, it follows that for all $P^{\otimes n}$,

$$\sum_{i=1}^n \int \Psi(z_i, \psi(P^{\otimes n})) dP_i(z_i) = 0.$$

Taking $P^{\otimes n}$ to be $P_t^{(g_1, \dots, g_n)}$, we get for all $t \in \mathbb{R}$,

$$\sum_{i=1}^n c(t, g_i) \Psi(z_i, \psi(P_t^{(g_1, \dots, g_n)})) K(tg_i(z_i)) dP_i(z_i) = 0.$$

Differentiating with respect to t and taking $t = 0$, it follows that

$$\begin{aligned} \frac{d}{dt} \psi(P_t^{(g_1, \dots, g_n)}) &= \left(\sum_{i=1}^n \mathbb{E} \left[\dot{\Psi}(Z_i, \psi(P^{\otimes n})) \right] \right)^{-1} \mathbb{E} \left[\sum_{i=1}^n \Psi(Z_i, \psi(P^{\otimes n})) g_i(Z_i) \right] \\ &= \mathbb{E} \left[\sum_{i=1}^n \tilde{\psi}_{P_i}(Z_i) g_i(Z_i) \right]. \end{aligned} \quad (3.32)$$

Here

$$\tilde{\psi}_{P_i}(z_i) := \left(\sum_{i=1}^n \mathbb{E} \left[\dot{\Psi}(Z_i, \psi(P^{\otimes n})) \right] \right)^{-1} \{ \Psi(z_i, \psi(P^{\otimes n})) - \mathbb{E}[\Psi(Z_i, \psi(P^{\otimes n}))] \},$$

and the properties $c(0, g_i) = 1, c'(0, g_i) = 0, K(0) = 1, K'(0) = 1$ are used. The function $\tilde{\psi}_{P_i}(\cdot)$ is called the “efficient influence function” for the iid case. Substituting (3.31) and (3.32) in the Cramer-Rao lower bound (Lemma 3) and maximizing with respect to all $g_i, 1 \leq i \leq n$ with finite second moment implies

$$\sup_{\substack{(g_i)_{1 \leq i \leq n}: \\ \mathbb{E}(g_i(Z_i))=0, \mathbb{E}(g_i^2(Z_i))<\infty}} \left(\sum_{i=1}^n \mathbb{E} [g_i^2(Z_i)] \right)^{-1} \left(\sum_{i=1}^n \mathbb{E} \left[\tilde{\psi}_{P_i}(Z_i) g_i(Z_i) \right] \right)^2 = \mathbb{E} \left[\sum_{i=1}^n \tilde{\psi}_{P_i}^2(Z_i) \right].$$

This maximum is attained for $g_i(z_i) = \tilde{\psi}_{P_i}(z_i)$. By a semiparametric extension of regular estimator, this implies that any regular efficient estimator T_n must have an

asymptotic linear representation given by

$$\sqrt{n} (T_n - \psi(P^{\otimes n})) = \sqrt{n} \sum_{i=1}^n \tilde{\psi}_{P_i}(Z_i) + o_p(1).$$

Note that $\mathbb{E}[\tilde{\psi}_{P_i}(Z_i)] = 0$ and $\text{Var}(\tilde{\psi}_{P_i}(Z_i)) < \infty$. By an application of Lindeberg-Feller theorem, it follows that T_n under suitable normalization has an asymptotic normal distribution under the Lindeberg condition.

3.A.1 Application to Linear Regression

For the case of linear regression, $Z_i = (X_i, Y_i)$ and $\rho(z, \theta) = (y - x^\top \theta)^2$. Therefore,

$$\Psi(z, \theta) = -x(y - x^\top \theta) \quad \text{and} \quad \dot{\Psi}(z, \theta) = -xx^\top.$$

Hence, any efficient regular estimator T_n of the target of estimator β_n must have an asymptotic linear representation given by

$$\begin{aligned} \sqrt{n} (T_n - \beta_n) &= \sqrt{n} \sum_{i=1}^n \left(\sum_{i=1}^n \mathbb{E}[X_i X_i^\top] \right)^{-1} \{X_i(Y_i - X_i^\top \beta_n) - \mathbb{E}[X_i(Y_i - X_i^\top \beta_n)]\} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma_n^{-1} \{X_i(Y_i - X_i^\top \beta_n) - \mathbb{E}[X_i(Y_i - X_i^\top \beta_n)]\} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Sigma_n^{-1} X_i(Y_i - X_i^\top \beta_n) + o_p(1). \end{aligned}$$

Realize from Equation (3.28) that the least squares linear regression estimator $\hat{\beta}_n$ satisfies the linear representation and so the least squares estimator is a semiparametrically efficient estimator of β_n . Similar calculations holds for M -estimators obtained from generalized linear models.

End of Chapter 3.

Valid Post-selection Inference in Assumption-lean Linear Regression

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, Edward I.
George, and Linda H. Zhao

University of Pennsylvania
e-mail: arunku@wharton.upenn.edu

Abstract: Construction of valid statistical inference for estimators based on data-driven selection has received a lot of attention in the recent times. Berk et al. (2013) is possibly the first work to provide valid inference for Gaussian homoscedastic linear regression with fixed covariates under arbitrary covariate/variable selection. The setting is unrealistic and is extended by Bachoc et al. (2016) by relaxing the distributional assumptions. A major drawback of the aforementioned works is that the construction of valid confidence regions is computationally intensive. In this paper, we first prove that post-selection inference is equivalent to simultaneous inference and then construct valid post-selection confidence regions which are computationally simple. Our construction is based on deterministic inequalities and apply to independent as well as dependent random variables without the requirement of correct distributional assumptions. Finally, we compare the volume of our confidence regions with the existing ones and show that under non-stochastic covariates, our regions are much smaller.

1. Introduction and Motivation

1.1. Motivation of the Problem

In recent times, there has been a crisis in the sciences because too many research results are found to lack replicability and reproducibility. Some of this crisis has been attributed to a failure of statistical methods to account for data-dependent exploration and modeling that precedes statistical inference. Data-dependent actions such as selection of subsets of cases, of covariates, of responses, of transformations and of model types has been aptly named “researcher degrees of freedom” (Simmons et al., 2011), and these may well be a significant contributing factor in the current crisis. Classical statistics does not account for them because it is built on a framework where all modeling decisions are to be made *independently of the data on which inference is to be based*. But if the data are in fact used to this end prior to statistical inference, then such inference loses its justifications and the ensuing validity conferred on it by classical theories. It is therefore critical that the theory of statistical inference be brought up to date to account for

Figure 3.1: The following chapter is based on [Kuchibhotla et al. \(2020\)](#)

Chapter 4

Post-selection Inference in Linear Regression

In this chapter, we solve (in multiple ways) the problem of post-selection inference for linear regression¹. All these solutions are provided under the assumption-lean framework introduced in Chapter 3. As shown in Chapter 2, there has been a rich literature about post-selection inference arising from various types of selection. This chapter only focuses on a particular type of selection, namely, variable selection in linear regression. Relevant literature on this type of selection is reviewed at appropriate parts of this chapter. The structure of the objective function in linear regression allows for a special construction of confidence regions for valid post-selection inference. At present, it is not clear if this construction lends itself for general M -estimation settings. The main focus of these confidence regions is not optimality in terms of getting the smallest (or near smallest) volume confidence regions but validity and quick computation.

The setting or framework of Chapter 3 allows for p (the total number of covariates)

¹This Chapter is based on [Kuchibhotla et al. \(2020\)](#)

changing with n as can be seen from the multivariate Berry-Esseen bounds. However, p has to diverging (if at all) at a rate much slower than n for the asymptotics there to work. As is the current focus of high-dimensional statistics literature, variable selection is mostly related to the cases where p is larger than n and diverging much faster than n . In this chapter, the results are presented with this ideology in the background.

The remaining chapter is organized as follows. Section 4.1 provides the required notation for formulating the valid post model-selection inference problem on a rigorous footing. In Section 4.2, the problem of post-selection inference is shown to be equivalent to the problem of simultaneous inference. The first approach (or strategy) for valid post-selection inference is presented in Section 4.3 along with its main features. Section 4.4 describes an implementation method based on the multiplier bootstrap. A simple generalization to linear regression-type problems is presented in Section 4.5. Section 4.6 presents an interesting connection between the post-selection confidence regions to the estimators proposed in the high-dimensional linear regression literature. Finally, in Section 4.7, we discuss various advantages and disadvantages of the approach presented in this chapter.

Many of the proofs are deferred to Appendices 4.A, 4.B and 4.C. Most of the chapter is based on the unified framework developed in Chapter 3, although comments about applicability with dependent random vectors are given at appropriate places. Appendix 4.D provides the theoretical background and some new results related to high-dimensional central limit theorem and consistency of multiplier bootstrap. These results are required for computation of joint quantiles for the confidence regions discussed. Appendix 4.E describes a specific dependence structure where the computation of required quantiles is not very much different from that of the independent setting.

4.1 Notation and Problem Formulation

Let $(X_i^\top, Y_i)^\top \in \mathbb{R}^p \times \mathbb{R}, 1 \leq i \leq n$ represent a sample of n observations. The covariate vectors X_i are column vectors. In case p varies with n , this should be interpreted as a triangular array. Throughout, the term “model” is used to specify the subset of covariates present in the regression. We do not assume a linear model (in any sense) to be true anywhere for any choice of covariates in this or in the subsequent sections. This is in accordance with the assumption-lean framework of Chapter 3.

For any vector $v \in \mathbb{R}^q (q \geq 1)$ and $1 \leq j \leq q$, $v(j)$ denotes the j -th coordinate of v . For any non-empty model M given by a subset of $\{1, 2, \dots, q\}$, $v(M)$ denotes a sub-vector of v with indices in M . The notation $|M|$ is used to denote the cardinality of M . For instance, if $M = \{2, 4\}$ and $q \geq 4$, then $v(M) = (v(2), v(4))$. If $M = \{j\}$ is a singleton then $v(j)$ is used instead of $v(\{j\})$. For any non-empty model $M \subseteq \{1, 2, \dots, q\}$ and any symmetric matrix $A \in \mathbb{R}^{q \times q}$, let $A(M)$ denote the sub-matrix of A with indices in $M \times M$ and for $1 \leq j, k \leq q$, let $A(j, k)$ denotes the value at the j -th row and the k -th column of A . Define the r -norm of a vector $v \in \mathbb{R}^q$ for $1 \leq r \leq \infty$ as

$$\|v\|_r := \left(\sum_{j=1}^q |v(j)|^r \right)^{1/r}, \quad \text{for } 1 \leq r < \infty, \quad \text{and} \quad \|v\|_\infty := \max_{1 \leq j \leq q} |v(j)|.$$

Let $\|v\|_0$ denote the number of non-zero entries in v (note this is not a norm). For any matrix A , let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A . Also, let the elementwise maximum and the operator norm be defined, respectively, as

$$\|A\|_\infty := \max_{1 \leq j, k \leq q} |A(j, k)|, \quad \text{and} \quad \|A\|_{op} := \sup_{\|\delta\|_2 \leq 1} \|A\delta\|_2.$$

The following inequalities will be used throughout without any special mention. For

any matrix $A \in \mathbb{R}^{q \times q}$ and $u, v \in \mathbb{R}^q$,

$$\|v\|_1 \leq \|v\|_0^{1/2} \|v\|_2, \quad \|Av\|_\infty \leq \|A\|_\infty \|v\|_1, \quad \text{and} \quad |u^\top Av| \leq \|A\|_\infty \|u\|_1 \|v\|_1. \quad (4.1)$$

For any $1 \leq k \leq p$, define the set of models

$$\mathcal{M}(k) := \{M : M \subseteq \{1, 2, \dots, p\}, 1 \leq |M| \leq k\},$$

so that $\mathcal{M}(p)$ is the power set of $\{1, 2, \dots, p\}$ with the deletion of the empty set. The set $\mathcal{M}(k)$ denotes the set of all non-empty models of size bounded by k .

Traditionally, it is common to include an intercept term when fitting the linear regression. To avoid extra notation, we assume that all covariates under consideration are included in the vectors X_i . So, take the first coordinate of all X_i 's to be 1, that is, $X_i(1) = 1$ for all $1 \leq i \leq n$, if one wants the intercept. To proceed further, assume that the observations are independent but possibly non-identically distributed. This is exactly the assumption-lean framework introduced in Section 3.4 of Chapter 3. The notations \mathbb{E} and \mathbb{P} are used to denote expectation and probability computed with respect to all the randomness involved.

For any $M \in \{1, 2, \dots, p\}$, define the ordinary least squares empirical risk (or objective) function as

$$\hat{R}_n(\theta; M) := \frac{1}{n} \sum_{i=1}^n \{Y_i - X_i^\top(M)\theta\}^2, \quad \text{for } \theta \in \mathbb{R}^{|M|}. \quad (4.2)$$

Using this, define the expected risk (or objective) function as

$$R_n(\theta; M) := \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\{Y_i - X_i^\top(M)\theta\}^2 \right], \quad \text{for } \theta \in \mathbb{R}^{|M|}. \quad (4.3)$$

Define the related matrices and vectors as

$$\begin{aligned}\widehat{\Sigma}_n &:= \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \widehat{\Gamma}_n := \frac{1}{n} \sum_{i=1}^n X_i Y_i \in \mathbb{R}^p, \\ \Sigma_n &:= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i X_i^\top] \in \mathbb{R}^{p \times p}, \quad \text{and} \quad \Gamma_n := \frac{1}{n} \sum_{i=1}^n \mathbb{E} [X_i Y_i] \in \mathbb{R}^p.\end{aligned}$$

All the solutions for post-selection inference to be discussed in the forthcoming sections depend on two error norms related to these matrices and vectors. Define the estimation errors of Σ_n and Γ_n as:

$$\begin{aligned}\mathcal{D}_{1n}^\Gamma &:= \left\| \widehat{\Gamma}_n - \Gamma_n \right\|_\infty = \sup_{M \in \mathcal{M}(1)} \left\| \widehat{\Gamma}_n(M) - \Gamma_n(M) \right\|_\infty, \\ \mathcal{D}_{2n}^\Sigma &:= \left\| \widehat{\Sigma}_n - \Sigma_n \right\|_\infty = \sup_{M \in \mathcal{M}(2)} \left\| \widehat{\Sigma}_n(M) - \Sigma_n(M) \right\|_\infty.\end{aligned}\tag{4.4}$$

Recall the notation $\Sigma_n(M)$ and $\Gamma_n(M)$ as sub-matrix and sub-vector of Σ_n and Γ_n , respectively. Finally, define the least squares estimator and the corresponding target for model M as

$$\widehat{\beta}_{n,M} := \arg \min_{\theta \in \mathbb{R}^{|M|}} \widehat{R}_n(\theta; M), \quad \text{and} \quad \beta_{n,M} := \arg \min_{\theta \in \mathbb{R}^{|M|}} R_n(\theta; M),\tag{4.5}$$

for all $M \subseteq \{1, 2, \dots, p\}$. Observe that $\widehat{\beta}_{n,M}$ and $\beta_{n,M}$ are vectors in $\mathbb{R}^{|M|}$. They are not sub-vectors of any fixed vector in \mathbb{R}^p . This is the reason we specifically write M as a subscript and not in parenthesis. See Section 3.1 of [Berk et al. \(2013\)](#) for a related discussion.

Remark 4.1.1 Note that the objective functions $\widehat{R}_n(\theta; M)$ and $R_n(\theta; M)$ defined in (4.2) and (4.3) are convex quadratic functions of $\theta \in \mathbb{R}^{|M|}$ and so the minimizers always exist. Under the assumption of strict positive definiteness of Σ_n (or just $\Sigma_n(M)$), $R_n(\theta; M)$ is strictly convex and implies a unique minimizer, $\beta_{n,M}$. This

uniqueness does not require any specific relation between model size $|M|$ and the sample size n . In other words, $\beta_{n,M}$ is well-defined even if $|M| > n$ as long as $\Sigma_n(M)$ is non-singular. However, $\hat{\beta}_{n,M}$ can only be well-defined for the case $|M| \leq n$ since $\hat{\Sigma}_n(M)$ has rank at most $\min\{|M|, n\}$. \diamond

As shown in Chapter 3, under very mild assumptions, $\hat{\beta}_{n,M} - \beta_{n,M}$ converges to zero as n tends to infinity for any fixed, non-random model M . In view of this fact, $\hat{\beta}_{n,M}$ is estimating $\beta_{n,M}$. So, $\beta_{n,M}$ is the target of estimation when using $\hat{\beta}_{n,M}$. Also, for a fixed M , $\hat{\beta}_{n,M}$ has an asymptotic normal distribution, i.e.,

$$n^{1/2} \left(\hat{\beta}_{n,M} - \beta_{n,M} \right) \xrightarrow{\mathcal{L}} N_{|M|} (0, AV_M),$$

for some positive definite matrix AV_M that depends on M and some moments of (X, Y) . See the linear representation (3.28) of Chapter 3. The notation $\xrightarrow{\mathcal{L}}$ denotes the convergence in law (or distribution) and $N_{|M|}(0, AV_M)$ denotes the multivariate normal distribution on $\mathbb{R}^{|M|}$ with mean zero and covariance matrix AV_M . This asymptotic normality allows for a construction of a $(1 - \alpha)$ -confidence region $\hat{\mathcal{R}}_{n,M}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\beta_{n,M} \in \hat{\mathcal{R}}_{n,M} \right) \geq 1 - \alpha,$$

for any fixed $\alpha \in [0, 1]$. This implies that for any two fixed, non-random models M_1 and M_2 ,

$$\liminf_{n \rightarrow \infty} \inf_{M \in \{M_1, M_2\}} \mathbb{P} \left(\beta_{n,M} \in \hat{\mathcal{R}}_{n,M} \right) \geq 1 - \alpha. \quad (4.6)$$

Suppose now the data is used to select between the models M_1 and M_2 by some criterion formal or otherwise. Let the final model (which is now random) be denoted by \hat{M} . About this random model, the only available information is

$\mathbb{P}(\widehat{M} \in \{M_1, M_2\}) = 1$. Observe that

$$\begin{aligned} \mathbb{P}(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{\widehat{M}}) &= \mathbb{P}\left(\beta_{n,M_1} \in \widehat{\mathcal{R}}_{M_1} \mid \widehat{M} = M_1\right) \mathbb{P}(\widehat{M} = M_1) \\ &\quad + \mathbb{P}\left(\beta_{n,M_2} \in \widehat{\mathcal{R}}_{M_2} \mid \widehat{M} = M_2\right) \mathbb{P}(\widehat{M} = M_2). \end{aligned} \quad (4.7)$$

There are a few comments to be made from this equation.

- Even after choosing a random model \widehat{M} , the target of linear regression estimator $\widehat{\beta}_{n,\widehat{M}}$ is $\beta_{n,\widehat{M}}$, since this is what happens if \widehat{M} is degenerate on a particular model. This fact requires a proof as shown in Lemma 4 and Lemma 5. It is crucial to recognize that $\beta_{n,\widehat{M}}$ is a random quantity for a random model \widehat{M} .
- Having a guarantee as stated in the confidence bound (4.6) does not imply anything (in general) about the conditional probabilities in Equation (4.7) unless \widehat{M} does not depend on the data. Note that if \widehat{M} is chosen independently of the data used to construct $\widehat{\mathcal{R}}_{n,M}$ then the conditional probabilities in (4.7) become the marginal probabilities and so valid post-selection inference is obtained. The recent paper [Rinaldo et al. \(2016\)](#) uses sample splitting as a method of inference after model-selection. One important disadvantage of sample splitting is its inability to extend to dependent observations even though for independent observations it gives very general and powerful results. For more comments, see section 7 of [Rinaldo et al. \(2016\)](#).
- The conditional inference put forward by [Lee et al. \(2016\)](#), [Tibshirani et al. \(2016\)](#) and [Tian et al. \(2016\)](#) among others tries to bound the conditional probabilities on the right hand side of (4.7) for some fixed method of choosing \widehat{M} . [Berk et al. \(2013\)](#) lower bounds the left hand side of (4.7) under the assumption of fixed covariates and homoscedastic Gaussian errors for arbitrary method of

choosing \widehat{M} . See Chapter 2 for more details on both these approaches.

The problem of valid post model-selection inference is to construct the set of confidence regions $\{\widehat{\mathcal{R}}_{n,M} : M \in \mathcal{M}\}$ for some fixed (non-random) set of models \mathcal{M} such that for any random model \widehat{M} (possibly) depending on the same data satisfying $\mathbb{P}(\widehat{M} \in \mathcal{M}) \rightarrow 1$,

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}) \geq 1 - \alpha. \quad (4.8)$$

Throughout the chapter, $\alpha \in [0, 1]$ is fixed and the confidence regions $\widehat{\mathcal{R}}_{n,M}$ do depend on α . The guarantee (4.8) requires the confidence asymptotically since we believe that no finite sample confidence statements can be made without significantly loading the assumptions. The following remarks make the understanding of setting clearer.

Remark 4.1.2 If a sequence of confidence regions $\widehat{\mathcal{R}}_{n,M}$ that satisfy (4.8) can be constructed, then we are equipped to perform valid post-selection inference since by duality of confidence regions and hypothesis testing, one can also formally test hypothesis related to the randomly chosen model \widehat{M} . \diamond

Remark 4.1.3 Comparing confidence statement (4.8) and (4.6), we note a major difference that in (4.8), we are trying to cover a random parameter with a random confidence region while in (4.6), we are trying to cover a fixed (non-random) parameter with a random confidence region. Often it is useful to have infimum over a class of data-generating distributions before the probability in (4.8) which is referred to as honest confidence in Rinaldo et al. (2016), that is,

$$\liminf_{n \rightarrow \infty} \inf_{P^{\otimes n} \in \mathcal{P}_n} \mathbb{P}(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}) \geq 1 - \alpha,$$

for some class of probability distributions \mathcal{P}_n , as in Section 3.A. This “honesty” holds for our results too, however, we do not stress on this further. \diamond

Remark 4.1.4 In order for the results stated in the following sections to be valid, one should *not* look at the dataset to decide whether or not to include some covariate in model-selection or if some models can be included in \mathcal{M} . For example one should *not* do the following. Let the sample of observations be $\{(Y_i, X_i(1), X_i(2)) : 1 \leq i \leq n\}$. Choose by any method some model \widehat{M} from the set of models $\{\{1\}, \{2\}, \{1, 2\}\}$. Now make a residual plot and, say, the residuals seem correlated with the covariates in \widehat{M} . So, add the quadratic terms for covariates in \widehat{M} . This gives $2|\widehat{M}|$ covariates in total. Then perform another model selection with all possible combinations of $2|\widehat{M}|$ new covariates. Continue until the residual plot “looks good enough”. This procedure does not have a fixed value of p and a fixed set \mathcal{M} in our notation. This scenario is possible when trying to fit a polynomial regression where the data analyst goes on adding higher powers until satisfied with the residual plot. This comment also helps clarify why one cannot consider \mathcal{M} to be the set of models generated by the LASSO path because the models that constitute the LASSO path are data-dependent. See [Bellec \(2016\)](#) for a possible way to solve this problem. \diamond

4.2 Equivalence of Post-selection and Simultaneous Inference

The first step towards achieving the goal of constructing a set of confidence regions $\{\widehat{\mathcal{R}}_{n,M} : M \in \mathcal{M}\}$ satisfying (4.8) is to convert the post-selection inference problem into a simultaneous inference problem. This conversion is provided in Theorem 3 and this (one of the implications) is also the basis of the method in [Berk et al. \(2013\)](#). The following theorem is proved finite sample and the version with \liminf follows readily from this result.

Theorem 3. *For any set of confidence regions $\{\widehat{\mathcal{R}}_{n,M} : M \in \mathcal{M}\}$ and $\alpha \in [0, 1]$, the*

following two statements are equivalent:

1. the post-selection inference problem is solved, that is,

$$\mathbb{P}\left(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}\right) \geq 1 - \alpha,$$

for all random models $\widehat{M} \in \mathcal{M}$ depending on the data.

2. the simultaneous inference problem is solved, that is,

$$\mathbb{P}\left(\bigcap_{M \in \mathcal{M}} \left\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\right\}\right) \geq 1 - \alpha.$$

Proof. (2) \Rightarrow (1): Define for every $M \in \mathcal{M}$, the event $\mathcal{A}_M = \{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\}$. Since $\mathbb{P}(\widehat{M} \in \mathcal{M}) = 1$, extending the equality (4.7) implies

$$\begin{aligned} \mathbb{P}\left(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}\right) &= \mathbb{P}\left(\mathcal{A}_{\widehat{M}}\right) = \sum_{M' \in \mathcal{M}} \mathbb{P}\left(\mathcal{A}_{M'} \cap \{\widehat{M} = M'\}\right) \\ &\geq \sum_{M' \in \mathcal{M}} \mathbb{P}\left(\left\{\bigcap_{M \in \mathcal{M}} \mathcal{A}_M\right\} \cap \{\widehat{M} = M'\}\right) \\ &= \mathbb{P}\left(\bigcap_{M \in \mathcal{M}} \mathcal{A}_M\right) = \mathbb{P}\left(\bigcap_{M \in \mathcal{M}} \left\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\right\}\right). \end{aligned}$$

Summarizing these inequalities proves (2) \Rightarrow (1).

(1) \Rightarrow (2): To prove this implication, note that

$$\mathbb{P}\left(\bigcap_{M \in \mathcal{M}} \left\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\right\}\right) = \mathbb{E}\left[\min_{M \in \mathcal{M}} \mathbb{1}\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\}\right],$$

where $\mathbb{1}\{A\}$ denotes the indicator of event A . Now, take a random model \widehat{M} such

that

$$\widehat{M} \in \arg \min_{M \in \mathcal{M}} \mathbb{1}\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\} \quad \Rightarrow \quad \mathbb{1}\{\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}\} = \min_{M \in \mathcal{M}} \mathbb{1}\{\beta_{n,M} \in \widehat{\mathcal{R}}_{n,M}\}.$$

Taking expectation on both sides of this equality proves (1) \Rightarrow (2). \square

Remark 4.2.1 Theorem 3 is very useful not just in providing pathways to solving the post-selection inference problem but also in showing some interesting infeasibility results. For instance, in the examples mentioned in Chapter 1, we have shown that the practice of statistics is free flowing where the previously obtained information is used to guide further exploration. Instead of solving the PoSI problem with a fixed universe \mathcal{M} as in Theorem 3, one might be interested in solving the problem with increasing universes. More precisely, one might define the sequence of universes $\{\mathcal{M}_k, k \geq 1\}$ where

$$\mathcal{M}_k := \bigcup_{s=1}^k \{M : M \subseteq \{1, 2, \dots, p\}^{\otimes s}\}.$$

This means that \mathcal{M}_k contains all subsets of covariates obtained through interactions of order at most k of the initial set of variables. The intended methodology is that the analyst explores the data with the initial set of variables. If the initial covariates seem insufficient to explain the data, the analyst will consider pairwise interactions of these covariates and continues to explore the data. This process continues until the analyst is satisfied with the goodness of fit. Essentially, the analyst picks a model \widehat{M} in the universe $\mathcal{M}_{\widehat{k}}$ where $\widehat{k} \geq 1$ is itself based on the data. The inference problem becomes

$$\mathbb{P}\left(\beta_{n,\widehat{M}} \in \widehat{\mathcal{R}}_{n,\widehat{M}}\right) \geq 1 - \alpha,$$

for all random models \widehat{M} in all random universes $\mathcal{M}_{\widehat{k}}$. Now applying Theorem 3

yields that we need confidence regions such that

$$\mathbb{P} \left(\bigcap_{k \geq 1} \bigcap_{M \in \mathcal{M}_k} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\} \right) \geq 1 - \alpha.$$

Because \hat{k} is obtained in a complicated fashion, there is no apparent gain from the structure of universes. One can, however, see the structure of universes as hierarchy and use them in the way of constructing simultaneous confidence regions. (Intuitively, think of Bonferroni as giving $\alpha/|\mathcal{M}|, \dots, \alpha/|\mathcal{M}|$ for each model $M \in \mathcal{M}$ but now with hierarchy use $\alpha_1, \dots, \alpha_M$ that are decreasing and add up to α .)

If the sequence of universes is complicated and big enough, then it is inevitable that some of the confidence regions have to be the whole parameter space to satisfy the simultaneity guarantee. \diamond

Remark 4.2.2 Note that in view of Theorem 3 valid post-selection inference is inherently a high-dimensional problem in the sense that the number of parameters we want to estimate and infer about is in general much larger than the sample size available. For illustration, consider the “not-so-bad” usual regression setting where there are $p = 10$ covariates and $n = 500$ observations. There are 50 observations per parameter in the full model which is considered a fixed-dimensional problem. Now, for the post-selection inference problem with all non-empty sub-models, there are $2^p - 1 = 1023$ vector parameters of varying dimensions. So, there are $p2^{p-1} = 5120$ many parameters from 500 observations which falls into the high-dimensional category. \diamond

Theorem 3 shows that in order to achieve the goal (4.8) of valid post-selection inference, it is enough and necessary to construct a set of confidence regions $\hat{\mathcal{R}}_{n,M}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}} \{\beta_{n,M} \in \hat{\mathcal{R}}_{n,M}\} \right) \geq 1 - \alpha. \quad (4.9)$$

All of our solutions for post-selection inference problem in this chapter and the following are solved based on (4.9). Note that the equivalence proved in Theorem 3 has nothing to do with linear regression and is true for any set of targets $\{\beta_{n,M} : M \in \mathcal{M}\}$.

4.3 First Approach for Post-Selection Inference

Now that we are equipped with the required notation, we proceed to construct confidence regions $\widehat{\mathcal{R}}_{n,M}$ for linear regression. Recall from Equation (3.12) of Chapter 3 that the least squares estimator and target given in (4.5) can be written as

$$\begin{aligned}\widehat{\beta}_{n,M} &= \arg \min_{\theta \in \mathbb{R}^{|M|}} \left\{ \theta^\top \widehat{\Sigma}_n(M) \theta - 2\theta^\top \widehat{\Gamma}_n(M) \right\}, \quad \text{and,} \\ \beta_{n,M} &= \arg \min_{\theta \in \mathbb{R}^{|M|}} \left\{ \theta^\top \Sigma_n(M) \theta - 2\theta^\top \Gamma_n(M) \right\}.\end{aligned}\tag{4.10}$$

Define the confidence regions

$$\widehat{\mathcal{R}}_{n,M} = \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \widehat{\Sigma}_n(M) \left\{ \widehat{\beta}_{n,M} - \theta \right\} \right\|_\infty \leq C_{1n}^\Gamma(\alpha) + C_{2n}^\Sigma(\alpha) \|\theta\|_1 \right\}, \tag{4.11}$$

$$\widehat{\mathcal{R}}_{n,M}^\dagger = \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \widehat{\Sigma}_n(M) \left\{ \widehat{\beta}_{n,M} - \theta \right\} \right\|_\infty \leq C_{1n}^\Gamma(\alpha) + C_{2n}^\Sigma(\alpha) \left\| \widehat{\beta}_{n,M} \right\|_1 \right\} \tag{4.12}$$

for every $M \in \mathcal{M}(p)$, where $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$ are bivariate joint quantiles of \mathcal{D}_{1n}^Γ and \mathcal{D}_{2n}^Σ , that is,

$$\mathbb{P} \left(\mathcal{D}_{1n}^\Gamma \leq C_{1n}^\Gamma(\alpha) \quad \text{and} \quad \mathcal{D}_{2n}^\Sigma \leq C_{2n}^\Sigma(\alpha) \right) \geq 1 - \alpha. \tag{4.13}$$

The constants $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$, of course, depend on the true distribution of the observations and the number of full set of covariates p . These dependencies are suppressed throughout. It is important to realize that these constants are not unique and any choice that satisfies the probability guarantee is allowed. These constants are not known and should be estimated from the data. A bootstrap procedure to

estimate these bivariate quantiles is described in Section 4.4. The estimation errors \mathcal{D}_{1n}^Γ and \mathcal{D}_{2n}^Σ converge by the law of large numbers to zero as $n \rightarrow \infty$ under mild conditions. Therefore, $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$ converge to zero as $n \rightarrow \infty$.

The following two theorems prove validity (4.9) with the confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}_{n,M}^\dagger$ given above for the set of models $\mathcal{M}(p)$ and $\mathcal{M}(k)$ (for some $k \leq p$), respectively.

Theorem 4. *The set of confidence regions $\{\hat{\mathcal{R}}_{n,M} : M \in \mathcal{M}(p)\}$ defined in (4.11) satisfies*

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \{ \beta_{n,M} \in \hat{\mathcal{R}}_{n,M} \} \right) \geq 1 - \alpha, \quad (4.14)$$

Furthermore, for any random model \hat{M} with $\mathbb{P}(\hat{M} \in \mathcal{M}(p)) = 1$,

$$\mathbb{P} \left(\beta_{n,\hat{M}} \in \hat{\mathcal{R}}_{n,\hat{M}} \right) \geq 1 - \alpha.$$

Proof. The proof is surprisingly elementary and involves simple manipulation of the estimating equations. From the definitions of the estimator and target $\hat{\beta}_{n,M}$, $\beta_{n,M}$ given in (4.10), these vectors satisfy the equations

$$\hat{\Sigma}_n(M) \hat{\beta}_{n,M} - \hat{\Gamma}_n(M) = 0, \quad \text{for all } M \in \mathcal{M}(p), \quad (4.15)$$

$$\Sigma_n(M) \beta_{n,M} - \Gamma_n(M) = 0, \quad \text{for all } M \in \mathcal{M}(p). \quad (4.16)$$

Adding and subtracting $\beta_{n,M}$ from $\hat{\beta}_{n,M}$ in equation (4.15) implies

$$\hat{\Sigma}_n(M) \{ \hat{\beta}_{n,M} - \beta_{n,M} \} = \hat{\Gamma}_n(M) - \hat{\Sigma}_n(M) \beta_{n,M} \quad \text{for all } M \in \mathcal{M}(p).$$

Subtracting Equation (4.16) from this equation leads to

$$\left[\hat{\Gamma}_n(M) - \Gamma_n(M) \right] - \left[\hat{\Sigma}_n(M) - \Sigma_n(M) \right] \beta_{n,M} = \hat{\Sigma}_n(M) \left\{ \hat{\beta}_{n,M} - \beta_{n,M} \right\},$$

for all $M \in \mathcal{M}(p)$. Taking $\|\cdot\|_\infty$ on both sides and using the triangle inequality,

$$\left\| \hat{\Sigma}_n(M) \left\{ \hat{\beta}_{n,M} - \beta_{n,M} \right\} \right\|_\infty \leq \left\| \hat{\Gamma}_n(M) - \Gamma_n(M) \right\|_\infty + \left\| \left[\hat{\Sigma}_n(M) - \Sigma_n(M) \right] \beta_{n,M} \right\|_\infty,$$

holds for all $M \in \mathcal{M}(p)$. By an application of the second inequality in (4.1), for all $M \in \mathcal{M}(p)$,

$$\left\| \hat{\Sigma}_n(M) \left\{ \hat{\beta}_{n,M} - \beta_{n,M} \right\} \right\|_\infty \leq \left\| \hat{\Gamma}_n(M) - \Gamma_n(M) \right\|_\infty + \left\| \hat{\Sigma}_n(M) - \Sigma_n(M) \right\|_\infty \|\beta_{n,M}\|_1. \quad (4.17)$$

Since $\hat{\Gamma}_n(M) - \Gamma_n(M)$, $\hat{\Sigma}_n(M) - \Sigma_n(M)$ are sub-vector and sub-matrix of $\hat{\Gamma}_n - \Gamma_n$ and $\hat{\Sigma}_n - \Sigma_n$, this inequality implies,

$$\left\| \hat{\Sigma}_n(M) \left\{ \beta_{n,M} - \hat{\beta}_{n,M} \right\} \right\|_\infty \leq \left\| \hat{\Gamma}_n - \Gamma_n \right\|_\infty + \left\| \hat{\Sigma}_n - \Sigma_n \right\|_\infty \|\beta_{n,M}\|_1, \text{ for } M \in \mathcal{M}(p). \quad (4.18)$$

Note that all the equations and inequalities above are deterministic and hold for any sample. Changing this into a probability one statement and using the definitions of \mathcal{D}_{1n}^Γ and \mathcal{D}_{2n}^Σ in (4.4),

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \left\| \Sigma_n(M) \left\{ \beta_{n,M} - \hat{\beta}_{n,M} \right\} \right\|_\infty \leq \mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1 \right\} \right) = 1. \quad (4.19)$$

The definition of $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ in (4.13) proves the required result (4.14). The second result follows by an application of Theorem 3. \square

Remark 4.3.1 (Validity Guarantee) It is interesting to note that the guarantee (4.14) in Theorem 4 is valid for every sample size n and any number of covariates p . In particular, $p \gg n$ and $p = \infty$ can be treated without any difficulty. It is also important to observe that $\hat{\Sigma}_n(M)$ becomes singular for $|M| > n$ and this makes the confidence region $\hat{\mathcal{R}}_{n,M}$ infinitely wide.

The finite sample guarantee (4.14) only holds if $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ satisfy (4.13) for all $p, n \geq 1$. In general, these bivariate quantiles can only be estimated consistently in the asymptotic sense as explained in Section 4.4. \diamond

Remark 4.3.2 (Independence of Observations) For simplicity in the discussion above, the assumption of independence of random vectors $(X_i, Y_i), 1 \leq i \leq n$ is used. Theorem 4 holds without this assumption since no use of this assumption was made in its proof. Validity of post-selection guarantee holds as long as $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$ are valid quantiles in the sense of (4.13). In light of this remark, the applicability of Theorem 4 holds way beyond the unified framework of Chapter 3. \diamond

Remark 4.3.3 (Restriction of Models for Selection) The validity guarantee (4.14) in Theorem 4 implies that one can use all models in the variable selection procedure. In this case, the confidence region $\hat{\mathcal{R}}_{n,M}$ does not depend on the set of models used in selection procedure. It might be of practical importance to consider the post-selection inference problem when the set of models \mathcal{M} used in selection is different from (or smaller than) $\mathcal{M}(p)$. Following the proof of Theorem 4, the transition from inequality (4.17) to (4.18) should be changed as follows:

$$\left\| \hat{\Sigma}_n(M) \left\{ \hat{\beta}_{n,M} - \beta_{n,M} \right\} \right\|_\infty \leq \mathcal{D}_{1n}^\Gamma(\mathcal{M}) + \mathcal{D}_{2n}^\Sigma(\mathcal{M}) \|\beta_{n,M}\|_1, \quad \text{for all } M \in \mathcal{M},$$

where

$$\mathcal{D}_{1n}^\Gamma(\mathcal{M}) := \sup_{M \in \mathcal{M}} \left\| \widehat{\Gamma}_n(M) - \Gamma_n(M) \right\|_\infty \quad \text{and} \quad \mathcal{D}_{2n}^\Sigma(\mathcal{M}) := \sup_{M \in \mathcal{M}} \left\| \widehat{\Sigma}_n(M) - \Sigma_n(M) \right\|_\infty.$$

Note from definition (4.4) that as long as \mathcal{M} contains all models of size 2, then $\mathcal{D}_{1n}^\Gamma(\mathcal{M}) = \mathcal{D}_{1n}^\Gamma$ and $\mathcal{D}_{2n}^\Sigma(\mathcal{M}) = \mathcal{D}_{2n}^\Sigma$. For the final confidence region, let $C_{1n}^\Gamma(\alpha; \mathcal{M})$ and $C_{2n}^\Sigma(\alpha; \mathcal{M})$ be the bivariate quantiles of $\mathcal{D}_{1n}^\Gamma(\mathcal{M})$ and $\mathcal{D}_{2n}^\Sigma(\mathcal{M})$, that is,

$$\mathbb{P} \left(\mathcal{D}_{1n}^\Gamma(\mathcal{M}) \leq C_{1n}^\Gamma(\alpha; \mathcal{M}) \quad \text{and} \quad \mathcal{D}_{2n}^\Sigma(\mathcal{M}) \leq C_{2n}^\Sigma(\alpha; \mathcal{M}) \right) \geq 1 - \alpha,$$

and set for $M \in \mathcal{M}$,

$$\widehat{\mathcal{R}}_{n,M} := \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \widehat{\Sigma}_n(M) \left\{ \widehat{\beta}_{n,M} - \theta \right\} \right\|_\infty \leq C_{1n}^\Gamma(\alpha; \mathcal{M}) + C_{2n}^\Sigma(\alpha; \mathcal{M}) \|\theta\|_1 \right\}.$$

The guarantee of Theorem 4 holds true with these definitions for any *non-random* set of models \mathcal{M} . A similar remark holds for all the confidence regions defined below and will not be repeated. \diamond

The shape of the confidence region $\widehat{\mathcal{R}}_{n,M}$ is hard to visualize from the definition (4.11) and it is also difficult to study the Lebesgue measure of this confidence region. With a different parametrization of $\widehat{\mathcal{R}}_{n,M}$, Belloni et al. (2017) proves this confidence region is a convex polyhedral set. See Equation (42) of the supplementary material of Belloni et al. (2017). For these reasons, we also prove the asymptotic validity of the confidence region $\widehat{\mathcal{R}}_{n,M}^\dagger$ defined in (4.12). To state the theorem, consider the following assumption:

(A1)(k) The estimation error \mathcal{D}_{2n}^Σ in connection with $1 \leq k \leq p$ satisfy

$$k\mathcal{D}_{2n}^\Sigma = o_p(\Lambda_n(k)) \quad \text{as} \quad n \rightarrow \infty,$$

where

$$\Lambda_n(k) := \min_{M \in \mathcal{M}(k)} \lambda_{\min}(\Sigma_n(M)).$$

Some comments can be helpful. This assumption is used for uniform consistency of least squares estimator in $\|\cdot\|_1$ -norm as shown in Lemma 4. The rate of convergence of \mathcal{D}_{2n}^Σ to zero imply a rate constraint on k . Here too $k = k_n$ is allowed to be a sequence depending on n . As can be expected, the dependence structure between the random vectors $(X_i, Y_i), 1 \leq i \leq n$ and their moments decide the rate at which \mathcal{D}_{2n}^Σ converges to zero. See Lemma 5 for more details. The theorem is stated with this high level assumption so that it is more widely applicable in particular to various structural dependencies on observations. Note that assumption (A1)(k) allows for the minimum eigenvalue of Σ_n to converge to zero as $n \rightarrow \infty$ if $p = p_n$ changes with n .

Before proceeding to the proof of validity of $\hat{\mathcal{R}}_{n,M}^\dagger$ as valid post-selection confidence regions, we prove uniform-in-model consistency of $\hat{\beta}_{n,M}$ to $\beta_{n,M}$. See Appendix 4.A for a detailed proof.

Lemma 4. *For all $k \geq 1$ satisfying $k\mathcal{D}_{2n}^\Sigma \leq \Lambda_n(k)$ and for all $M \in \mathcal{M}(k)$,*

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_1 \leq \frac{|M| (\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1)}{\Lambda_n(k) - k\mathcal{D}_{2n}^\Sigma}. \quad (4.20)$$

The following theorem proves the validity for $\hat{\mathcal{R}}_{n,M}^\dagger$.

Theorem 5. *For every $1 \leq k \leq p$ that satisfies (A1)(k), the confidence regions $\{\hat{\mathcal{R}}_{n,M}^\dagger : M \in \mathcal{M}(p)\}$ defined in (4.12) satisfies*

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(k)} \left\{ \beta_{n,M} \in \hat{\mathcal{R}}_{n,M}^\dagger \right\} \right) \geq 1 - \alpha.$$

Proof. The starting point of this proof is Equation (4.19). Under assumption (A1)(k), Lemma 4 (inequality (4.20)) implies that for all $M \in \mathcal{M}(k)$,

$$\begin{aligned} \left| \frac{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \hat{\beta}_{n,M} \right\|_1}{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \beta_{n,M} \right\|_1} - 1 \right| &\leq \frac{\mathcal{D}_{2n}^\Sigma \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_1}{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \beta_{n,M} \right\|_1} \\ &\leq \frac{\mathcal{D}_{2n}^\Sigma}{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \beta_{n,M} \right\|_1} \times \frac{|M| \left\{ \mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \beta_{n,M} \right\|_1 \right\}}{\Lambda_n(k) - |M| \mathcal{D}_{2n}^\Sigma} \\ &\leq \frac{k \mathcal{D}_{2n}^\Sigma}{\Lambda_n(k) - k \mathcal{D}_{2n}^\Sigma}. \end{aligned}$$

Therefore, for $1 \leq k \leq p$ satisfying assumption (A1)(k),

$$\sup_{M \in \mathcal{M}(k)} \left| \frac{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \hat{\beta}_{n,M} \right\|_1}{\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \beta_{n,M} \right\|_1} - 1 \right| \leq \frac{k \mathcal{D}_{2n}^\Sigma / \Lambda(k)}{1 - (k \mathcal{D}_{2n}^\Sigma / \Lambda(k))} = o_p(1).$$

Hence,

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(k)} \left\{ \left\| \Sigma_n(M) \left\{ \beta_{n,M} - \hat{\beta}_{n,M} \right\} \right\|_\infty \leq \mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \left\| \hat{\beta}_M \right\|_1 \right\} \right) = 1.$$

The definition of $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ in (4.13) proves the required result. \square

Remark 4.3.4 (Shape of $\hat{\mathcal{R}}_{n,M}^\dagger$) It is easy to see that the confidence region $\hat{\mathcal{R}}_{n,M}^\dagger$ is described by $2|M|$ linear inequalities (with random coefficients) and is a polyhedral set. The general shape is described as parallelepiped. The Lebesgue measure of this confidence region is much easier to study than that of the region $\hat{\mathcal{R}}_{n,M}$ as presented in Lemma 6 below. \diamond

Remark 4.3.5 (Centering and Scaling) It is not difficult to see that the confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}_{n,M}^\dagger$ are not equivariant with respect to linear transformation of covariates or response. Equivariance is an important feature for practical interpretation.

A simple way to obtain equivariance with respect to diagonal linear transformations of the random vectors would be to use linear regression with covariates centered and scaled to have sample mean zero and sample variance 1. Since the validity of confidence regions does not require independence, as mentioned in Remark 4.E.1, this centering and scaling based on the data will not effect the post-selection guarantee as long as marginal means and variances are estimated consistently. This might also have an effect on the volume of the confidence regions not in terms of rate but in terms of constants since the intercept is not longer needed in $\|\beta_{n,M}\|_1$. See Section 4.7 for more details. \diamond

Remark 4.3.6 (Case of Fixed Covariates) Since most of the post-selection inference literature as reviewed in Section 4.1 deals with the case of fixed covariates, it is of particular interest to understand how our confidence regions behave in this case. For fixed covariates, the distribution of X_i is degenerate at the observed value X_i and hence,

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i X_i^\top] = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top = \hat{\Sigma}_n.$$

Therefore, $\mathcal{D}_{2n}^\Sigma = \|\hat{\Sigma}_n - \Sigma_n\|_\infty = 0$ and so, $C_2(\alpha) = 0$. Also, note that in this case

$$\beta_{n,M} = \left(\frac{1}{n} \sum_{i=1}^n X_i(M) X_i^\top(M) \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i(M) \mathbb{E}[Y_i] \right).$$

Hence, in case of fixed covariates,

$$\hat{\mathcal{R}}_{n,M} = \hat{\mathcal{R}}_{n,M}^\dagger = \left\{ \left\| \Sigma_n(M) \left\{ \hat{\beta}_{n,M} - \beta_{n,M} \right\} \right\|_\infty \leq C_{1n}^\Gamma(\alpha) \right\}.$$

Note that under fixed covariates assumption (A1)(k) is trivially satisfied since \mathcal{D}_{2n}^Σ . Thus by Theorem 4 (or 5), finite sample valid post-selection inference holds for all

model sizes in case of fixed covariates under no model or distributional assumptions as were required in Berk et al. (2013).

A nice feature of the methodology proposed in Berk et al. (2013) is that the inference is tight in the sense there exists a model selection procedure such that the post-selection confidence interval has coverage exactly $1 - \alpha$. Even though the confidence region $\widehat{\mathcal{R}}_{n,M}$ is derived under a more general framework, this tightness holds in this generality. This can be easily seen by noting that

$$\begin{aligned} \sup_{M \in \mathcal{M}(p)} \left\| \Sigma_n(M) \left\{ \widehat{\beta}_{n,M} - \beta_{n,M} \right\} \right\|_\infty &= \sup_{M \in \mathcal{M}(p)} \left| \frac{1}{n} \sum_{i=1}^n X_i(M) (Y_i - \mathbb{E}[Y_i]) \right| \\ &= \sup_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n X_i(j) (Y_i - \mathbb{E}[Y_i]) \right| = \mathcal{D}_{1n}^\Gamma. \end{aligned}$$

Take $\widehat{M} = \{\widehat{j}\}$, where

$$\widehat{j} \in \arg \max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^n X_i(j) (Y_i - \mathbb{E}[Y_i]) \right|.$$

For this random model \widehat{M} , the coverage of $\widehat{\mathcal{R}}_{n,\widehat{M}}$ is exactly equal to $(1 - \alpha)$. \diamond

Before proceeding further with the study of the confidence regions, it might be useful to understand the rates at which \mathcal{D}_{1n}^Γ and \mathcal{D}_{2n}^Σ converge to zero under some assumptions on the initial random vectors (X_i, Y_i) , $1 \leq i \leq n$. As mentioned in Remark 4.E.1, the validity of post-selection coverage guarantee does not require independence of random vectors and so, a rate result under “physical dependence” is presented in Appendix 4.E. Set $Z_i = (X_i^\top, Y_i)^\top$ for $1 \leq i \leq n$ and define

$$\widehat{\Omega}_n := \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top, \quad \text{and} \quad \Omega_n := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i Z_i^\top] \in \mathbb{R}^{(p+1) \times (p+1)}.$$

Observe that

$$\max\{\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma\} \leq \left\| \hat{\Omega}_n - \Omega_n \right\|_\infty.$$

The following lemma from [Kuchibhotla and Chakraborty \(2018\)](#) proves a finite sample bound for the expected value of the maximum absolute value of $\hat{\Omega}_n - \Omega_n$. For this result, set for $\alpha > 0$ and any random variable W ,

$$\|W\|_{\psi_\alpha} := \inf \left\{ C > 0 : \mathbb{E} \left[\psi_\alpha \left(\frac{|W|}{C} \right) \right] \leq 1 \right\},$$

where $\psi_\alpha(x) = \exp(x^\alpha) - 1$ for $x \geq 0$. For $0 < \alpha < 1$, $\|\cdot\|_{\psi_\alpha}$ is not a norm but is a quasi-norm. A random variable W satisfying $\|W\|_{\psi_\alpha} < \infty$ is called a sub-Weibull random variable of order α . The special cases $\alpha = 1$ and $\alpha = 2$ correspond to the well-known classes of sub-exponential and sub-Gaussian random variables.

Lemma 5. *Fix $n, p \geq 2$. Suppose the random vectors $Z_i, 1 \leq i \leq n$ are independent and satisfy for some $\alpha \geq 0$*

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p+1} \|Z_i(j)\|_{\psi_\alpha} \leq K_{n,p}, \quad (4.21)$$

for some positive constant $K_{n,p}$. Then

$$\mathbb{E} \left[\sqrt{n} \left\| \hat{\Omega}_n - \Omega_n \right\|_\infty \right] \leq C_\alpha \left\{ A_{n,p} \sqrt{\log p} + K_{n,p}^2 (\log p \log n)^{2/\alpha} n^{-1/2} \right\},$$

where C_α is a positive universal constant that grows at the rate of $(1/\alpha)^{1/\alpha}$ as $\alpha \downarrow 0$ and

$$A_{n,p}^2 := \max_{1 \leq j \leq k \leq p+1} \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i(j)X_i(k)).$$

Proof. See Section 4.1 of [Kuchibhotla and Chakraborty \(2018\)](#). □

The confidence regions proposed through approach 1 are simple parallelepipeds and can be seen as linear transformations of $\|\cdot\|_\infty$ -norm balls. Hence, their Lebesgue measures can be computed exactly. Since that confidence regions are valid over a large number of models, we present a relative Lebesgue measure result uniform over a set of models. For $A \subseteq \mathbb{R}^q$ with $q \geq 1$, let $\mathbf{Leb}(A)$ denote the Lebesgue measure of A with the measure supported on \mathbb{R}^q . For convenience, we do not use different notations for the Lebesgue measure for different $q \geq 1$.

Lemma 6. *For any $k \geq 1$ such that assumption (A1)(k) are satisfied, the uniform relative Lebesgue measure result holds:*

$$\sup_{M \in \mathcal{M}(k)} \frac{\mathbf{Leb}(\widehat{\mathcal{R}}_{n,M}^\dagger) \Lambda_n^{|M|}(k)}{(C_{1n}^\Gamma(\alpha) + C_{2n}^\Sigma(\alpha) \|\beta_{n,M}\|_1)^{|M|}} = O_p(1).$$

Hence, it can be said that $\mathbf{Leb}(\widehat{\mathcal{R}}_{n,M}^\dagger) = O_p(\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1)^{|M|}$ uniformly for $M \in \mathcal{M}(k)$ if $\Lambda_n^{-1}(k) = O(1)$. Moreover, additionally under the setting of Lemma 5,

$$\mathbf{Leb}(\widehat{\mathcal{R}}_{n,M}^\dagger) = O_p\left(\sqrt{\frac{|M| \log p}{n}}\right)^{|M|} \quad \text{uniformly for } M \in \mathcal{M}(k),$$

if p and n satisfy

$$(\log p)^{2/\alpha} (\log n)^{2/\alpha - 1/2} = o(n^{1/2}). \quad (4.22)$$

Proof. See Appendix 4.B for a detailed proof. □

4.4 Computation by Multiplier Bootstrap

All the confidence regions defined in the previous section (and the ones to be defined in the forthcoming sections) depend only on the available data except for the (joint) quantiles $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$. Computation or estimation of joint bivariate quantiles

$C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$ is the most important component of an application of approach 1 for valid post-selection inference. In this section, we apply the high-dimensional central limit theorem and multiplier bootstrap for estimating these quantiles. We note that either a classical bootstrap or the recently popularized method of multiplier bootstrap works for estimating these joint quantiles in the setting described in Lemma 5. See Chernozhukov et al. (2017a) and Zhang and Cheng (2014) for a detailed discussion. For simplicity, we will only describe the method of multiplier bootstrap for the case of independent random vectors. The discussion here applies the central limit theorem and multiplier bootstrap result proved in Appendix 4.D. And we refer to Zhang and Cheng (2014) for the case of dependent settings described in Appendix 4.E.

Define vectors $W_i \in \mathbb{R}^q$ for $1 \leq i \leq n$ containing

$$(\{X_i(j)Y_i\}, 1 \leq j \leq p; \{X_i(l)X_i(m)\}, 1 \leq l \leq m \leq p),$$

with

$$q = 2p + \frac{p(p-1)}{2} = O(p^2).$$

As shown in Equation (4.38), for any $t_1, t_2 \in \mathbb{R}^+ \cup \{0\}$, the set

$$\{\mathcal{D}_{1n}^\Gamma \leq t_1, \mathcal{D}_{2n}^\Sigma \leq t_2\},$$

can be written as a rectangle in terms of

$$S_n^W := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{W_i - \mathbb{E}[W_i]\}.$$

In the unified framework of linear regression, (X_i, Y_i) are possibly non-identically

distributed and so, $\mathbb{E}[W_i]$ are not all equal. Let e_1, e_2, \dots, e_n be independent standard normal random variables and define

$$S_n^{eW} := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i(W_i - \bar{W}_n), \quad \text{where} \quad \bar{W}_n := \frac{1}{n} \sum_{i=1}^n W_i.$$

Write $S_n^{eW}(\mathbf{I})$ for the first p coordinates of S_n^{eW} and $S_n^{eW}(\mathbf{II})$ for the remaining coordinates of S_n^{eW} . The following algorithm gives the pseudo-program for implementing the multiplier bootstrap.

1. Generate N random vectors from $N_n(0, I_n)$, with I_n denoting the identity matrix of dimension n . Let these be denoted by $\{e_{i,j} : 1 \leq i \leq n, 1 \leq j \leq N\}$.
2. Compute the j -th replicate of S_n^{eW} as

$$S_{n,j}^* := \frac{1}{n} \sum_{i=1}^n e_{i,j}(W_i - \bar{W}_n), \quad \text{for } 1 \leq j \leq N.$$

3. Find any two numbers $(\hat{C}_{1n}^\Gamma, \hat{C}_{2n}^\Sigma)$ such that

$$\frac{1}{N} \sum_{j=1}^N \mathbb{1} \left\{ \|S_{n,j}^*(\mathbf{I})\|_\infty \leq \hat{C}_{1n}^\Gamma, \|S_{n,j}^*(\mathbf{II})\|_\infty \leq \hat{C}_{2n}^\Sigma \right\} \geq 1 - \alpha.$$

Here $\mathbb{1}\{A\}$ is the indicator function of a set A .

The following theorem proves the validity of multiplier bootstrap under assumption (4.21) of Lemma 5. Note that we only prove asymptotic conservativeness instead of consistency which does not hold. See Remark 4.D.1 in Appendix 4.D. This inconsistency can be easily understood by noting that $\mathbb{E}[W_i]$ is replaced by the average \bar{W}_n which is not a consistent estimator. Define

$$L_{n,p} := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [|W_i(j) - \mathbb{E}[W_i(j)]|^3].$$

Theorem 6. Suppose $(X_i^\top, Y_i)^\top, 1 \leq i \leq n$ are independent random variables satisfying

$$\min_{1 \leq i \leq n} \min_{1 \leq j \leq q} \text{Var}(W_i) \geq B > 0,$$

and

$$\max_{1 \leq i \leq n} \max \left\{ \max_{1 \leq j \leq p} \|X_i(j)\|_{\psi_\alpha}, \|Y_i\|_{\psi_\alpha} \right\} \leq K_{n,p}. \quad (4.23)$$

If $n, p \geq 1$ are such that

$$\max \left\{ L_{n,p} K_{n,p}^6 (\log p)^{1+6/\alpha}, L_{n,p}^2 \log^7 p, K_{n,q}^{12} \log p \right\} = o(n),$$

then the multiplier bootstrap described above provides a conservative inference in the sense that

$$\liminf_{n \rightarrow \infty} \inf_{t_1, t_2 \geq 0} \left(\mathbb{P}(\mathcal{D}_{1n}^\Gamma \leq t_1, \mathcal{D}_{2n}^\Sigma \leq t_2) - \mathbb{P}(\|S_{n,j}^{eW}(\mathbf{I})\|_\infty \leq t_1, \|S_{n,j}^{eW}(\mathbf{II})\|_\infty \leq t_2 | \mathcal{Z}_n) \right) \geq 0,$$

where $\mathcal{Z}_n := \{(X_i^\top, Y_i)^\top : 1 \leq i \leq n\}$.

Proof. Theorems 10 and 11 (stated in Appendix 4.D) apply in the setting above since under assumption (4.23),

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq q} \|W_i(j)\|_{\psi_{\alpha/2}} \leq \max_{1 \leq i \leq n} \max \left\{ \max_{1 \leq j \leq p} \|X_i(j)\|_{\psi_\alpha}, \|Y_i\|_{\psi_\alpha} \right\}^2 \leq K_{n,p}^2.$$

And the rate restriction on n and p ensure that the bounds in Theorem 10 and 11 both converge to zero. \square

By Theorem 6, the estimates $(\hat{C}_{1n}^\Gamma, \hat{C}_{2n}^\Sigma)$ are consistent for some quantities that can replace the quantiles $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ of $(\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma)$ in (4.13).

4.5 A Generalization for Linear Regression-type Problems

A simple generalization of Theorems 4 and 5 as stated in Theorem 7 allows valid post-selection inference in linear regression-type problems. The importance of this generalization can be seen from Remark 4.5.1 and the discussion in Section 4.7. To describe this generalization, consider the following setting. Let $\hat{\Sigma}_n^*, \Sigma_n^*$ be two p -dimensional matrices and $\hat{\Gamma}_n^*, \Gamma_n^*$ be two p -dimensional vectors. Consider the error norms

$$\mathcal{D}_{1n}^{\Gamma^*} := \left\| \hat{\Gamma}_n^* - \Gamma_n^* \right\|_{\infty} \quad \text{and} \quad \mathcal{D}_{2n}^{\Sigma^*} := \left\| \hat{\Sigma}_n^* - \Sigma_n^* \right\|_{\infty}.$$

Define for every $M \in \mathcal{M}(p)$, the estimator and the corresponding target as

$$\begin{aligned} \hat{\xi}_{n,M} &:= \arg \min_{\theta \in \mathbb{R}^{|M|}} \left\{ \theta^{\top} \hat{\Sigma}_n^*(M) \theta - 2\theta^{\top} \hat{\Gamma}_n^*(M) \right\}, \\ \xi_{n,M} &:= \arg \min_{\theta \in \mathcal{R}^{|M|}} \left\{ \theta^{\top} \Sigma_n^*(M) \theta - 2\theta^{\top} \Gamma_n^*(M) \right\}. \end{aligned}$$

Consider for any $M \in \mathcal{M}(p)$, the confidence regions $\hat{\mathcal{R}}_{n,M}^*$ and $\hat{\mathcal{R}}_{n,M}^{\dagger}$, analogues to those before, as

$$\begin{aligned} \hat{\mathcal{R}}_{n,M}^* &:= \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \hat{\Sigma}_n^*(M) \left(\hat{\xi}_{n,M} - \theta \right) \right\|_{\infty} \leq C_{1n}^{\Gamma^*}(\alpha) + C_{2n}^{\Sigma^*}(\alpha) \|\theta\|_1 \right\}, \\ \hat{\mathcal{R}}_{n,M}^{\dagger} &:= \left\{ \theta \in \mathbb{R}^{|M|} : \left\| \hat{\Sigma}_n^*(M) \left(\hat{\xi}_{n,M} - \theta \right) \right\|_{\infty} \leq C_{1n}^{\Gamma^*}(\alpha) + C_{2n}^{\Sigma^*}(\alpha) \left\| \hat{\xi}_{n,M} \right\|_1 \right\}. \end{aligned}$$

where $C_{1n}^{\Gamma^*}(\alpha)$ and $C_{2n}^{\Sigma^*}(\alpha)$ are constants (or joint quantiles) that satisfy,

$$\mathbb{P} \left(\mathcal{D}_{1n}^{\Gamma^*} \leq C_{1n}^{\Gamma^*}(\alpha) \quad \text{and} \quad \mathcal{D}_{2n}^{\Sigma^*} \leq C_{2n}^{\Sigma^*}(\alpha) \right) \geq 1 - \alpha.$$

Finally, let $\Lambda_n^*(k) = \min\{\lambda_{\min}(\Sigma_n^*(M)) : M \in \mathcal{M}(k)\}$.

Theorem 7. *The set of confidence regions $\{\hat{\mathcal{R}}_{n,M}^* : M \in \mathcal{M}(p)\}$ satisfies*

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \xi_{0,M} \in \hat{\mathcal{R}}_{n,M}^* \right\} \right) \geq 1 - \alpha,$$

and if for any $1 \leq k \leq p$ that satisfies $k\mathcal{D}_{2n}^* = o_p(\Lambda^*(k)) = o_p(1)$,

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(k)} \left\{ \xi_{0,M} \in \hat{\mathcal{R}}_{n,M}^{*\dagger} \right\} \right) \geq 1 - \alpha.$$

Proof. The proof is exactly the same as for Theorems 4 and 5. The reader just has to realize that we did not use any structure of $\hat{\Sigma}_n, \hat{\Gamma}_n$ or that they are unbiased estimators of Σ_n, Γ_n respectively, in the proof there. \square

Remark 4.5.1 The result in Theorem 7 allows one to deal with the case of missing data or outliers in linear regression setting. In case of missing data or when the data is suspected of containing outliers, it might be more useful to use estimators of Σ_n and Γ_n that take this concern into account. For the case of missing data/errors-in-covariates/multiplicative noise, see [Loh and Wainwright \(2012, Examples 1, 2 and 3\)](#) and references therein for estimators other than $\hat{\Sigma}_n$ and $\hat{\Gamma}_n$. For the case of outliers either in the classical sense or in the adversarial corruption setting, see [Chen et al. \(2013\)](#). For correct usage of this theorem, it is crucial that the sub-matrix and sub-vector of Σ_n^* and Γ_n^* , respectively are used for sub-models. For example, if we use full covariate imputation in case of missing data, then the sub-model estimator should be based on a sub-matrix of this full covariate imputation. \diamond

4.6 Connection to High-dimensional Regression

The confidence regions $\widehat{\mathcal{R}}_{n,M}$ and $\widehat{\mathcal{R}}_{n,M}^\dagger$ have a very close connection to the well-known estimator in the high-dimensional linear regression literature called the Dantzig Selector proposed by [Candes and Tao \(2007\)](#) and the closely related ones by [Rosenbaum and Tsybakov \(2010\)](#) and [Chen et al. \(2013\)](#). These papers or methods are not related to post-selection inference and were proposed under a linear model assumption. The Dantzig selector estimates $\beta_0 \in \mathbb{R}^p$, using observations $(X_i^\top, Y_i), 1 \leq i \leq n$ that satisfy $Y_i = X_i^\top \beta_0 + \varepsilon_i$ for independent and identically distributed errors ε_i with a mean zero normal distribution. [Candes and Tao \(2007\)](#) as many others assumed fixed covariates $X_i, 1 \leq i \leq n$. In our notation, the Dantzig selector is defined by the optimization problem

$$\text{minimize } \|\beta\|_1 \quad \text{subject to } \|\Gamma_n - \Sigma_n \beta\|_\infty \leq \lambda_n,$$

for some tuning parameter λ_n that converges to zero as n increases. To relate this to our confidence regions $\widehat{\mathcal{R}}_{n,M}^\dagger$ (in [\(4.11\)](#)), note that for $\beta = \beta_0$ in the constraint set, the quantity inside the norm is $\Sigma_n(\widehat{\beta} - \beta_0)$ where $\widehat{\beta}$ is any least squares estimator. See also Equation [\(3.15\)](#) in Chapter 3. The estimator defined in [Chen et al. \(2013\)](#) resembles

$$\text{minimize } \|\beta\|_1 \quad \text{subject to } \|\Gamma_n - \Sigma_n \beta\|_\infty \leq \lambda_n + \delta_n \|\beta\|_1,$$

for some tuning parameters λ_n and δ_n both converging to zero as n increases. This constraint set corresponds to our confidence regions $\widehat{\mathcal{R}}_{n,M}$ in [Theorem 4](#).

The following theorem proves that there exist valid post-selection confidence regions that resemble the objective functions of lasso ([Tibshirani \(1996\)](#)) and sqrt-lasso ([Belloni et al. \(2011\)](#)). The proof is deferred to [Appendix 4.C](#). These relations to

high-dimensional linear regression literature poses the interesting question: “is there a more deeper connection between post-selection inference and high-dimensional estimation?”. Other than the results in linear regression, we do not yet have an answer to this interesting question.

Define for every $M \in \mathcal{M}(p)$, the confidence regions

$$\check{\mathcal{R}}_M := \{\theta \in \mathbb{R}^{|M|} :$$

$$\hat{R}_n(\theta; M) \leq \hat{R}_n(\hat{\beta}_{n,M}; M) + 2C_{1n}^\Gamma(\alpha) \left[\|\hat{\beta}_{n,M}\|_1 + \|\theta\|_1 \right] + C_{2n}^\Sigma(\alpha) \left[\|\hat{\beta}_{n,M}\|_1^2 + \|\theta\|_1^2 \right] \},$$

$$\check{\mathcal{R}}_M^\dagger := \left\{ \theta \in \mathbb{R}^{|M|} : \hat{R}_n(\theta; M) \leq \hat{R}_n(\hat{\beta}_{n,M}; M) + 4C_{1n}^\Gamma(\alpha) \|\hat{\beta}_M\|_1 + 2C_{2n}^\Sigma(\alpha) \|\hat{\beta}_M\|_1^2 \right\},$$

$$\check{\mathcal{R}}_M := \{\theta \in \mathbb{R}^{|M|} :$$

$$\hat{R}_n^{1/2}(\theta; M) \leq \hat{R}_n^{1/2}(\hat{\beta}_{n,M}; M) + C_n^{1/2}(\alpha) (1 + \|\theta\|_1) + C_n^{1/2}(\alpha) \left(1 + \|\hat{\beta}_M\|_1 \right) \},$$

$$\check{\mathcal{R}}_M^\dagger := \left\{ \theta \in \mathbb{R}^{|M|} : \hat{R}_n^{1/2}(\theta; M) \leq \hat{R}_n^{1/2}(\hat{\beta}_{n,M}; M) + 2C_n^{1/2}(\alpha) \left(1 + \|\hat{\beta}_M\|_1 \right) \right\},$$

where $\hat{R}_n(\cdot; M)$ is the empirical least squares objective function defined in Equation (4.2) and $C_n(\alpha)$ is the $(1 - \alpha)$ -upper quantile of $\max\{\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma\}$.

Theorem 8. *For any $n \geq 1, p \geq 1$, the following simultaneous inference guarantee holds:*

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \beta_{n,M} \in \check{\mathcal{R}}_M \right\} \right) \geq 1 - \alpha, \quad (4.24)$$

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \beta_{n,M} \in \check{\mathcal{R}}_M^\dagger \right\} \right) \geq 1 - \alpha, \quad (4.25)$$

and for any $1 \leq k \leq p$ satisfying (A1)(k), we have

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \beta_{n,M} \in \check{\mathcal{R}}_M^\dagger \right\} \right) \geq 1 - \alpha, \quad (4.26)$$

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(p)} \left\{ \beta_{n,M} \in \check{\mathcal{R}}_M \right\} \right) \geq 1 - \alpha, \quad (4.27)$$

Remark 4.6.1 (Intersection of Confidence Regions) All our confidence regions are based on deterministic inequalities as mentioned before. This implies that the intersection of the confidence regions $\hat{\mathcal{R}}_{n,M}$, $\hat{\mathcal{R}}_{n,M}^\dagger$ and $\check{\mathcal{R}}_M$ provides a valid simultaneous and post-selection inference. That means, for any $1 \leq k \leq p$ such that (A1)(k) holds,

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(k)} \left\{ \hat{\mathcal{R}}_{n,M} \cap \hat{\mathcal{R}}_{n,M}^\dagger \cap \check{\mathcal{R}}_M \right\} \right) \geq 1 - \alpha. \quad (4.28)$$

To prove this, let $\hat{\mathcal{C}}_M$, $\hat{\mathcal{C}}_M^\dagger$ and $\check{\mathcal{C}}_M$ represent the confidence sets $\hat{\mathcal{R}}_{n,M}$, $\hat{\mathcal{R}}_{n,M}^\dagger$ and $\check{\mathcal{R}}_M$ with $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ replaced by $(\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma)$. From the proofs of Theorems 4, 5 and 8, it is clear that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{M \in \mathcal{M}(k)} \left\{ \hat{\mathcal{C}}_M \cap \hat{\mathcal{C}}_M^\dagger \cap \check{\mathcal{C}}_M \right\} \right) = 1.$$

And by the definition of $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$ (in (4.13)), the result (4.28) follows. Provably the intersection of confidence regions is smaller. By the same argument it is possible to include the confidence regions $\check{\mathcal{R}}_M^\dagger$, $\check{\mathcal{R}}_M$, and $\check{\mathcal{R}}_M^\dagger$ in the intersection. \diamond

Remark 4.6.2 (Usefulness of Lasso-based Regions) The confidence regions discussed in this section are given solely for the purpose of illustrating and making solid the connection between post-selection inference and high-dimensional linear regression. The shape of all these confidence regions is ellipsoid and have larger volume

than the confidence region $\widehat{\mathcal{R}}_M^\dagger$ in terms of the rate. This result is not presented here but is not difficult to prove. This rate comparison is only asymptotic and the intersection argument presented in Remark 4.6.1 might still be useful in finite samples.

◇

4.7 Pros and Cons of Approach 1

The confidence regions $\widehat{\mathcal{R}}_{n,M}$ and $\widehat{\mathcal{R}}_{n,M}^\dagger$ constitute what we call approach 1. Various advantages and disadvantages of this approach are discussed in this section. Some of these comments also apply to the confidence regions mentioned in Theorem 7.

The following are some of the advantages of this approach. The confidence regions are asymptotically valid for post-selection inference. This is the first work that provides valid post-selection inference in this generality. The confidence region for any model M depend only on the joint quantiles $C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha)$ and the least squares linear regression estimator corresponding to the model $M, \widehat{\beta}_{n,M}$. So, the computational complexity of these confidence regions is no more than a multiple of the computational complexity of $\widehat{\beta}_{n,M}$. Computation of $C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha)$ takes no more than a linear function of p operations, as shown in Section 4.4. This computational complexity is in sharp contrast to the valid post-selection inference method proposed by Berk et al. (2013) or Bachoc et al. (2016) which requires essentially solving for the least squares estimators of all the models for a confidence region with some model M . Therefore, implementation of their procedure is NP-hard, in general. The Lebesgue measure of the confidence regions $\widehat{\mathcal{R}}_{n,M}^\dagger$ converges to zero at a rate that is the minimax rate in high-dimensional linear regression literature. So, we suspect this might be the optimal rate here too but at present we do not have a proof or even an optimality framework. Note that the volume of the confidence region for model M is computed with respect to the Lebesgue on $\mathbb{R}^{|M|}$.

There is one more advantage which might not seem like one at first glance. The confidence region for $\beta_{n,M}$ for a particular model does not require information on how many models are being used for model selection. The volume of the confidence region for $\beta_{n,M}$ depends only on the features of the model M except for the quantiles. This implies that the confidence regions $\hat{\mathcal{R}}_{n,M}^\dagger, M \in \mathcal{M}(k)$ can often have much smaller volumes than the ones produced using the approach of [Berk et al. \(2013\)](#).

There are some disadvantages and some irking factors associated with this approach. Firstly, notice that the confidence regions are not invariant under linear transformations of the observations as briefed in Remark [4.3.5](#). Most methods in high-dimensional linear regression procedures that induce sparsity also share this feature. Even from a naive point of view, invariance under change of units for all variables involved is crucial for interpretation. This translates to invariance under diagonal linear transformations of the observations. Normalizing all the variables involved to have a unit standard deviation is a commonly suggested method to attain invariance under diagonal transformations. Formally, this means one should use

$$X_i^* = \left(\frac{X_i(1) - \bar{X}(1)}{s_n(1)}, \dots, \frac{X_i(p) - \bar{X}_i(p)}{s_n(p)} \right), \quad Y_i^* = \frac{Y_i - \bar{Y}}{s_n(0)},$$

in place of $(X_i, Y_i), 1 \leq i \leq n$, where for $1 \leq j \leq p$,

$$\bar{X}(j) = \frac{1}{n} \sum_{i=1}^n X_i(j), \quad \text{and} \quad s_n^2(j) = \frac{1}{n} \sum_{i=1}^n [X_i(j) - \bar{X}(j)]^2,$$

and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \text{and} \quad s_n^2(0) = \frac{1}{n} \sum_{i=1}^n [Y_i - \bar{Y}]^2.$$

This leads to the matrix and vector,

$$\Sigma_n^* = \frac{1}{n} \sum_{i=1}^n X_i^* X_i^{*\top}, \quad \text{and} \quad \Gamma_n^* = \frac{1}{n} \sum_{i=1}^n X_i^* Y_i^*.$$

Note that the observations (X_i^*, Y_i^*) , $1 \leq i \leq n$ are not independent even if we start with independent observations (X_i, Y_i) . This is one of the reasons why we did not assume independence for Theorems 4, 5 and 7. Of course one needs to prove the rates for the error norms $\mathcal{D}_{1n}^{\Gamma^*}$ and $\mathcal{D}_{2n}^{\Sigma^*}$ in this case for an application of these results. We leave it to the reader to verify that the rates are exactly the same obtained in Lemma 5 (one needs to use a Slutsky-type argument). See [Cui et al. \(2016\)](#) for a similar derivation. We conjecture that much weaker conditions than listed in Lemma 5 are enough for those same rates, in particular, exponential moments are not required. See [van de Geer and Muro \(2014, Theorem 5.3\)](#) for a result in this direction. Getting back to invariance under arbitrary linear transformations, we do not know if it is possible come up with a procedure that retains the computational complexity of approach 1 while satisfying this invariance. We conjecture that this is not possible and that there is a strict trade-off between computational efficiency and affine invariance. On this point, the general approach developed in the next chapter obeys this affine invariance but are generally NP-hard in computation.

Another disadvantage of approach 1 is that it is mostly based on deterministic inequalities. As the reader may have suspected, this might lead to some conservativeness of the method. Note that non-identical distributions of the observations already introduces some conservativeness. The confidence regions $\hat{\mathcal{R}}_{n,M}$ and $\hat{\mathcal{R}}_{n,M}^\dagger$ cover $\beta_{n,M}$ with probability (at least) $1 - \alpha$ asymptotically. In particular, these confidence regions provide valid post-selection inference for the full vector $\beta_{n,M}$ instead of each of the coordinates of $\beta_{n,M}$. The region $\hat{\mathcal{R}}_{n,M}^\dagger$ is defined by a system of linear inequalities

and hence the local inference (or inference on coordinates) for $\beta_{n,M}(j), 1 \leq j \leq |M|$ can be obtained by solving a linear program. However, these can be very conservative for local inference guarantees.

We emphasize before ending this section that the main focus of approach 1 is validity and better computational complexity not optimality. However, optimality holds for our confidence regions as mentioned in Remark 4.3.6 for fixed covariates. It should be understood that without validity there is no point in proving any kind of optimality properties about the size of confidence region. Validity and optimality are the focus of the unified approach presented in Chapter 5.

4.8 Numerical examples

In this section, we demonstrate some properties of PoSI and compare with other confidence intervals with post-selection guarantee by Berk et al. (2013), Bachoc et al. (2016), Tibshirani et al. (2016). We consider the following data generating model for numerical examples.

$$Y_i = X_i^\top \beta_0 + \varepsilon_i, \quad 1 \leq i \leq n \quad \text{with} \quad \beta_0 = \mathbf{0}_p, \quad \text{and} \quad \varepsilon_i \stackrel{iid}{\sim} N(0, 1). \quad (4.29)$$

The following three settings of covariates will be considered:

1. **Setting A (orthogonal design):** X_i are chosen so that $\hat{\Sigma} = n^{-1} \sum_{i=1}^n X_i X_i^\top = I_p$, the identity matrix in p dimensions. The data is generated by starting with a random matrix with iid Gaussian entries and applying Gram-Schmidt to satisfy $\hat{\Sigma} = I_p$.
2. **Setting B (exchangeable design):** X_i are such that $\hat{\Sigma} = I_p + \alpha \mathbf{1}_p \mathbf{1}_p^\top$ with $\alpha = -1/(p+2)$, which is close to the degenerate case attained for $\alpha = -1/p$. The data is first generated as in Setting A and then multiplied by $\hat{\Sigma}^{1/2}$.

3. **Setting C (worst-case design):** X_i are such that

$$\hat{\Sigma} := \begin{bmatrix} I_{p-1} & c\mathbf{1}_{p-1} \\ c\mathbf{1}_{p-1}^\top & 1 \end{bmatrix}, \text{ where } c^2 = \frac{1}{2(p-1)}.$$

Settings A and B lead to the best rate for the “max- $|t|$ ” approach, while Setting C leads to the worst rate. See Berk et al. (2013, Sections 6.1 and 6.2) for results in these three settings.

Figure 4.1 shows the comparison of Approach 1 with Berk style PoSI and also selective inference confidence intervals. We also present the projection of Approach 1 regions to a rectangle; rectangular confidence regions are most interpretable and of course, they are bigger than the original approach 1 region. This simulation is taken from Kuchibhotla et al. (2020, Section 9) and we refer the reader to this paper for more details. This simulation is reproducible through <https://github.com/post-selection-inference/R>.

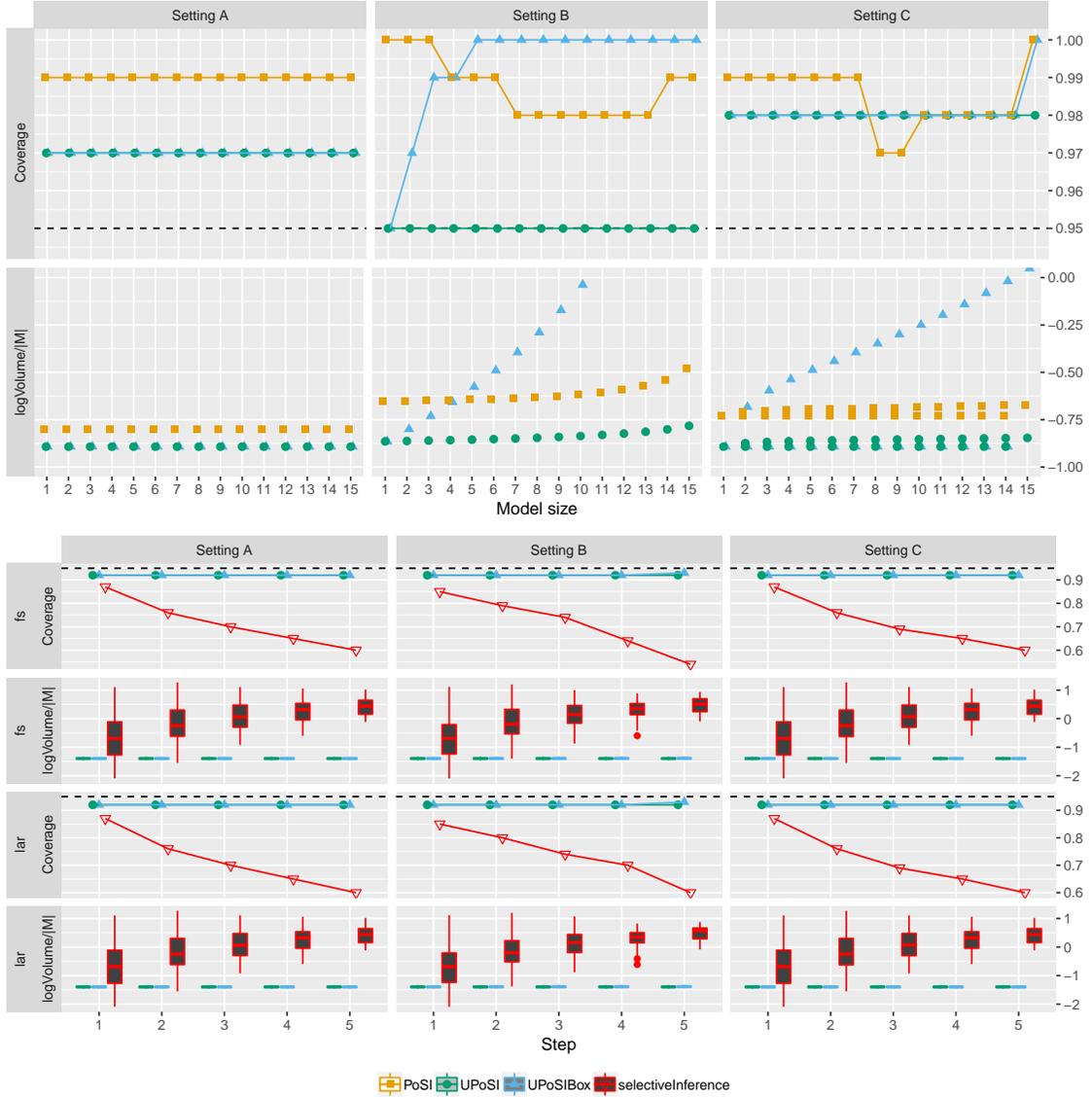


Figure 4.1: Comparison of “UPoSI” with “PoSI” (Berk et al., 2013) and “selective Inference” (Tibshirani et al., 2016). Methods included are the “UPoSI” confidence regions $\hat{\mathcal{R}}_{n,M}^\dagger$ (4.12) and the projected Box regions: “UPoSIBox” regions. The first two plots provide comparisons with the “PoSI” regions (5.16) of Berk et al. (2013). The next four plots show comparisons with “selective Inference.” Rather than providing overall simultaneous coverage, we show simultaneous coverage for different model sizes separately: $1 \leq |M| \leq 15$ for comparison with “PoSI” and $1 \leq |M| \leq 5$ for comparison with “selective Inference.” Because the volume of a region in $|M|$ dimensions scales like $C^{|M|}$ for some constant C , we plot $\log(\text{Leb}(\hat{\mathcal{R}}_{n,M}^\dagger))/|M|$, which allows comparison across different model sizes. Recall that in Setting C models fall into two groups: those that contain the last covariate, and those that don’t. This is the reason for showing two dots for each model size in Setting C. The size of dots indicates the proportion of models in each group. The dashed lines in the coverage plots show the nominal confidence level 0.95.

Figure 4.2 shows the comparison of Approach 1 with sample splitting, the simplest approach for post-selection inference.

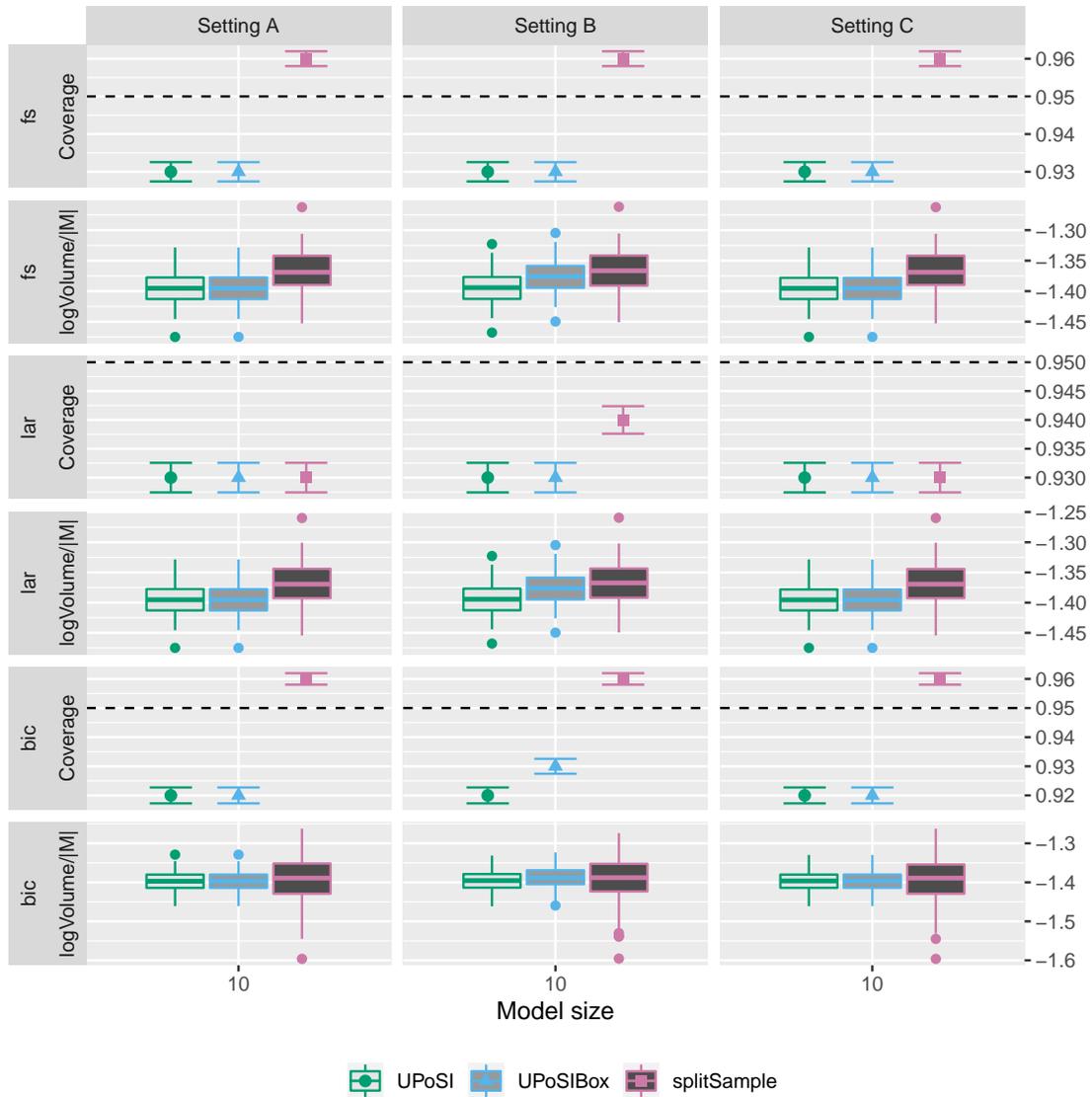


Figure 4.2: Comparison of coverage and volume of UPoSI with sample splitting. In all cases the volume of our confidence regions are at least as good as sample splitting. The latter is slightly more conservative in coverage in some cases, but not dramatically so.

APPENDIX

4.A Proof of Lemma 4

Fix $M \in \mathcal{M}(k)$ with $k\mathcal{D}_{2n}^\Sigma \leq \Lambda_n(k)$. Observe that the least squares estimator satisfies

$$\hat{\beta}_{n,M} - \beta_{n,M} = (\Sigma_n(M))^{-1} \left(\left[\hat{\Gamma}_n(M) - \Gamma_n(M) \right] - \left[\hat{\Sigma}_n(M) - \Sigma_n(M) \right] \beta_{n,M} \right),$$

and for all $M \in \mathcal{M}(k)$,

$$\begin{aligned} \left\| \hat{\Sigma}_n(M) - \Sigma_n(M) \right\|_{op} &\leq \sup_{\substack{\|\theta\|_0 \leq k, \\ \|\theta\|_2 \leq 1}} \left| \theta^\top (\hat{\Sigma}_n - \Sigma_n) \theta \right| \\ &\leq k \left\| \hat{\Sigma}_n - \Sigma_n \right\|_\infty = k\mathcal{D}_{2n}^\Sigma. \end{aligned} \quad (4.30)$$

Thus, for all $M \in \mathcal{M}(k)$,

$$\Lambda_n(k) - k\mathcal{D}_{2n}^\Sigma \leq \|\Sigma_n(M)\|_{op} - k\mathcal{D}_{2n}^\Sigma \leq \left\| \hat{\Sigma}_n(M) \right\|_{op} \leq \|\Sigma_n(M)\|_{op} + k\mathcal{D}_{2n}^\Sigma.$$

Hence, for k satisfying $k\mathcal{D}_{2n}^\Sigma \leq \Lambda_n(k)$,

$$\begin{aligned} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2 &\leq \frac{\left\| \hat{\Gamma}_n(M) - \Gamma_n(M) \right\|_2 + \left\| [\hat{\Sigma}_n(M) - \Sigma_n(M)] \beta_{n,M} \right\|_2}{\Lambda_n(k) - k\mathcal{D}_{2n}^\Sigma} \\ &\leq |M|^{1/2} \frac{\left\| \hat{\Gamma}_n(M) - \Gamma_n(M) \right\|_\infty + \left\| [\hat{\Sigma}_n(M) - \Sigma_n(M)] \beta_{n,M} \right\|_\infty}{\Lambda_n(k) - k\mathcal{D}_{2n}^\Sigma} \\ &\leq \frac{|M|^{1/2} (\mathcal{D}_{1n}^\Gamma + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1)}{\Lambda_n(k) - k\mathcal{D}_{2n}^\Sigma}. \end{aligned}$$

Now applying

$$\left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_1 \leq |M|^{1/2} \left\| \hat{\beta}_{n,M} - \beta_{n,M} \right\|_2,$$

the result follows.

4.B Proof of Lemma 6

For any fixed model M , the Lebesgue measure of the confidence region is given by

$$\mathbf{Leb}(\hat{\mathcal{R}}_M^\dagger) = |\Sigma_n(M)|^{-1} \left(C_{1n}^\Gamma(\alpha) + C_{2n}^\Sigma(\alpha) \left\| \hat{\beta}_M \right\|_1 \right)^{|M|}, \quad (4.31)$$

which converges to zero as n tends to infinity. Here for any matrix $A \in \mathbb{R}^{p \times p}$, $|A|$ denotes the determinant of A . This equality follows since the confidence region $\hat{\mathcal{R}}_M^\dagger$ can be written as

$$\hat{\mathcal{R}}_M^\dagger = \left\{ [\Sigma_n(M)]^{-1} (\theta + \hat{\beta}_M) : \|\theta\|_\infty \leq \left(C_{1n}^\Gamma(\alpha) + C_{2n}^\Sigma(\alpha) \left\| \hat{\beta}_{n,M} \right\|_1 \right) \right\}.$$

By inequality (4.30), for all $M \in \mathcal{M}(k)$

$$|\Sigma_n(M)|^{-1} \leq (\Lambda(k) - k\mathcal{D}_{2n}^\Sigma)^{-|M|}.$$

We know that $C_{1n}^\Gamma(\alpha)$ and $C_{2n}^\Sigma(\alpha)$ converge to zero at a rate depending on the tails of the joint distribution of (X_i, Y_i) . The result now follows from equation (4.31) and uniform consistency of $\hat{\beta}_{n,M}$ in the $\|\cdot\|_1$ -norm as shown in Lemma 4 under (A1)(k).

To prove the second result, first note that from Lemma 5,

$$\max\{C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha)\} = O\left(\sqrt{\frac{\log p}{n}}\right),$$

since the second term in the expectation bound in Lemma 5 is of lower order than the first term under the assumption (4.22) of Lemma 6. The result is now proved if

we prove that for all $M \in \mathcal{M}(k)$,

$$\|\beta_{n,M}\|_1^2 \leq \frac{|M|}{\Lambda_n(k)} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right). \quad (4.32)$$

By definition of $\beta_{n,M}$ it follows that

$$0 \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] - \frac{1}{n} \sum_{i=1}^n \beta_{n,M}^\top \mathbb{E}[X_i(M)X_i^\top(M)] \beta_{n,M} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(Y_i - X_i^\top(M)\beta_{n,M})^2].$$

Therefore, by definition of $\Lambda_n(k)$,

$$\Lambda_n(k) \|\beta_{n,M}\|_2^2 \leq \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i^2] \right).$$

Now using the inequality $\|\beta_{n,M}\|_1 \leq \sqrt{|M|} \|\beta_{n,M}\|_2$, inequality (4.32) follows.

4.C Proof of Theorem 8

Only the proof of (4.24) and (4.26) is provided and the steps to prove (4.25) and (4.27) are sketched since the proof is similar.

It is easy to verify that for any $M \subseteq \mathcal{M}(p)$ and $\theta \in \mathbb{R}^{|M|}$

$$\left| \theta^\top \hat{\Sigma}_n(M)\theta - 2\theta^\top \hat{\Gamma}_n(M) - \theta^\top \Sigma_n(M)\theta + 2\theta^\top \Gamma_n(M) \right| \leq \|\theta\|_1^2 \mathcal{D}_{2n}^\Sigma + 2\|\theta\|_1 \mathcal{D}_{1n}^\Gamma. \quad (4.33)$$

Therefore, for every $M \in \mathcal{M}(p)$,

$$\begin{aligned}
& \beta_{n,M} \widehat{\Sigma}_n(M) \beta_{n,M} - 2\beta_{n,M}^\top \widehat{\Gamma}_n(M) \\
& \leq \beta_{n,M} \Sigma_n(M) \beta_{n,M} - 2\beta_{n,M}^\top \Gamma_n(M) + 2\mathcal{D}_{1n}^\Gamma \|\beta_{n,M}\|_1 + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1^2 \\
& \leq \widehat{\beta}_{n,M} \Sigma_n(M) \widehat{\beta}_{n,M} - 2\widehat{\beta}_{n,M}^\top \Gamma_n(M) + 2\mathcal{D}_{1n}^\Gamma \|\beta_{n,M}\|_1 + \mathcal{D}_{2n}^\Sigma \|\beta_{n,M}\|_1^2 \\
& \leq \widehat{\beta}_{n,M} \widehat{\Sigma}_n(M) \widehat{\beta}_{n,M} - 2\widehat{\beta}_{n,M}^\top \Gamma_n(M) + 2\mathcal{D}_{1n}^\Gamma \left[\|\widehat{\beta}_{n,M}\|_1 + \|\beta_{n,M}\|_1 \right] \\
& \quad + \mathcal{D}_{2n}^\Sigma \left[\|\widehat{\beta}_{n,M}\|_1^2 + \|\beta_{n,M}\|_1^2 \right].
\end{aligned}$$

Here the first inequality follows from inequality (4.33) with $\theta = \beta_{n,M}$, the second inequality follows from the definition of $\beta_{n,M}$ (see Equation (4.10)) and the third inequality follows from inequality (4.33) with $\theta = \widehat{\beta}_{n,M}$. Adding the sample average of $\{Y_i^2 : 1 \leq i \leq n\}$ on both sides, we get for all $M \in \mathcal{M}(p)$,

$$\widehat{R}_n(\beta_{n,M}; M) \leq \widehat{R}_n(\widehat{\beta}_{n,M}; M) + 2\mathcal{D}_{1n}^\Gamma \left[\|\widehat{\beta}_{n,M}\|_1 + \|\beta_{n,M}\|_1 \right] + \mathcal{D}_{2n}^\Sigma \left[\|\widehat{\beta}_{n,M}\|_1^2 + \|\beta_{n,M}\|_1^2 \right]. \quad (4.34)$$

This implies the first result (4.24). To prove the second result (4.26), note that

$$\begin{aligned}
& \left| \frac{\mathcal{D}_{1n}^2 + 2\mathcal{D}_{1n}\mathcal{D}_{2n} \|\widehat{\beta}_M\|_1 + \mathcal{D}_{2n}^2 \|\widehat{\beta}_M\|_1^2}{\mathcal{D}_{1n}^2 + 2\mathcal{D}_{1n}\mathcal{D}_{2n} \|\beta_{0,M}\|_1 + \mathcal{D}_{2n}^2 \|\beta_{0,M}\|_1^2} - 1 \right| \\
& = \left| \left(\frac{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\widehat{\beta}_M\|_1}{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\beta_{0,M}\|_1} \right)^2 - 1 \right| \\
& \leq \left| \frac{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\widehat{\beta}_M\|_1}{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\beta_{0,M}\|_1} - 1 \right|^2 + 2 \left| \frac{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\widehat{\beta}_M\|_1}{\mathcal{D}_{1n} + \mathcal{D}_{2n} \|\beta_{0,M}\|_1} - 1 \right|,
\end{aligned}$$

which converges to zero under assumption (A1)(k), following the proof of Theorem

5. This implies that the error

$$\left[2\mathcal{D}_{1n} \left\| \hat{\beta}_{n,M} \right\|_1 + \mathcal{D}_{2n} \left\| \hat{\beta}_{n,M} \right\|_1^2 \right] - \left[2\mathcal{D}_{1n} \|\beta_{n,M}\|_1 + \mathcal{D}_{2n} \|\beta_{n,M}\|_1^2 \right],$$

is of smaller order than each of the terms uniformly in $M \in \mathcal{M}(k)$. The second result (4.26) then follows trivially by substituting the estimated parameters for the targets in inequality (4.34) and using the definition of $(C_{1n}^\Gamma(\alpha), C_{2n}^\Sigma(\alpha))$.

To prove the results with square-root lasso based regions, note that from inequality (4.34)

$$\begin{aligned} \hat{R}_n^{1/2}(\beta_{n,M}; M) &\leq \hat{R}_n^{1/2}(\hat{\beta}_{n,M}; M) + \max\{\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma\}^{1/2} \left(1 + \left\| \hat{\beta}_{n,M} \right\|_1 \right) \\ &\quad + \max\{\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma\}^{1/2} \left(1 + \|\beta_{n,M}\|_1 \right). \end{aligned}$$

4.D High-dimensional CLT and Bootstrap Consistency

Suppose $W_i, 1 \leq i \leq n$ are independent random vectors in \mathbb{R}^q with mean zero and finite second moment. Let $G_i, 1 \leq i \leq n$ be independent Gaussian random vectors in \mathbb{R}^q with mean zero satisfying

$$\mathbb{E}[G_i G_i^\top] = \mathbb{E}[W_i W_i^\top] \quad \text{for all } 1 \leq i \leq n.$$

Set

$$S_n^W := \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \quad \text{and} \quad S_n^G := \frac{1}{\sqrt{n}} \sum_{i=1}^n G_i.$$

Let \mathcal{A}^r denote the set of all rectangles in \mathbb{R}^q , that is, \mathcal{A}^r consists of all sets A of the form

$$A = \{z \in \mathbb{R}^q : a(j) \leq z(j) \leq b(j) \text{ for all } 1 \leq j \leq q\},$$

for some vectors $a, b \in \mathbb{R}^q$. Define

$$L_{n,q} := \max_{1 \leq j \leq q} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [|W_i(j)|^3], \quad (4.35)$$

and for $\phi \geq 1$, set

$$M_{n,W}(\phi) := \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\max_{1 \leq j \leq q} |W_i(j)|^3 \mathbb{1} \left\{ \max_{1 \leq j \leq q} |W_i(j)| \geq \sqrt{n}/(4\phi \log q) \right\} \right]. \quad (4.36)$$

Similarly, define $M_{n,G}(\phi)$ with $W_i(j)$'s replaced by $G_i(j)$'s in (4.36) and let

$$M_n(\phi) := M_{n,W}(\phi) + M_{n,G}(\phi).$$

Finally, set for any class \mathcal{A} of (Borel) sets in \mathbb{R}^q ,

$$\rho_n(\mathcal{A}) := \sup_{A \in \mathcal{A}} \left| \mathbb{P}(S_n^W \in A) - \mathbb{P}(S_n^G \in A) \right|.$$

To proceed further, we now present Theorem 2.1 of [Chernozhukov et al. \(2017a\)](#).

Theorem 9 (Theorem 2.1 of [Chernozhukov et al. \(2017a\)](#)). *Suppose that there exists some constant $B > 0$ such that*

$$\min_{1 \leq i \leq n} \min_{1 \leq j \leq q} \mathbb{E} [|W_i(j)|^2] \geq B.$$

Then there exist constants $K_1, K_2 > 0$ depending only on B such that for every

constant $L \geq L_{n,q}$,

$$\rho_n(\mathcal{A}^r) \leq K_1 \left[\left(\frac{L^2 \log^7 q}{n} \right)^{1/6} + \frac{M_n(\phi_n)}{L} \right], \quad (4.37)$$

with

$$\phi_n := K_2 \left(\frac{L^2 \log^4 q}{n} \right)^{-1/6}.$$

Before deriving the exact rate under the assumption (4.21) of Lemma 5, we prove that a bound of the form (4.37) proves a central limit theorem for $(\mathcal{D}_{1n}^\Gamma, \mathcal{D}_{2n}^\Sigma)$. Observe that for any $t_1, t_2 \in \mathbb{R}^+ \cup \{0\}$,

$$\begin{aligned} & \{ \mathcal{D}_{1n}^\Gamma \leq t_1, \mathcal{D}_{2n}^\Sigma \leq t_2 \} \\ &= \left\{ -t_1 \leq \frac{1}{n} \sum_{i=1}^n \{X_i(j)Y_i - \mathbb{E}[X_i(j)Y_i]\} \leq t_1 \text{ for all } 1 \leq j \leq p \right\} \cap \\ & \quad \left\{ -t_2 \leq \frac{1}{n} \sum_{i=1}^n \{X_i(l)X_i(m) - \mathbb{E}[X_i(l)X_i(m)]\} \leq t_2 \text{ for all } 1 \leq l \leq m \leq p \right\}. \end{aligned} \quad (4.38)$$

The right hand side here is a rectangle in terms of the vector S_n^W with mean zero vectors W_i containing

$$(\{X_i(j)Y_i - \mathbb{E}[X_i(j)Y_i]\}, 1 \leq j \leq p; \{X_i(l)X_i(m) - \mathbb{E}[X_i(l)X_i(m)]\}, 1 \leq l \leq m \leq p).$$

Note that W_i 's are vectors in \mathbb{R}^q with

$$q = 2p + \frac{p(p-1)}{2}.$$

Getting back to the central limit theorem under assumption (4.21), we need to bound $M_n(\phi)$. The following is a generalization (in terms of the tail assumption) of

Lemma C.1 of [Chernozhukov et al. \(2017a\)](#).

Lemma 7. *Let ξ be a nonnegative random variable such that for some constants $A, B > 0$ and $0 < \alpha < 3$,*

$$\mathbb{P}(\xi > x) \leq A \exp\left(-\frac{x^\alpha}{B^\alpha}\right) \quad \text{for all } x \geq 0.$$

Then for every $t \geq (6/\alpha)^{1/\alpha} B$,

$$\mathbb{E}[\xi^3 \mathbb{1}\{\xi \geq t\}] \leq \left(\frac{2+\alpha}{\alpha}\right) A t^3 \exp\left(-\frac{t^\alpha}{B^\alpha}\right).$$

Proof. Observe that

$$\begin{aligned} \mathbb{E}[\xi^3 \mathbb{1}\{\xi > t\}] &= 3 \int_0^t \mathbb{P}(\xi > t) x^2 dx + 3 \int_t^\infty \mathbb{P}(\xi > x) x^2 dx \\ &= \mathbb{P}(\xi > t) t^3 + 3 \int_t^\infty \mathbb{P}(\xi > x) x^2 dx. \end{aligned}$$

It is easy to derive that

$$\int_t^\infty x^2 \exp\left(-\frac{x^\alpha}{B^\alpha}\right) dx = \frac{B^3}{\alpha} \Gamma\left(\frac{3}{\alpha}, \frac{t^\alpha}{B^\alpha}\right),$$

for the upper incomplete gamma function

$$\Gamma(a, z) := \int_z^\infty \exp(-x) x^{a-1} dx.$$

Now using equation (1.5) of [Borwein and Chan \(2009\)](#), it follows that

$$\int_t^\infty x^2 \exp\left(-\frac{x^\alpha}{B^\alpha}\right) dx \leq \frac{2B^3}{\alpha} \left(\frac{t^\alpha}{B^\alpha}\right)^{3/\alpha-1} \exp(-t^\alpha/B^\alpha) \leq \frac{2t^3}{\alpha} \exp(-t^\alpha/B^\alpha),$$

if

$$t \geq 2^{1/\alpha}(3/\alpha - 1)^{1/\alpha}B.$$

Therefore,

$$\mathbb{E} \left[\xi^3 \mathbb{1}_{\{\xi \geq t\}} \right] \leq A \left[1 + \frac{6}{\alpha} \right] t^3 \exp \left(-\frac{t^\alpha}{B^\alpha} \right).$$

□

Using Theorem 9 and Lemma 7, we prove the following theorem under assumption (4.21). Recall the definition of $L_{n,q}$ from (4.35).

Theorem 10. *Suppose that the mean zero random vectors $W_i, 1 \leq i \leq n$ in \mathbb{R}^q satisfying for some $\beta, B > 0$,*

$$\min_{1 \leq j \leq p} \min_{1 \leq i \leq n} \mathbb{E} [W_i^2(j)] \geq B, \quad (4.39)$$

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq q} \|W_i(j)\|_{\psi_\beta} \leq K_{n,q}, \quad (4.40)$$

and if

$$\frac{1}{8K_2K_{n,q}} \left(\frac{n}{L_{n,q} \log q} \right)^{1/3} \geq 2^{1/\beta} \log^{1/\beta} q + (12/\beta)^{1/\beta} \quad (4.41)$$

then there exist constants K_1 , depending only on B and C_β depending only on B, β such that

$$\rho_n(\mathcal{A}^r) \leq K_1 \left(\frac{L_{n,q}^2 \log^7 q}{n} \right)^{1/6} + C_\beta \frac{K_{n,q}^6 \log q}{n}.$$

Proof. Set

$$\phi_n = K_2 \left(\frac{n}{L_{n,q}^2 \log^4 q} \right)^{1/6} \Rightarrow \frac{\sqrt{n}}{4\phi_n \log q} = \frac{n^{1/3} L_{n,q}^{1/3}}{4K_2 \log^{1/3} q} = \frac{1}{4K_2} \left(\frac{nL_{n,q}}{\log q} \right)^{1/3}.$$

Under assumption(4.40), it follows that

$$\mathbb{P} \left(\max_{1 \leq j \leq q} |W_i(j)| \geq 2^{1/\beta} K_{n,q} t^{1/\beta} + 2^{1/\beta} K_{n,q} \log^{1/\beta} q \right) \leq 2 \exp(-t). \quad (4.42)$$

Define

$$\Delta_i := 2^{-1/\beta} K_{n,p}^{-1} \left(\max_{1 \leq j \leq q} |W_i(j)| - 2^{1/\beta} K_{n,q} \log^{1/\beta} q \right)_+ \quad \text{with } (x)_+ = \max\{x, 0\}.$$

By an application of the tail bound (4.42), for all $1 \leq i \leq n$,

$$\mathbb{P}(\Delta_i \geq t) \leq 2 \exp(-t^\beta) \quad \text{for all } t \geq 0.$$

To bound $M_{n,W}(\phi_n)$, note that under assumption (4.41)

$$\begin{aligned} & 2^{-3/\beta} K_{n,p}^{-3} M_{n,W}(\phi_n) \\ & \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left(\log^{1/\beta} q + \Delta_i \right)^3 \mathbb{1} \left\{ \max_{1 \leq j \leq q} |W_i(j)| \geq \sqrt{n}/(4\phi_n \log q) \right\} \right], \\ & \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\left(\log^{1/\beta} q + \Delta_i \right)^3 \mathbb{1} \{ \Delta_i \geq n^{1/3} L_{n,q}^{1/3} / (2^{1/\beta} 8 K_{n,p} K_2 \log^{1/3} q) \} \right] \\ & \leq \left[8 \log^{3/\beta} q + \left(\frac{24 + 4\beta}{\beta} \right) \frac{n L_{n,q}}{2^{9+3/\beta} K_{n,q}^3 K_2^3 \log q} \right] \exp \left(- \frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q} \right) \\ & \leq \left(\frac{24 + 5\beta}{\beta} \right) \frac{n L_{n,q}}{2^{9+3/\beta} K_{n,q}^3 K_2^3 \log q} \exp \left(- \frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q} \right). \end{aligned}$$

Following the same argument for $M_{n,G}(\phi_n)$ and noting that G_i satisfy assumption (4.40) with $\beta = 2$, we obtain

$$2^{-3/2} K_{n,q}^{-3} M_{n,G}(\phi_n) \leq \frac{17n L_{n,q}}{2^{9+3/2} K_{n,q}^3 K_2^3 \log q} \exp \left(- \frac{n^{2/3} L_{n,q}^{2/3}}{2^7 K_{n,q}^2 K_2^2 \log^{2/3} q} \right).$$

Therefore, for all $n, q \geq 1$ satisfying (4.40),

$$M_n(\phi_n) \leq \left(\frac{24 + 5\beta}{2^9 \beta} \right) \frac{nL_{n,q}}{K_2^3 \log q} \exp \left(- \frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q} \right) \\ + \frac{17nL_{n,q}}{2^9 K_2^3 \log q} \exp \left(- \frac{n^{2/3} L_{n,q}^{2/3}}{2^7 K_{n,q}^2 K_2^2 \log^{2/3} q} \right),$$

and

$$\frac{M_n(\phi_n)}{L_{n,q}} \leq \left(\frac{24 + 5\beta}{2^9 \beta} \right) \frac{n}{K_2^3 \log q} \exp \left(- \frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q} \right) \\ + \frac{17n}{2^9 K_2^3 \log q} \exp \left(- \frac{n^{2/3} L_{n,q}^{2/3}}{2^7 K_{n,q}^2 K_2^2 \log^{2/3} q} \right),$$

Substituting these bounds in the bound (4.37), we get

$$\rho_n(\mathcal{A}^r) \leq K_1 \left(\frac{L_{n,q}^2 \log^7 q}{n} \right)^{1/6} + \left(\frac{24 + 5\beta}{2^9 \beta} \right) \frac{nK_1}{K_2^3 \log q} \exp \left(- \frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q} \right) \\ + \frac{17nK_1}{2^9 K_2^3 \log q} \exp \left(- \frac{n^{2/3} L_{n,q}^{2/3}}{2^7 K_{n,q}^2 K_2^2 \log^{2/3} q} \right).$$

By a direct calculation, it is easy to derive that for $\nu_1, \nu_2 > 0$,

$$x^{\nu_1} \exp(-x/\nu_2) \leq \nu_2^{\nu_1} \nu_1^{\nu_1} \exp(-\nu_1). \quad (4.43)$$

Using inequality (4.43) with

$$x = \left(\frac{nL_{n,q}}{K_{n,q}^3 K_2^3 \log q} \right)^{\beta/3}, \quad \nu_2 = 2^{1+3\beta} \quad \text{and} \quad \nu_1 = 6/\beta,$$

implies

$$\frac{nK_1}{K_2^3 \log q} \exp\left(-\frac{n^{\beta/3} L_{n,q}^{\beta/3}}{2^{1+3\beta} K_{n,q}^\beta K_2^\beta \log^{\beta/3} q}\right) \leq \frac{K_{n,q}^6 K_2^3 K_1 \log q}{nL_{n,q}^2} \left(\frac{12}{e\beta}\right)^{6/\beta} 2^{18}.$$

Substituting this bound along with its analogue for $\beta = 2$, we obtain

$$\begin{aligned} \rho_n(\mathcal{A}^r) &\leq K_1 \left(\frac{L_{n,q}^2 \log^7 q}{n}\right)^{1/6} + \left(\frac{24 + 5\beta}{\beta}\right) 2^9 \left(\frac{12}{e\beta}\right)^{6/\beta} \frac{K_{n,q}^6 K_2^3 K_1 \log q}{nL_{n,q}^2} \\ &\quad + \frac{2^9 17 K_{n,q}^6 K_2^3 K_1 \log q}{nL_{n,q}^2}. \end{aligned}$$

Under the lower bound assumption on the variance of $W_i(j)$, it follows that

$$L_{n,q} \geq B^{3/2}.$$

Thus, the result follows by replacing the constant for the last two terms by C_β (a constant depending only on β and B). \square

4.D.1 Bootstrap Consistency

In this sub-section, we consider the consistency of multiplier bootstrap based on Section 4.1 of [Chernozhukov et al. \(2017a\)](#). It is also possible to consider the empirical bootstrap in high-dimensions and prove its consistency based on the proof of Proposition 4.3 of [Chernozhukov et al. \(2017a\)](#). We do not prove it here as the proof techniques are the same.

Let e_1, e_2, \dots, e_n be a sequence of independent standard normal random variables independent of $\mathcal{W}_n := \{W_1, \dots, W_n\}$. Set

$$\bar{W}_n := \frac{1}{n} \sum_{i=1}^n W_i \in \mathbb{R}^q,$$

and consider the normalized sum

$$S_n^{eW} := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i (W_i - \bar{W}_n).$$

Note that

$$S_n^{eW} | \mathcal{W}_n \sim N \left(0, \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}_n) (W_i - \bar{W}_n)^\top \right) \in \mathbb{R}^q.$$

To prove consistency of multiplier bootstrap, we bound a quantity similar to $\rho_n(\mathcal{A}^{re})$, defined as

$$\rho_n^{\text{MB}}(\mathcal{A}^{re}) := \sup_{A \in \mathcal{A}^{re}} |\mathbb{P}(S_n^{eW} \in A | \mathcal{W}_n) - \mathbb{P}(S_n^G \in A)|.$$

Define

$$\begin{aligned} \Delta_{n,q} &:= \left\| \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}_n) (W_i - \bar{W}_n)^\top - \frac{1}{n} \sum_{i=1}^n \text{Var}(W_i) \right\|_\infty \\ &= \left\| \frac{1}{n} \sum_{i=1}^n \{W_i W_i^\top - \mathbb{E}[W_i W_i^\top]\} - \bar{W}_n \bar{W}_n^\top \right\|_\infty. \end{aligned}$$

Based on Theorem 4.1 and Remark 4.1 of [Chernozhukov et al. \(2017a\)](#), we prove the following theorem under assumption (4.40).

Theorem 11. *If $W_i, 1 \leq i \leq n$ are independent mean zero random vectors, then under assumptions (4.39) and (4.40),*

$$\mathbb{E} \left[\rho_n^{\text{MB}}(\mathcal{A}^{re}) \right] \leq C \log^{2/3} q \left[A_{n,q}^{1/3} \left(\frac{\log q}{n} \right)^{1/6} + K_{n,q}^{2/3} \frac{(\log q \log n)^{\frac{2}{3\beta}}}{n^{1/3}} \right],$$

for some constant C depending only on β, B . Here

$$A_{n,q} := \max_{1 \leq l \leq m \leq q} \frac{1}{n} \sum_{i=1}^n \text{Var}(W_i(l)W_i(m)).$$

Proof. As proved in Remark 4.1 of [Chernozhukov et al. \(2017a\)](#), we have

$$\rho_n^{\text{MB}}(\mathcal{A}^{re}) \leq C \Delta_{n,q}^{1/3} \log^{2/3} q.$$

So, to prove the result, all we need is to prove

$$\mathbb{E} [\Delta_{n,q}^{1/3}] \leq M_\beta \left[A_{n,q} \sqrt{\frac{\log q}{n}} + K_{n,q}^2 (\log q \log n)^{2/\beta} n^{-1} \right]^{1/3},$$

for some constant M_β . Observe that

$$\Delta_{n,q}^{1/3} \leq \left\| \frac{1}{n} \sum_{i=1}^n \{W_i W_i^\top - \mathbb{E}[W_i W_i^\top]\} \right\|_\infty^{1/3} + \|\bar{W}_n\|_\infty^{2/3}.$$

The bound on the expectation of the first term on the right hand side follows readily from Lemma 5. The bound on the expectation of the second term follows from Remark 4.2 of [Kuchibhotla and Chakraborty \(2018\)](#) and can be seen to be smaller order than the bound on the expectation of the first term. \square

Remark 4.D.1 (Inconsistency under unknown unequal means) Section 3.5 of Chapter 3 proved the inconsistency of bootstrap when the expectations are unknown/different (or more generally non-identically distributed). The same comment applies to the high-dimensional multiplier bootstrap. When $\mathbb{E}[W_i] = 0$ for all $1 \leq i \leq n$, then the second equality in the definition of $\Delta_{n,q}$ is true and converges to zero as $n \rightarrow \infty$. In general, if $\mathbb{E}[W_i] \neq 0$ but $\mathbb{E}[\bar{W}_n] = 0$, then the second equality of $\Delta_{n,q}$ should actually read

$$\Delta_{n,q} = \left\| \frac{1}{n} \sum_{i=1}^n \{W_i W_i^\top - \mathbb{E}[W_i W_i^\top]\} - \bar{W}_n \bar{W}_n^\top + \frac{1}{n} \sum_{i=1}^n \mathbb{E}[W_i] \mathbb{E}[W_i^\top] \right\|_\infty.$$

The difference

$$\left\| \bar{W}_n \bar{W}_n^\top - \frac{1}{n} \sum_{i=1}^n \mathbb{E} [W_i] \mathbb{E} [W_i^\top] \right\|_\infty,$$

does not converge to zero unless all the expectations are the same. So, $\Delta_{n,q}$ does not converge to zero. However,

$$\left\| \frac{1}{n} \sum_{i=1}^n (W_i - \bar{W}_n) (W_i - \bar{W}_n)^\top - \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[(W_i - \mathbb{E} [\bar{W}_n]) (W_i - \mathbb{E} [\bar{W}_n])^\top \right] \right\|_\infty,$$

does converge to zero as $n \rightarrow \infty$ and

$$\frac{1}{n} \sum_{i=1}^n \text{Var} (W_i) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[(W_i - \mathbb{E} [\bar{W}_n]) (W_i - \mathbb{E} [\bar{W}_n])^\top \right].$$

Hence again by Anderson's Lemma multiplier bootstrap provides an asymptotically conservative inference, in general. Observe that the sets in (4.38) are centrally convex symmetric sets and so, Anderson's Lemma applies. \diamond

4.E Bounds on $\|\widehat{\Omega}_n - \Omega_n\|_\infty$ under Dependence

In this section, we derive rate of convergence of $\|\widehat{\Omega}_n - \Omega_n\|_\infty$ under dependence. We first describe some classical notions of dependence that include well-known dependent processes as special cases. The description is essentially taken from ?. Let $\{\xi_t : t \in \mathbb{Z}\}$ be a stochastic process on some measure space. Let $\mathcal{F}_{m,n}$ (for $m < n$) be the σ -field generated by $\{\xi_i : m \leq i \leq n\}$ with possibility of $m = -\infty$ and $n = \infty$ included. Define

$$\alpha(j) := \sup_{k \in \mathcal{Z}} \sup \{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{F}_{-\infty,j}, B \in \mathcal{F}_{k+j,\infty} \},$$

$$\phi(j) := \sup_{k \in \mathcal{Z}} \sup \{ |\mathbb{P}(B|A) - \mathbb{P}(B)| : A \in \mathcal{F}_{-\infty,j}, B \in \mathcal{F}_{k+j,\infty}, \mathbb{P}(A) > 0 \}.$$

If $\alpha(j)$ (or correspondingly $\phi(j)$) converges to zero as j approaches infinity then the process $\{\xi_t : t \in \mathbb{Z}\}$ is called α -mixing (or correspondingly ϕ -mixing). It is clearly seen that every ϕ -mixing process is α -mixing since for any event A with $\mathbb{P}(A) > 0$,

$$|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \leq \mathbb{P}(A)|\mathbb{P}(B|A) - \mathbb{P}(B)|.$$

A process $\{\xi_t : t \in \mathbb{Z}\}$ is said to be m -dependent if $\alpha(j) = 0$ for all $j \geq m$. Evidently, m -dependent processes for any m are ϕ -mixing and so α -mixing too. One very useful feature of α -mixing processes is that measurable functions of finitely many elements of the process themselves α -mixing.

The dependence notion used in this section and the thesis is the one called functional dependence introduced by [Wu \(2005\)](#). It is possible to derive the results under the classical dependence notions like α -, ρ - mixing too, however, verifying the mixing assumptions can often be hard and many well-known processes do not satisfy them. See [Wu \(2005\)](#) for more details. It has also been shown that many econometric time series can be studied under the notion of functional dependence; see [Wu and Mielniczuk \(2010\)](#), [Liu et al. \(2013\)](#) and [Wu and Wu \(2016\)](#).

The dependence notion of [Wu \(2005\)](#) is written in terms of an input-output process that is easy to analyze in many settings. The process is defined as follows. Let $\{\varepsilon_i, \varepsilon'_i : i \in \mathbb{Z}\}$ denote a sequence of independent and identically distributed random variables on some measurable space $(\mathcal{E}, \mathcal{B})$. Let the q -dimensional (stochastic) process W_i has a causal representation as

$$W_i = G_i(\dots, \varepsilon_{i-1}, \varepsilon_i) \in \mathbb{R}^q,$$

for some vector-valued function $G_i(\cdot) = (g_{i1}(\cdot), \dots, g_{iq}(\cdot))$. By Wold representation

theorem for stationary processes, this causal representation holds in many cases. Define the non-decreasing filtration

$$\mathcal{F}_i := \sigma(\dots, \varepsilon_{i-1}, \varepsilon_i).$$

Using this filtration, we also use the notation $W_i = G_i(\mathcal{F}_i)$. To measure the strength of dependence, define for $r \geq 1$ and $1 \leq j \leq q$, the **functional dependence measure**

$$\delta_{s,r,j} := \max_{1 \leq i \leq n} \|W_i(j) - W_{i,s}(j)\|_r, \quad \text{and} \quad \Delta_{m,r,j} := \sum_{s=m}^{\infty} \delta_{s,r,j},$$

where

$$W_{i,s}(j) := g_{ij}(\mathcal{F}_{i,i-s}) \quad \text{with} \quad \mathcal{F}_{i,i-s} := \sigma(\dots, \varepsilon_{i-s-1}, \varepsilon'_{i-s}, \varepsilon_{i-s+1}, \dots, \varepsilon_{i-1}, \varepsilon_i). \quad (4.44)$$

The σ -field $\mathcal{F}_{i,i-s}$ represents a coupled version of \mathcal{F}_i . The quantity $\delta_{s,r,j}$ measures the dependence using the distance in terms of $\|\cdot\|_r$ -norm between $g_{ij}(\mathcal{F}_i)$ and $g_{ij}(\mathcal{F}_{i,i-s})$. In other words, it is quantifying the impact of changing input ε_{i-s} on the output $g_{ij}(\mathcal{F}_i)$; see Definition 1 of [Wu \(2005\)](#). The **dependence adjusted norm** for j -th coordinate is given by

$$\|\{W(j)\}\|_{r,\nu} := \sup_{m \geq 0} (m+1)^\nu \Delta_{m,r,j}, \quad \nu \geq 0.$$

To summarize these measures for the vector-valued process, define

$$\|\{W\}\|_{r,\nu} := \max_{1 \leq j \leq q} \|\{W(j)\}\|_{r,\nu} \quad \text{and} \quad \|\{W\}\|_{\psi_\alpha,\nu} := \sup_{r \geq 2} r^{-1/\alpha} \|\{W\}\|_{r,\nu}.$$

Remark 4.E.1 (Independent Sequences) Any notion of dependence should at

least include independent random variables. It might be helpful to understand how independent random variables fits into this framework of dependence. For independent random vectors W_i , the causal representation reduces to

$$W_i = G_i(\dots, \varepsilon_{i-1}, \varepsilon_i) = G_i(\varepsilon_i) \in \mathbb{R}^q.$$

It is not a function of any of the previous $\varepsilon_j, j < i$. This implies by the definition (4.44) that

$$W_{i,s} = \begin{cases} G_i(\varepsilon_i) = W_i, & \text{if } s \geq 1, \\ G_i(\varepsilon'_i) =: W'_i, & \text{if } s = 0. \end{cases}$$

Here W'_i represents an independent and identically distributed copy of W_i . Hence,

$$\delta_{s,r,j} = \begin{cases} 0, & \text{if } s \geq 1, \\ \|W_i(j) - W'_i(j)\|_r \leq 2 \|W_i(j)\|_r, & \text{if } s = 0. \end{cases}$$

It is now clear that for any $\nu > 0$,

$$\|\{W\}\|_{r,\nu} = \sup_{m \geq 0} (m+1)^\nu \Delta_{m,r} = \Delta_{0,r} \leq 2 \max_{1 \leq j \leq q} \|W_i(j)\|_r.$$

Hence, if the independent sequence W_i satisfies $\|W_i(j)\|_r \leq Cr^{1/\alpha}$ for some $C > 0$ and for all $r \geq 1$, then $\|\{W\}\|_{\psi_\alpha,\nu} < \infty$ for all $\nu > 0$, in particular for $\nu = \infty$. Therefore, independence corresponds to $\nu = \infty$. As ν decreases to zero, the random vectors become more and more dependent. \diamond

Recall that

$$\left\| \hat{\Omega}_n - \Omega_n \right\|_\infty := \max_{1 \leq j, k \leq p+1} \left| \frac{1}{n} \sum_{i=1}^n (Z_i(j)Z_i(k) - \mathbb{E}[Z_i(j)Z_i(k)]) \right|,$$

which is a maximum of $(p + 1)^2$ many averages. To prove rate of convergence for this with p possible increasing, we need a tail bound for each average. The following result, which is an extension of Theorem 2 of [Wu and Wu \(2016\)](#), provides a tail bound for an average of mean zero functionally dependent real-valued random variables with exponential tails. For proving these moment bounds, we need a few preliminary results. Set $q = 1$ and so the causal representation becomes

$$W_i = g_i(\dots, \varepsilon_{i-1}, \varepsilon_i), \quad (4.45)$$

for some real valued function g_i . We write $\delta_{k,r} = \|W_i - W_{i,k}\|_r$. The following proposition bounds the r -th moment of W_i in terms of $\|\{W\}\|_{r,\nu}$. This is based on the calculation shown after Equation (2.8) in [Wu and Wu \(2016\)](#).

Proposition 1. *Consider the setting above. If $\mathbb{E}[W_i] = 0$ for $1 \leq i \leq n$, then*

$$\|W_i\|_r \leq \|\{W\}\|_{r,0} \leq \|\{W\}\|_{r,\nu}, \quad \text{for any } r \geq 1 \quad \text{and } \nu > 0.$$

Proof. Assuming $\mathbb{E}[W_i] = 0$ for $1 \leq i \leq n$, it follows that

$$W_i = \sum_{\ell=-\infty}^i (\mathbb{E}[W_i|\mathcal{F}_\ell] - \mathbb{E}[W_i|\mathcal{F}_{\ell-1}]),$$

and so,

$$\|W_i\|_r \leq \sum_{\ell=-\infty}^i \|\mathbb{E}[W_i|\mathcal{F}_\ell] - \mathbb{E}[W_i|\mathcal{F}_{\ell-1}]\|_r = \sum_{\ell=-\infty}^i \|\mathbb{E}[W_i - W_{i,i-\ell}|\mathcal{F}_{-\ell}]\|_r \leq \sum_{\ell=0}^{\infty} \delta_{\ell,r}.$$

The last inequality follows from Jensen's inequality and noting that the last bound equals $\Delta_{0,r}$, it follows that $\|W_i\|_r \leq \Delta_{0,r} = \|\{W\}\|_{r,0}$. \square

The following lemma provides a bound on the moments of martingales in terms

of the moments of the martingale difference sequence. This result is an improvement over the classical Burkholder's inequality.

Lemma 8 (Theorem 2.1 of [Rio \(2009\)](#)). *Let $\{S_n : n \geq 0\}$ be a martingale sequence with $S_0 = 0$ adapted with respect to some non-decreasing filtration $\mathcal{F}_n, n \geq 0$. Let $E_k = S_k - S_{k-1}$ denote the corresponding martingale difference sequence. Then for any $p \geq 2$,*

$$\|S_n\|_p \leq \sqrt{p-1} \left(\sum_{k=1}^n \|E_k\|_p^2 \right)^{1/2}.$$

We are now ready to state and prove the theorem about tail bound of sum of functionally dependent random variables. Define the functions

$$s(\lambda) := (1/2 + 1/\lambda)^{-1}, \quad \text{and} \quad T_1(\lambda) := \min\{\lambda, 1\} \quad \text{for all } \lambda > 0. \quad (4.46)$$

Theorem 12. *Suppose W_1, \dots, W_n are elements of the causal process (4.45) with mean zero. If for some $\alpha > 0$, and $\nu > 0$,*

$$\|\{W\}\|_{\psi_{\alpha, \nu}} = \sup_{p \geq 2} \sup_{m \geq 0} p^{-1/\alpha} (m+1)^\nu \Delta_{m,p} < \infty. \quad (4.47)$$

Define

$$\Omega_n(\nu) := 2^\nu \times \begin{cases} 5/(\nu - 1/2)^3, & \text{if } \nu > 1/2, \\ 2(\log_2 n)^{5/2}, & \text{if } \nu = 1/2, \\ 5(2n)^{(1/2-\nu)}/(1/2 - \nu)^3, & \text{if } \nu < 1/2. \end{cases}$$

Then for any $p \geq 2$,

$$\left\| \sum_{i=1}^n W_i \right\|_p \leq \sqrt{pn} \|\{W\}\|_{\psi_{\alpha, \nu}} C_1(\nu) + C_\alpha \|\{W\}\|_{\psi_{\alpha, \nu}} (\log n)^{1/s(\alpha)} p^{1/T_1(s(\alpha))} \Omega_n(\nu), \quad (4.48)$$

where C_α is a constant depending only on α , $C_1(\nu)$ is a constants depending only on ν given by

$$C_1(\nu) := \left[\sqrt{6} + \frac{20\sqrt{2}\pi^3 2^\nu}{3\nu^3} \right].$$

Furthermore, it follows by Markov's inequality that for all $t \geq 0$,

$$\mathbb{P} \left(\left| \sum_{i=1}^n W_i \right| \geq e\sqrt{tn} \|\{W\}\|_{2,\nu} C_1(\nu) + C_\alpha \|\{W\}\|_{\psi_\alpha,\nu} t^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \Omega_n(\nu) \right) \leq 8e^{-t}.$$

Here C_α is different from the one in the moment bound (4.48).

Proof. Define

$$S_n := \sum_{i=1}^n W_i, \quad L = \left\lfloor \frac{\log n}{\log 2} \right\rfloor, \quad \text{and} \quad \xi_\ell = \begin{cases} 2^\ell, & \text{if } 0 \leq \ell < L, \\ n, & \text{if } \ell = L. \end{cases}$$

Define for $m \geq 0$,

$$W_i^{(m)} := \mathbb{E} [W_i | \varepsilon_{i-m}, \dots, \varepsilon_i], \quad \text{and} \quad M_{i,\ell} := \sum_{k=1}^i \left(W_k^{(\xi_\ell)} - W_k^{(\xi_{\ell-1})} \right).$$

Let

$$S_{n,m} := \sum_{i=1}^n W_i^{(m)},$$

and consider the decomposition

$$S_n = S_{n,0} + (S_n - S_{n,n}) + \sum_{\ell=1}^L (S_{n,\xi_\ell} - S_{n,\xi_{\ell-1}}) := \mathbf{I} + \mathbf{II} + \mathbf{III}. \quad (4.49)$$

We prove the moment bound (4.48) by bounding the moments of each term in the decomposition (4.49).

Bounding I: Regarding the first term **I**, observe that $S_{n,0}$ is a sum of independent

random variables $W_i^{(0)}$ satisfying the tail assumption of Theorem 3.2 of [Kuchibhotla and Chakraborty \(2018\)](#) with $\beta = \alpha$. This verification follows by noting that

$$\left\| W_i^{(0)} \right\|_p \stackrel{(a)}{\leq} \|W_i\|_p \stackrel{(b)}{\leq} \|\{W\}\|_{p,\nu} \stackrel{(c)}{\leq} p^{1/\alpha} \|\{W\}\|_{\psi_{\alpha,\nu}}.$$

Inequality (a) follows from Jensen's inequality, (b) follows from Proposition 1 and (c) follows assumption (4.47). Hence, we get that for any $p \geq 1$,

$$\begin{aligned} \|\mathbf{I}\|_p &= \left\| \sum_{i=1}^n \mathbb{E} [W_i | \varepsilon_i] \right\|_p \\ &\leq \sqrt{6p} \left(\sum_{i=1}^n \mathbb{E} [W_i^2] \right)^{1/2} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/T_1(\alpha)} (\log n)^{1/\alpha}, \end{aligned}$$

for some constant C_α depending only on α . Here Jensen's inequality is used to bound the variance of $\mathbb{E} [W_i | \varepsilon_i]$. By Proposition 1, $\|W_i\|_2 \leq \|\{W\}\|_{2,\nu}$ and hence

$$\|S_{n,0}\|_p \leq \sqrt{6pn} \|\{W\}\|_{2,\nu} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/T_1(\alpha)} (\log n)^{1/\alpha}. \quad (4.50)$$

Bounding II: For the second term, note that

$$S_n = \sum_{i=1}^n W_i = \sum_{i=1}^n \mathbb{E} [W_i | \varepsilon_i, \varepsilon_{i-1}, \dots] = S_{n,\infty},$$

and hence,

$$S_n - S_{n,n} = \sum_{m=n}^{\infty} (S_{n,m+1} - S_{n,m}).$$

Substituting the definition of $S_{n,m}$, we have

$$S_{n,m+1} - S_{n,m} = \sum_{k=1}^n (\mathbb{E} [W_k | \varepsilon_k, \dots, \varepsilon_{k-m-1}] - \mathbb{E} [W_k | \varepsilon_k, \dots, \varepsilon_{k-m}]).$$

We now prove that the summands above form a martingale difference sequence with respect to a filtration. The following construction is taken from the proof of Lemma 1 of [Liu and Wu \(2010\)](#). Define

$$D_{k,m+1} := \mathbb{E} [W_k | \varepsilon_k, \dots, \varepsilon_{k-m-1}] - \mathbb{E} [W_k | \varepsilon_k, \dots, \varepsilon_{k-m}],$$

and the non-decreasing filtration

$$\mathcal{G}_{k,m+1} := \sigma(\varepsilon_{k-m-1}, \varepsilon_{k-m}, \dots).$$

It is easy to see that

$$\mathbb{E} [D_{n-k+1,m+1} | \mathcal{G}_{k-1,m+1}] = 0. \quad (4.51)$$

Therefore, $\{(D_{n-k+1,m+1}, \mathcal{G}_{k,m+1}) : 1 \leq k \leq n\}$ forms a martingale difference sequence. This implies that $S_{n,m+1} - S_{n,m}$ is a martingale and hence by Lemma 8 we get for $p \geq 2$,

$$\|S_{n,m+1} - S_{n,m}\|_p^2 \leq p \sum_{k=1}^n \|D_{k,m+1}\|_p^2.$$

To further bound the right hand side, note that for $p \geq 2$,

$$\|D_{k,m+1}\|_p = \|\mathbb{E} [W_k - g(\dots, \varepsilon'_{k-m-1}, \varepsilon_{k-m}, \dots, \varepsilon_k) | \varepsilon_k, \dots, \varepsilon_{k-m-1}]\|_p \leq \delta_{m+1,p}. \quad (4.52)$$

Hence, for $p \geq 2$,

$$\|S_{n,m+1} - S_{n,m}\|_p \leq \sqrt{pn} \delta_{m+1,p},$$

and

$$\|S_n - S_{n,n}\|_p \leq \sum_{m=n}^{\infty} \|S_{n,m+1} - S_{n,m}\|_p \leq \sqrt{pn} \sum_{m=n}^{\infty} \delta_{m+1,p} = \sqrt{pn} \Delta_{n+1,p}.$$

Under assumption (4.47), we obtain

$$\|\mathbf{II}\|_p = \|S_n - S_{n,n}\|_p \leq \|\{W\}\|_{\psi_{\alpha,\nu}} \frac{n^{1/2} p^{1/2+1/\alpha}}{(n+2)^\nu} = \|\{W\}\|_{\psi_{\alpha,\nu}} n^{1/2-\nu} p^{1/2+1/\alpha}. \quad (4.53)$$

Bounding III: To bound **III**, note by definition of $M_{i,\ell}$ that

$$\mathbf{III} = \sum_{\ell=1}^L \sum_{k=1}^n \left(W_k^{(\xi_\ell)} - W_k^{(\xi_{\ell-1})} \right) = \sum_{\ell=1}^L M_{n,\ell}.$$

Now observe that the summands of $M_{n,\ell}$,

$$\mathcal{D}_{k,\ell} := \left(W_k^{(\xi_\ell)} - W_k^{(\xi_{\ell-1})} \right),$$

are ξ_ℓ -dependent in the sense that $\mathcal{D}_{k,\ell}$ and $\mathcal{D}_{s,\ell}$ are independent if $|s - k| > \xi_\ell$. This can be proved as follows. By definition $\mathcal{D}_{k,\ell}$ is only a function of $(\varepsilon_k, \dots, \varepsilon_{k-\xi_\ell})$ and by independence of $\varepsilon_k, k \in \mathbb{Z}$, the claim follows. Now a blocking technique can be used to convert $M_{n,\ell}$ into a sum of independent variables. See Corollary A.1 of [Romano and Wolf \(2000\)](#) for a similar use. Define

$$\mathcal{A}_\ell := \{2\xi_\ell i + j : i \in \mathbb{Z}, 1 \leq j \leq \xi_\ell\},$$

$$\mathcal{B}_\ell := \{2\xi_\ell i + \xi_\ell + j : i \in \mathbb{Z}, 1 \leq j \leq \xi_\ell\}.$$

Consider the decomposition of $M_{n,\ell}$ as

$$M_{n,\ell} = \sum_{k=1}^n \mathcal{D}_{k,\ell} = A_{n,\ell} + B_{n,\ell},$$

where

$$A_{n,\ell} := \sum_{1 \leq k \leq n, k \in \mathcal{A}} \mathcal{D}_{k,\ell} \quad \text{and} \quad B_{n,\ell} := \sum_{1 \leq k \leq n, k \in \mathcal{B}} \mathcal{D}_{k,\ell}.$$

We now provide moment bounds for $M_{n,\ell}$ by giving moment bounds for $A_{n,\ell}$ and $B_{n,\ell}$ which is in turn done by separating the summands of $A_{n,\ell}$ and $B_{n,\ell}$ to form an independent sum. Note that

$$\begin{aligned} A_{n,\ell} &= \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} \left(\sum_{j=1}^{\xi_\ell} \mathcal{D}_{2\xi_\ell i+j,\ell} \right) = \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} \left(\sum_{k=2\xi_\ell i+1}^{2\xi_\ell i+\xi_\ell} \left(W_k^{(\xi_\ell)} - W_k^{(\xi_\ell-1)} \right) \right) \\ &= \sum_{i=1}^{\lfloor \frac{n}{2\xi_\ell} \rfloor} (M_{2\xi_\ell i+\xi_\ell,\ell} - M_{2\xi_\ell i,\ell}). \end{aligned} \tag{4.54}$$

By the ξ_ℓ -independence of the summands of $M_{n,\ell}$, we get that the summands in the final representation of $A_{n,\ell}$ are independent and so Theorem 3.2 of [Kuchibhotla and Chakraborty \(2018\)](#) applies. In the following, we verify the assumption of Theorem 3.2 of [Kuchibhotla and Chakraborty \(2018\)](#). For $1 \leq i < j \leq n$, it is clear that

$$\begin{aligned} M_{j,\ell} - M_{i,\ell} &= \sum_{k=i+1}^j \left(W_k^{(\xi_\ell)} - W_k^{(\xi_\ell-1)} \right) \\ &= \sum_{k=i+1}^j \left(\sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left(W_k^{\xi_\ell} - W_k^{(\xi_\ell-1)} \right) \right) \\ &= \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left(\sum_{k=i+1}^j \left(W_k^{(t)} - W_k^{(t-1)} \right) \right). \end{aligned}$$

By triangle inequality

$$\|M_{j,\ell} - M_{i,\ell}\|_p \leq \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \left\| \sum_{k=i+1}^j (W_k^{(t)} - W_k^{(t-1)}) \right\|_p. \quad (4.55)$$

As proved in (4.51), the summation for each t represents a martingale and hence by Lemma 8, we get for $p \geq 2$ that

$$\left\| \sum_{k=i+1}^j (W_k^{(t)} - W_k^{(t-1)}) \right\|_p^2 \leq p \sum_{k=i+1}^j \|W_k^{(t)} - W_k^{(t-1)}\|_p^2 \leq p \sum_{k=i+1}^j \delta_{t,p}^2 = p(j-i)\delta_{t,p}^2.$$

Here we used inequality (4.52). Substituting this in inequality (4.55) and using $\xi_{\ell-1} \geq \xi_\ell/2$, we get

$$\begin{aligned} \|M_{j,\ell} - M_{i,\ell}\|_p &\leq p^{1/2}(j-i)^{1/2} \sum_{t=1+\xi_{\ell-1}}^{\xi_\ell} \delta_{t,p} \leq p^{1/2}(j-i)^{1/2} \Delta_{1+\xi_{\ell-1},p} \\ &\leq \|\{W\}\|_{p,\nu} p^{1/2}(j-i)^{1/2}(2+\xi_{\ell-1})^{-\nu} \\ &\leq 2^\nu \|\{W\}\|_{p,\nu} p^{1/2}(j-i)^{1/2} \xi_\ell^{-\nu}. \end{aligned} \quad (4.56)$$

Under assumption (4.47), we get

$$\begin{aligned} \|M_{j,\ell} - M_{i,\ell}\|_p &\leq 2^\nu \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/2+1/\alpha}(j-i)^{1/2} \xi_\ell^{-\nu} \\ &= 2^\nu \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/s(\alpha)}(j-i)^{1/2} \xi_\ell^{-\nu}. \end{aligned}$$

See (4.46) for the definition of $s(\alpha)$. Thus, for all $1 \leq i \leq \lfloor \frac{n}{2\xi_\ell} \rfloor$,

$$\sup_{p \geq 2} p^{-1/s(\alpha)} \|M_{2\xi_\ell i + \xi_\ell, \ell} - M_{2\xi_\ell i, \ell}\|_p \leq 2^\nu \|\{W\}\|_{\psi_{\alpha,\nu}} \xi_\ell^{1/2-\nu}.$$

So, the summands of $A_{n,\ell}$ in the final representation in (4.54) are independent and satisfy the hypothesis of Theorem 3.2 of [Kuchibhotla and Chakraborty \(2018\)](#) with

$\beta = s(\alpha)$. Therefore, for $p \geq 2$,

$$\begin{aligned}
\|A_{n,\ell}\|_p &\leq \sqrt{6p} \left(\sum_{i=1}^{\lfloor n/(2\xi_\ell) \rfloor} \|M_{2\xi_\ell i + \xi_\ell, \ell} - M_{2\xi_\ell i, \ell}\|_2^2 \right)^{1/2} \\
&\quad + C_\alpha 2^\nu \|\{W\}\|_{\psi_{\alpha,\nu}} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2-\nu} p^{1/T_1(s(\alpha))} \\
&\leq \sqrt{12p} \|\{W\}\|_{2,\nu} \frac{2^\nu \xi_\ell^{1/2}}{\xi_\ell^\nu} \left(\frac{n}{2\xi_\ell} \right)^{1/2} \\
&\quad + C_\alpha 2^\nu \|\{W\}\|_{\psi_{\alpha,\nu}} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2-\nu} p^{1/T_1(s(\alpha))} \\
&\leq \frac{2^\nu}{\xi_\ell^\nu} \left[\|\{W\}\|_{2,\nu} \sqrt{6pn} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2} \right].
\end{aligned}$$

Here the second inequality follows from (4.56).

Similarly a representation for $B_{n,\ell}$ exists with independent summands satisfying the assumption of Theorem 3.2 of [Kuchibhotla and Chakraborty \(2018\)](#) with $\beta = s(\alpha)$ and so,

$$\|B_{n,\ell}\|_p \leq \frac{2^\nu}{\xi_\ell^\nu} \left[\|\{W\}\|_{2,\nu} \sqrt{6pn} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2} \right].$$

Combining the bounds for $A_{n,\ell}$ and $B_{n,\ell}$ implies the bound on $M_{n,\ell}$ as

$$\|M_{n,\ell}\|_p \leq \frac{2^{1+\nu}}{\xi_\ell^\nu} \left[\|\{W\}\|_{2,\nu} \sqrt{6pn} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} p^{1/T_1(s(\alpha))} (\log n)^{1/s(\alpha)} \xi_\ell^{1/2} \right]. \quad (4.57)$$

To complete bounding **III**, we need to bound the moments of the sum of $M_{n,\ell}$ over $1 \leq \ell \leq L$ which are all dependent. For this, define the sequence

$$\lambda_\ell = \begin{cases} 3\pi^{-2}\ell^{-2}, & \text{if } 1 \leq \ell \leq L/2, \\ 3\pi^{-2}(L+1-\ell)^{-2}, & \text{if } L/2 < \ell \leq L. \end{cases}$$

This positive sequence satisfies $\sum_{\ell=1}^L \lambda_\ell < 1$. It is easy to derive from Hölder's

inequality that

$$\left| \sum_{\ell=1}^L a_\ell \right|^p \leq \sum_{\ell=1}^L \frac{|a_\ell|^p}{\lambda_\ell^p}.$$

Substituting in this inequality $a_\ell = M_{n,\ell}$ and the moment bound (4.57), we get

$$\begin{aligned} \mathbb{E} \left[\left| \sum_{\ell=1}^L M_{n,\ell} \right|^p \right] &\leq 2^{(2+\nu)p} \|\{W\}\|_{2,\nu}^p (6pn)^{p/2} \sum_{\ell=1}^L \frac{1}{\lambda_\ell^p \xi_\ell^{p\nu}} \\ &\quad + C_\alpha 2^{(2+\nu)p} \|\{W\}\|_{\psi_{\alpha,\nu}}^p p^{p/T_1(s(\alpha))} (\log n)^{p/s(\alpha)} \sum_{\ell=1}^L \frac{\xi_\ell^{p/2}}{\lambda_\ell^p \xi_\ell^{p\nu}}. \end{aligned}$$

It follows from Lemma 9 and the definition of $\Omega_n(\nu)$ that for $p \geq 2$,

$$\begin{aligned} &\left\| \sum_{\ell=1}^L M_{n,\ell} \right\|_p \\ &\leq \frac{5\pi^3 2^2}{3\sqrt{3}} \left[\frac{2^\nu \|\{W\}\|_{2,\nu} \sqrt{6pn}}{\nu^3} + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} (\log n)^{1/s(\alpha)} \Omega_n(\nu) p^{1/T_1(s(\alpha))} \right]. \end{aligned} \tag{4.58}$$

Combining the moment bounds (4.50), (4.53) and (4.58), it follows that for $p \geq 2$,

$$\begin{aligned} \|S_n\|_p &\leq \sqrt{6pn} \|\{W\}\|_{\psi_{\alpha,\nu}} \left[1 + \frac{20\pi^3 2^\nu}{3\sqrt{3}\nu^3} \right] + \|\{W\}\|_{\psi_{\alpha,\nu}} n^{1/2-\nu} p^{1/s(\alpha)} \\ &\quad + C_\alpha \|\{W\}\|_{\psi_{\alpha,\nu}} (\log n)^{1/s(\alpha)} p^{1/T_1(s(\alpha))} \Omega_n(\nu). \end{aligned}$$

Here the inequalities $s(\alpha) \leq \alpha$ and $T_1(s(\alpha)) \leq T_1(\alpha)$ are used. Now noting that $\Omega_n(\nu) \geq n^{1/2-\nu}$ for all $\nu > 0$ and $p^{1/s(\alpha)} \leq p^{1/T_1(s(\alpha))}$, the result follows. \square

The following simple calculation is also used in Theorem 12. Define

$$L := \left\lfloor \frac{\log n}{\log 2} \right\rfloor \quad \text{and} \quad \lambda_\ell = \begin{cases} 3\pi^{-2}\ell^{-2}, & \text{if } 1 \leq \ell \leq L/2, \\ 3\pi^{-2}(L+1-\ell)^{-2}, & \text{if } L/2 < \ell \leq L. \end{cases}$$

Lemma 9. *The following inequalities hold true:*

(a) *For any $\beta > 0$ and $p \geq 2$,*

$$\sum_{\ell=1}^L \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} \leq 2 \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} \leq \left(\frac{5}{\beta^3}\right)^p \left(\frac{\pi^2}{3}\right)^{p+1}.$$

(b) *For any $\beta > 0$ and $p \geq 2$,*

$$\sum_{\ell=1}^L \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} \leq \left(\frac{\pi^2}{3}\right)^{p+1} \begin{cases} (5/(\beta - 1/2)^3)^p, & \text{if } \beta > 1/2, \\ 2(\log_2 n)^{2p+1}, & \text{if } \beta = 1/2, \\ (2n)^{(1/2-\beta)p} (5/(1/2 - \beta)^3)^p, & \text{if } \beta < 1/2. \end{cases}$$

Proof. (a) Note that for any $\beta > 0$,

$$\sup_{\ell>0} \ell^3 2^{-\ell\beta} = \ell^3 \exp(-(\log 2)\ell\beta) \leq \left(\frac{3}{e\beta \log 2}\right)^3 \leq \frac{5}{\beta^3},$$

and so,

$$\begin{aligned} \left(\frac{3}{\pi^2}\right)^p \sum_{\ell=1}^L \frac{1}{\lambda_\ell^p 2^{p\ell\beta}} &= \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p + \sum_{\ell=L/2+1}^L \left(\frac{(L+1-\ell)^2}{2^{\ell\beta}}\right)^p \\ &\leq \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p + 2^{-p\beta} \sum_{\ell=1}^{L/2} \left(\frac{\ell^2}{2^{\ell\beta}}\right)^p \\ &\leq 2 \left(\frac{5}{\beta^3}\right)^p \sum_{\ell=1}^{L/2} \frac{1}{\ell^p} \leq \frac{\pi^2}{3} \left(\frac{5}{\beta^3}\right)^p. \end{aligned}$$

Hence the result (a) follows.

(b) If $\beta > 1/2$, then

$$\sum_{\ell=1}^L \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} = \sum_{\ell=1}^L \frac{1}{\ell^p 2^{p\ell(\beta-1/2)}},$$

and so, the bound for this case follows from (a).

If $\beta = 1/2$, then

$$\sum_{\ell=1}^L \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} = \sum_{\ell=1}^L \frac{1}{\lambda_\ell^p} \leq 2 \left(\frac{\pi^2}{3}\right)^p \sum_{\ell=1}^{L/2} \ell^{2p} \leq 2 \left(\frac{\pi^2}{3}\right)^p \left(\frac{\log n}{\log 2}\right)^{2p+1}.$$

If $\beta > 1/2$, then

$$\begin{aligned} & \sum_{\ell=1}^L \frac{2^{p\ell(1/2-\beta)}}{\lambda_\ell^p} \\ &= \sum_{\ell=1}^{L/2} \frac{2^{\ell(1/2-\beta)p}}{\lambda_\ell^p} + 2^{(L+1)(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} \\ &\leq \sum_{\ell=1}^{L/2} \frac{2^{\ell(1/2-\beta)p}}{\lambda_\ell^p} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} \\ &\leq 2^{(L+1)(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{(L+1-\ell)(1/2-\beta)p}} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} \\ &\leq (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} + (2n)^{(1/2-\beta)p} \sum_{\ell=1}^{L/2} \frac{1}{\lambda_\ell^p 2^{\ell(1/2-\beta)p}} \\ &\leq (2n)^{(1/2-\beta)p} \left(\frac{5}{(1/2-\beta)^3}\right)^p \left(\frac{\pi^2}{3}\right)^{p+1}. \end{aligned}$$

Hence the result follows. □

End of Chapter 4.

All of Linear Regression

Arun K. Kuchibhotla, Lawrence D. Brown, Andreas Buja, and Junhui Cai

University of Pennsylvania

e-mail: arunku@upenn.edu, buja.at.wharton@gmail.com

Abstract: Least squares linear regression is one of the oldest and widely used data analysis tools. Although the theoretical analysis of ordinary least squares (OLS) estimator is as old, several fundamental questions are yet to be answered. Suppose regression observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ (not necessarily independent) are available. Some of the questions we deal with are as follows: under what conditions, does the OLS estimator converge and what is the limit? What happens if the dimension is allowed to grow with n ? What happens if the observations are dependent with dependence possibly strengthening with n ? How to do statistical inference under these kinds of misspecification? What happens to OLS estimator under variable selection? How to do inference under misspecification and variable selection?

We answer all the questions raised above with one simple deterministic inequality which holds for any set of observations and any sample size. This implies that all our results are finite sample (non-asymptotic) in nature. At the end, one only needs to bound certain random quantities under specific settings of interest to get concrete rates and we derive these bounds for the case of independent observations. In particular the problem of inference after variable selection is studied, for the first time, when d , the number of covariates increases (almost exponentially) with sample size n . We provide comments on the “right” statistic to consider for inference under variable selection and efficient computation of quantiles.

1. Introduction

Linear regression is one of the oldest and most widely practiced data analysis method. In many real data settings least squares linear regression leads to performance in par with state-of-the-art (and often far more complicated) methods while remaining amenable to interpretation. These advantages coupled with the argument “all models are wrong” warrants a detailed study of least squares linear regression estimator in settings that are close to the practical/realistic ones. Instead of proposing assumptions that we think are practical/realistic, we start with a clean slate. We start by not assuming anything about the observations

Figure 4.3: The following chapter is partly based on [Kuchibhotla et al. \(2019\)](#).

Chapter 5

Unified Framework for Post-selection Inference

In the previous chapter, we have discussed an approach for valid post-selection inference in (assumption-lean) linear regression. One of its great strengths is its validity and computational complexity. However, it has two main disadvantages (as mentioned in Section 4.7 in Chapter 4):

1. It is not clear if the confidence regions there are (asymptotically) tight, that is, there exists a random model-selection procedure that requires such confidence regions produced by Approach 1. Also, the confidence regions are not equivariant.
2. There is no clear generalization to other M -estimation problems.

In this chapter, we develop a unified recipe for valid post-selection inference (to tackle these disadvantages) based on the ideas of [Berk et al. \(2013\)](#) and [Bachoc et al. \(2016\)](#). The recipe is more widely applicable than either of these papers and uses the techniques of high-dimensional central limit theorem and multiplier bootstrap from [Cher-](#)

nozhlukov et al. (2017a). The main component of this idea is uniform asymptotic linear representation. Although the framework generalizes to dependent settings easily, we discuss the general recipe under independence and comments related to other dependent settings will be provided later.

The remaining chapter is organized as follows. Section 5.1 provides the unified framework for valid post-selection inference along with the main results proving validity. As a first example, valid post-selection inference in the linear regression problem is revisited in Section 5.2. Some discussion computation and approximate algorithms is given in Section 5.3.

5.1 General Recipe for Valid PoSI

To understand the ideology behind the general recipe for valid post-selection inference for multiple targets of estimation, let us first understand how one would do valid inference for one target of estimation. Suppose Z_1, \dots, Z_n are random vectors and the estimator is

$$\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^q} \frac{1}{n} \sum_{i=1}^n \rho(Z_i, \theta),$$

for some loss function $\rho(\cdot, \cdot)$ that is twice differentiable in the second argument. Under regularity conditions, $\hat{\theta}_n$ satisfies

$$\frac{1}{n} \sum_{i=1}^n \varphi(Z_i, \hat{\theta}_n) = 0, \quad \text{with} \quad \varphi(z, \theta) := \frac{\partial}{\partial t} \rho(z, t) \Big|_{t=\theta}.$$

Define θ_n through

$$\theta_n := \arg \min_{\theta \in \mathbb{R}^q} \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\rho(Z_i, \theta)].$$

This implies that $n^{-1} \sum_{i=1}^n \mathbb{E} [\varphi(Z_i, \theta_n)] = 0$. By a Taylor series expansion of $\varphi(\cdot, \cdot)$ in terms of the second argument around θ_n , we get

$$\frac{1}{n} \sum_{i=1}^n \dot{\varphi}(Z_i, \theta_n^*)(\hat{\theta}_n - \theta_n) = -\frac{1}{n} \sum_{i=1}^n \varphi(Z_i, \theta_n), \quad \text{with} \quad \dot{\varphi}(z, \theta) := \left. \frac{\partial}{\partial t} \varphi(z, t) \right|_{t=\theta}.$$

for some θ_n^* that lies on the line segment between $\hat{\theta}_n$ and θ_n . Under some regularity conditions, if $\hat{\theta}_n - \theta_n = o_p(1)$, then

$$\frac{1}{n} \sum_{i=1}^n \dot{\varphi}(Z_i, \theta_n^*) - \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\dot{\varphi}(Z_i, \theta_n)] = o_p(1),$$

and so,

$$\sqrt{n}(\hat{\theta}_n - \theta_n) = \left(\frac{1}{n} \sum_{i=1}^n \mathbb{E} [\dot{\varphi}(Z_i, \theta_n)] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(Z_i, \theta_n) + o_p(1). \quad (5.1)$$

This implies that $\sqrt{n}(\hat{\theta}_n - \theta_n)$ has an asymptotic normal distribution with mean zero. All the steps here can be made rigorous with not so unreasonable conditions on $\rho(\cdot, \cdot)$. See [Yuan and Jennrich \(1998\)](#) and [Kuchibhotla \(2018\)](#) for more details. Estimating the variance of this normal implies classical inference (such as confidence regions, hypothesis tests and so on).

From this analysis, we see that the main component of valid inference for θ_n is to prove an asymptotic linear representation (5.1). Also, this allows one to use the ordinary or score bootstrap since the estimator $\hat{\theta}_n$ behaves approximately like an average. This is the underlying ideology for the general recipe of valid post-selection inference. As shown in Section 4.2, valid post-selection inference is equivalent to valid simultaneous inference. In this general recipe, we require the asymptotic linear representation to hold uniformly over the set of all targets that one wants inference

for. See Figure 2.1 in Chapter 2.

Getting back to the general recipe for valid post-selection inference, let us consider a concrete problem with covariate selection. Let $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ be n random vectors taking values in \mathbb{R}^p and for each $M \subseteq \{1, 2, \dots, p\}$, let $\hat{\beta}_M$ denote the estimator

$$\hat{\beta}_M := \arg \min_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^n \rho(Y_i, X_{i,M}, \theta). \quad (5.2)$$

Here $X_{i,M}$ denotes the subvector of X_i with indices in M . When the analyst considers this estimator, he/she implicitly decides the target vector as

$$\beta_M := \arg \min_{\theta \in \mathbb{R}^{|M|}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\rho(Y_i, X_{i,M}, \theta)]. \quad (5.3)$$

We stress here that both $\hat{\beta}_M$ and β_M depend on the sample size n but for convenience, we drop this dependence in the notation. The previous discussion shows that for each fixed M , $\hat{\beta}_M$ is asymptotically normal centered at β_M . The post-selection inference problem in this case is as follows. Fix a universe \mathcal{M} of subsets of $\{1, 2, \dots, p\}$. Using the data the analyst is allowed to pick an element $\hat{M} \in \mathcal{M}$ and perform inference based on the estimator $\hat{\beta}_{\hat{j}, \hat{M}}$, where for $\hat{j} \in \hat{M}$, $\hat{\beta}_{\hat{j}, \hat{M}}$ represents the estimator corresponding to the \hat{j} -th covariate. Hence the PoSI problem is to construct a confidence region $\hat{\mathcal{R}}_{\hat{j}, \hat{M}}$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\beta_{\hat{j}, \hat{M}} \in \hat{\mathcal{R}}_{\hat{j}, \hat{M}} \right) \geq 1 - \alpha, \quad (5.4)$$

irrespective of how $\hat{M} \in \mathcal{M}$ and $\hat{j} \in \hat{M}$ are obtained. Theorem 3 of Chapter 4 proves that solving (5.4) is equivalent to solving

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{\substack{M \in \mathcal{M}, \\ j \in M}} \{ \beta_{j,M} \in \hat{\mathcal{R}}_{j,M} \} \right) \geq 1 - \alpha. \quad (5.5)$$

One of the simplest confidence regions satisfying the simultaneous guarantees is given by

$$\widehat{\mathcal{R}}_{j \cdot M}^{\text{unif-adj}} := \left\{ \theta \in \mathbb{R} : \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \theta)}{\widehat{\sigma}_{j \cdot M}} \right| \leq K_\alpha \right\}, \quad (5.6)$$

where K_α is the $(1 - \alpha)$ quantile of the “maximum statistic” or the “max-t statistic”

$$\max_{\substack{M \in \mathcal{M}, \\ j \in M}} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\widehat{\sigma}_{j \cdot M}} \right|.$$

Here $\widehat{\sigma}_{j \cdot M}$ is an estimate of the asymptotic standard deviation of $n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})$.

The confidence region in (5.6) will be referred to as uniform adjustment because the adjustment for simultaneity is uniform over all (j, M) . This confidence region is similar in spirit to the Tukey’s pairwise comparison test. The only hurdle in implementing the region (5.6) is the quantity K_α . The main crux of this chapter is spent on showing that the constant K_α can be estimated under

1. an assumption-lean setting: not requiring any parametric model or distributional assumptions;
2. only moment or tail assumptions on covariates and response;
3. the total number of covariate p possibly growing with the sample size almost exponentially;
4. both random and fixed covariates as in Chapter 3.

We now introduce the general assumptions under which the unified framework will be shown to work. These assumptions do not require that the estimator $\widehat{\beta}_M$ and target β_M are defined as in (5.2) and (5.3), but this will be our primary example. Further, we will verify the following assumptions for general loss functions $\rho(\cdot, \cdot)$ in the following sections.

Recall the observations are $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$. The unified framework is based on the following assumptions:

(B1) For the estimators $\hat{\beta}_M, M \in \mathcal{M}$, there exists targets $\beta_M, M \in \mathcal{M}$ such that

$$n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j \cdot M}(Z_i) + R_{j \cdot M}, \quad (5.7)$$

for functions $\psi_{j \cdot M}(\cdot, \cdot)$ satisfying

$$\mathbb{E} \left[\sum_{i=1}^n \psi_{j \cdot M}(Z_i) \right] = 0.$$

In (5.7), the constant $\sigma_{j \cdot M}$ represents the “asymptotic” standard deviation of $n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})$ defined by

$$\sigma_{j \cdot M}^2 := \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j \cdot M}(Z_i) \right).$$

Set

$$\Delta_{\mathcal{M}}^{\text{ULR}} := \max_{M \in \mathcal{M}, j \in M} \left| \frac{R_{j \cdot M}}{\sigma_{j \cdot M}} \right|.$$

(B2) There exists estimators $\hat{\sigma}_{j \cdot M}$ that are consistent for $\sigma_{j \cdot M}^*$ and

$$0 < \underline{\sigma} \leq \min_{M \in \mathcal{M}, j \in M} \frac{\sigma_{j \cdot M}}{\sigma_{j \cdot M}^*} \leq \max_{M \in \mathcal{M}, j \in M} \frac{\sigma_{j \cdot M}}{\sigma_{j \cdot M}^*} \leq 1. \quad (5.8)$$

Set

$$\Delta_{\mathcal{M}}^{\text{Var}} := \max_{M \in \mathcal{M}, j \in M} \left| \frac{\hat{\sigma}_{j \cdot M}}{\sigma_{j \cdot M}^*} - 1 \right|.$$

(B3) There exists estimators $\hat{\psi}_{j \cdot M}, M \in \mathcal{M}, j \in M$ estimating $\psi_{j \cdot M}(\cdot)$ on the sample.

Set

$$\Delta_{\mathcal{M}}^{\text{Inf}} := \max_{M \in \mathcal{M}, j \in M} \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}} \right)^2,$$

$$\Delta_{\mathcal{M}}^{\text{Boot}} := \max_{\substack{M \in \mathcal{M}, j \in M, \\ M' \in \mathcal{M}, j' \in M'}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} - \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} \right] \right\} \right|.$$

Some comments on the assumptions might be helpful. Firstly, none of these are assumptions; they are statements to give an indication of the quantities that are required to be bounded. Assumption (B1) is what we call **Uniform Asymptotic Linear Representation**. As shown in the analysis before, many reasonable estimators satisfy assumption (B1) in case \mathcal{M} is a singleton (Yuan and Jennrich, 1998). As described in Section 3.A, the functions $\psi_{j \cdot M}(\cdot)$ play the role of influence functions for $\hat{\theta}_{j \cdot M}$ with \mathcal{M} a singleton. So, all of assumption (B1) essentially implies is that the estimators $\hat{\theta}_{j \cdot M}$, $M \in \mathcal{M}, j \in M$ are approximately averages of n random vectors with the approximation errors disappearing uniformly over $M \in \mathcal{M}, j \in M$. Assumption (B2) although seems independent of (B1) will be applied with $\sigma_{j \cdot M}^*$ as a proxy for $\sigma_{j \cdot M}$. The reason for setting up the framework with a proxy instead of just $\sigma_{j \cdot M}$ is that in the unified framework of Chapter 3 there does not exist a consistent estimator for $\sigma_{j \cdot M}$. However, an asymptotically conservative estimator of $\sigma_{j \cdot M}$ exists and $\sigma_{j \cdot M}^*$ is the “upper bound” for $\sigma_{j \cdot M}$. We prove in Section 5.2 that both $\Delta_{\mathcal{M}}^{\text{ULR}}$ and $\Delta_{\mathcal{M}}^{\text{Var}}$ for linear regression and generalized linear models, respectively, under very general conditions allowing for the dependence of observations. Assumption (B3) requires uniformly consistent estimators of influence functions and this is used to prove the consistency of bootstrap that allows for valid estimation of quantiles. Based on Assumption (B3), a simple estimator $\hat{\sigma}_{j \cdot M}$ for the case of independent observations is

given by

$$\hat{\sigma}_{j \cdot M}^2 := \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{j \cdot M}^2(Z_i).$$

For this estimator, it is clear that

$$\left| \hat{\sigma}_{j \cdot M} - \sqrt{\frac{1}{n} \sum_{i=1}^n \psi_{j \cdot M}^2(Z_i)} \right| \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}^2(Z_i) \right)^2} \leq \sigma_{j \cdot M} (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2}.$$

Further, setting $\sigma_{j \cdot M}^* := \sqrt{n^{-1} \sum_{i=1}^n \mathbb{E}[\psi_{j \cdot M}^2(Z_i)]}$,

$$\left| \frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot M}^2(Z_i)}{\sigma_{j \cdot M}^2} - \frac{(\sigma_{j \cdot M}^*)^2}{\sigma_{j \cdot M}^2} \right| \leq \Delta_{\mathcal{M}}^{\text{Boot}}.$$

Note that $\sigma_{j \cdot M}^* \geq \sigma_{j \cdot M}$. This inequality holds because of the independence of Z_1, \dots, Z_n .

Combining these inequalities yields

$$\left| \frac{\hat{\sigma}_{j \cdot M}}{\sigma_{j \cdot M}} - \frac{\sigma_{j \cdot M}^*}{\sigma_{j \cdot M}} \right| \leq (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Boot}}.$$

Therefore, taking the maximum over all (j, M) , proves

$$\Delta_{\mathcal{M}}^{\text{Var}} \leq \max_{(j, M)} \frac{\sigma_{j \cdot M}}{\sigma_{j \cdot M}^*} \left[(\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Boot}} \right] \leq (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Boot}}. \quad (5.9)$$

This proves that in case Z_1, \dots, Z_n are independent, Assumption (B2) is implied by Assumption (B3).

There is a rich literature on uniform asymptotic linear representation and they have been used in optimal M -estimation problems. See condition (2.3) of Theorem 2.1 in Arcones (2005) and Section 10.2, 10.3, Equation (10.25) of Dodge and Jurevckova (2000) for examples where uniform asymptotic linear representation is proved for a large class of M -estimators indexed by a subset of \mathbb{R} (an uncountably infinite

index set). The main focus there is not inference but to choose a tuning parameter that asymptotically leads to an estimator with the “smallest” variance and to take into account this randomness in proving that final estimator with estimated tuning parameter has an asymptotic normal distribution with the “smallest” variance. This kind of tuning is very useful when using robust estimators since the statistician does not want to lose on asymptotic efficiency of the maximum likelihood estimator in case there is no contamination.

We now show that assumptions (B1) and (B2) imply a central limit theorem for $(t_{j \cdot M})_{M \in \mathcal{M}, j \in M}$ whenever $(n^{-1/2} \sum_{i=1}^n \psi_{j \cdot M}(Z_i))_{M \in \mathcal{M}, j \in M}$ satisfies a central limit theorem. Here the t -statistics $t_{j \cdot M}$ are given by

$$t_{j \cdot M} := \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}}.$$

Define a Gaussian random vector $(G_{j \cdot M})_{M \in \mathcal{M}, j \in M}$ satisfying $G_{j \cdot M} \sim N(0, 1)$ and

$$\text{Cov}(G_{j \cdot M}, G_{j' \cdot M'}) = \text{Cov} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}}, \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\psi_{j' \cdot M'}(Z_i)}{\sigma_{j' \cdot M'}} \right).$$

Define

$$\mathcal{Q} := \{(j, M) : M \in \mathcal{M}, j \in M\},$$

and $\Delta_{\mathcal{M}}^{\text{CLT}}$ as the error in the central limit theorem given by

$$\sup_{\substack{(a_{j \cdot M}), \\ (b_{j \cdot M})}} \left| \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \left\{ a_{j \cdot M} \leq \sum_{i=1}^n \frac{\psi_{j \cdot M}(Z_i)}{\sqrt{n} \sigma_{j \cdot M}} \leq b_{j \cdot M} \right\} \right) - \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \{a_{j \cdot M} \leq G_{j \cdot M} \leq b_{j \cdot M}\} \right) \right|.$$

Theorem 13. Under assumptions (B1) and (B2), for any $\delta_1, \delta_2 \in (0, 1)$,

$$\begin{aligned} & \sup_{\substack{(a_{j \cdot \mathcal{M}}), \\ (b_{j \cdot \mathcal{M}})}} \left| \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \right) - \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} \leq \frac{\sigma_{j \cdot \mathcal{M}} G_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \leq b_{j \cdot \mathcal{M}} \right\} \right) \right| \\ & \leq 4\Delta_{\mathcal{M}}^{\text{CLT}} + \mathbb{P}(\Delta_{\mathcal{M}}^{\text{ULR}} > \delta_1) + \mathbb{P}(\Delta_{\mathcal{M}}^{\text{Var}} > \delta_2) + \frac{24\sqrt{\log(e|\mathcal{Q}|)}}{\underline{\sigma}(1 - \delta_2)} \left[\delta_2 \sqrt{\log \left(\frac{|\mathcal{Q}|}{\Delta_{\mathcal{M}}^{\text{CLT}}} \right)} + \delta_1 \right]. \end{aligned}$$

Proof. See Appendix 5.A for a proof. \square

Controlling the Bound in Theorem 13. Theorem 13 is deterministic in nature, in the sense that we do not need any specific randomness assumptions on the data. Whatever is the nature of dependence between the observations, if we can verify that $\Delta_{\mathcal{M}}^{\text{CLT}} = o(1)$ and $\Delta_{\mathcal{M}}^{\text{Var}} \log(|\mathcal{Q}|) = o_p(1)$ and $\Delta_{\mathcal{M}}^{\text{ULR}} \sqrt{\log(|\mathcal{Q}|)} = o_p(1)$, then Theorem 13 implies that the set of all t -statistics will behave like Gaussian random variables (asymptotically).

The quantity $\Delta_{\mathcal{M}}^{\text{CLT}}$ can be bounded using many of the recent results in high-dimensional central limit theorem literature. We refer to Chernozhukov et al. (2017a); Koike (2019); Deng and Zhang (2017); Zhang and Wu (2017); Belloni et al. (2018); Zhang et al. (2018); Kuchibhotla et al. (2018); Chernozhukov et al. (2019); Fang and Koike (2020) just to mention a few. Most of these papers provide the results for the case of the average of mean zero random vectors with dimension almost exponential in the sample size. The notable exceptions are Zhang and Wu (2017); Zhang et al. (2018) which deal with the case of dependent random vectors using the notion of dependence developed by Wu (2005).

All the papers cited above work with the case of finite $|\mathcal{Q}|$. The way Theorem 13 is derived also uses this fact. Because of this the bound for $|\mathcal{Q}| = \infty$ is obsolete. We mention, however, that some extensions of the central limit theorems are available for the case of infinite dimensional spaces such as Banach spaces; see Chernozhukov et al.

(2014); Paulauskas and Rackauskas (1989); Statulevicius (2000) for some results. Theorem 13 can be extended to this case but we will not provide this extension but will refer to Kuchibhotla et al. (2018) for some related discussion.

The control of $\Delta_{\mathcal{M}}^{\text{ULR}}$ and $\Delta_{\mathcal{M}}^{\text{Var}}$ often is closely related “smoothness” of the estimator as a function of the empirical distribution of the data. We stress here that Theorem 13 holds true for arbitrary estimators $\hat{\beta}_{\mathcal{M}}$ and need not be defined through (5.2). In the following section we will provide control of $\Delta_{\mathcal{M}}^{\text{ULR}}$ and $\Delta_{\mathcal{M}}^{\text{Var}}$ for the case of linear and generalized linear regression model estimators. \square

From Theorem 13 readily yields that if $(a_{j \cdot \mathcal{M}})$ and $(b_{j \cdot \mathcal{M}})$ are chosen to satisfy

$$\mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} \leq \frac{\sigma_{j \cdot \mathcal{M}} G_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \leq b_{j \cdot \mathcal{M}} \right\} \right) \geq 1 - \alpha, \quad (5.10)$$

then asymptotically

$$\mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{ a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}} \} \right) \geq 1 - \alpha + o(1).$$

Although written $+o(1)$, it should be noted that this error could be positive or negative. Hence the finite sample coverage could be above or below $1 - \alpha$. There are two questions that follow:

1. Because the distribution of the data is unknown, we do not know the covariance operator of $(G_{j \cdot \mathcal{M}})$. How does one find the vectors $(a_{j \cdot \mathcal{M}})$, $(b_{j \cdot \mathcal{M}})$ satisfying (5.10).
2. Even if we assume that the distribution of $(\sigma_{j \cdot \mathcal{M}} G_{j \cdot \mathcal{M}} / \sigma_{j \cdot \mathcal{M}}^*)$ is known, what is the right form of $(a_{j \cdot \mathcal{M}})$ and $(b_{j \cdot \mathcal{M}})$? Being a multivariate distribution the quantiles are not unique and there are finite set of pair vectors $(a_{j \cdot \mathcal{M}})$, $(b_{j \cdot \mathcal{M}})$ such that (5.10) holds true.

The first question is relatively easier to answer through bootstrap. In the case $|\mathcal{Q}| = 1$ (the classical inference setting), we can estimate the distribution of the limiting Gaussian using bootstrap and this is what we follow to answer the first question. We first note that we cannot estimate the exact distribution of $(\sigma_{j\cdot M}G_{j\cdot M}/\sigma_{j\cdot M}^*)$ under the assumptions above because $\sigma_{j\cdot M}$ could not be estimated. The inflation factor $\sigma_{j\cdot M}^*/\sigma_{j\cdot M}$ cannot be estimated in general. This is similar to the conservativeness mentioned in Chapter 3 for the case of fixed covariate assumption-lean linear regression case.

We now describe the bootstrap algorithm that approximates the distribution of $(G_{j\cdot M})$. The following description is applicable only for the case of independent observations Z_1, \dots, Z_n . For the dependent case, we refer to [Zhang et al. \(2018\)](#).

Pseudocode:

1. Fix $B \geq 1$ the number of bootstrap replications. Generate mean zero variance one (real-valued) random variables e_1^b, \dots, e_n^b for $1 \leq b \leq B$. For example, e_i^b could be standard Gaussian or Rademacher random variables or Mammen's golden ratio random variable. We refer the reader to [Deng and Zhang \(2017\)](#) for details.
2. Compute the bootstrap version of $t_{j\cdot M}$ as

$$t_{j\cdot M}^{(b)} := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i^b \times \frac{\hat{\psi}_{j\cdot M}(Z_i)}{\hat{\sigma}_{j\cdot M}}, \quad \text{for } 1 \leq b \leq B.$$

Note that for all $(j, M) \in \mathcal{Q}$ the bootstrap version of $t_{j\cdot M}$ share the same multiplier e_i^b . This is done to preserve the dependence between different coordinates of $(t_{j\cdot M})$.

3. Report the empirical distribution of $(t_{j\cdot M}^{(b)})_{(j, M)}, 1 \leq b \leq B$.

Define

$$\mathcal{D}_n := \{Z_1, \dots, Z_n\},$$

and the Gaussian process $(G_{j \cdot M}^{\text{Boot}})$ satisfying

$$\mathbb{E}[G_{j \cdot M}^{\text{Boot}}] = 0, \quad \text{Cov}(G_{j \cdot M}^{\text{Boot}}, G_{j' \cdot M'}^{\text{Boot}}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right].$$

Note that such a Gaussian process exists because

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right] &= \text{Cov} \left(\frac{\sigma_{j \cdot M}}{\sigma_{j \cdot M}^*} G_{j \cdot M}, \frac{\sigma_{j' \cdot M'}}{\sigma_{j' \cdot M'}^*} G_{j' \cdot M'} \right) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}^*} \right] \mathbb{E} \left[\frac{\psi_{j' \cdot M'}(Z_i)}{\sigma_{j' \cdot M'}^*} \right]. \end{aligned}$$

Theorem 14 (Bootstrap Approximation). *Suppose e_1^1, \dots, e_n^1 are independent standard Gaussian random variables and assumption (B3) holds true. Further suppose that Z_1, \dots, Z_n are independent. Then on the event $\{\Delta_{\mathcal{M}}^{\text{Var}} \leq 1/2\} \cap \{\Delta_{\mathcal{M}}^{\text{Inf}} \leq 1\} \cap \{\Delta_{\mathcal{M}}^{\text{Boot}} \leq 1\}$, we have*

$$\begin{aligned} &\sup_{(a_{j \cdot M}), (b_{j \cdot M})} \left| \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \{a_{j \cdot M} \leq t_{j \cdot M}^{(1)} \leq b_{j \cdot M}\} \middle| \mathcal{D}_n \right) - \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \{a_{j \cdot M} \leq G_{j \cdot M}^{\text{Boot}} \leq b_{j \cdot M}\} \right) \right| \\ &\leq C \left(\Delta_{\mathcal{M}}^{\text{Boot}} + 12\Delta_{\mathcal{M}}^{\text{Var}} + 16(\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} \right)^{1/3} (\log(|\mathcal{Q}|))^{2/3}, \end{aligned}$$

for an absolute constant $C > 0$. Further, for all $(b_{j \cdot M}) \geq 0$,

$$\begin{aligned} &\inf_{(b_{j \cdot M})} \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \left\{ -b_{j \cdot M} \leq \frac{\sigma_{j \cdot M} G_{j \cdot M}}{\sigma_{j \cdot M}^*} \leq b_{j \cdot M} \right\} \right) - \mathbb{P} \left(\bigcap_{(j, M) \in \mathcal{Q}} \left\{ -b_{j \cdot M} \leq t_{j \cdot M}^{(1)} \leq b_{j \cdot M} \right\} \middle| \mathcal{D}_n \right) \\ &\geq -C \left(\Delta_{\mathcal{M}}^{\text{Boot}} + 12\Delta_{\mathcal{M}}^{\text{Var}} + 16(\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} \right)^{1/3} (\log(|\mathcal{Q}|))^{2/3}. \end{aligned}$$

Proof. See Appendix 5.B for a proof. □

Comments on Theorem 14. Unlike Theorem 13, Theorem 14 requires an assumption on the independence of Z_1, \dots, Z_n . Similar the obstacles encountered in the case of linear regression under fixed covariates, we get that the bootstrap distribution does not approximate the limiting Gaussian distribution of the t -statistics, but it converges to a distribution with a “larger” covariance operator. Fortunately, this allows us to perform valid inference albeit with some conservativeness. This conservativeness is shown in the second part of the result and formally, this part proves asymptotically,

$$\mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \{-b_{j \cdot M} \leq t_{j \cdot M} \leq b_{j \cdot M}\} \right) \geq \mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \{-b_{j \cdot M} \leq t_{j \cdot M}^{(1)} \leq b_{j \cdot M}\} \middle| \mathcal{D}_n \right) + o_p(1).$$

Hence finding $(b_{j \cdot M}) \geq 0$ such that the bootstrap coverage probability is $1 - \alpha$ yields a valid post-selection confidence region.

In order to apply Theorem 14, we need to show convergence to zero of $\Delta_{\mathcal{M}}^{\text{Var}}$, $\Delta_{\mathcal{M}}^{\text{Inf}}$, and $\Delta_{\mathcal{M}}^{\text{Boot}}$. The first one is also required for Theorem 13. Having consistent estimators of influence functions is also a requirement in the case of $|\mathcal{Q}| = 1$ (the classical inference setting). We will describe such influence function estimators for the case of the commonly used regression estimators in the following section. Bounding the quantity $\Delta_{\mathcal{M}}^{\text{Boot}}$ follows just from concentration inequalities for the averages of independent random vectors. These are similar to the bounds used for Lemma 5 in Chapter 4.

Finally, we remark that Theorem 14 is not the end of story corresponding to the application of bootstrap. We do not usually have access to the conditional distribution of bootstrap statistics because of computational constraints and hence one often uses the proxy

$$\frac{1}{B} \sum_{b=1}^B \mathbb{1} \left\{ \bigcap_{(j,M) \in \mathcal{Q}} \{-b_{j \cdot M} \leq t_{j \cdot M}^{(1)} \leq b_{j \cdot M}\} \right\},$$

where $B < \infty$ is a finite number representing the number of bootstrap replications. For general vectors $(b_{j,\mathcal{M}})$, the distance between the empirical bootstrap distribution (above) and the true bootstrap distribution is of order $\sqrt{|\mathcal{Q}|/B}$ because the VC dimension of the class of all hyperrectangles in $\mathbb{R}^{|\mathcal{Q}|}$ is of order $|\mathcal{Q}|$. This bound is not so useful for our purposes because $|\mathcal{Q}|$ is much larger than the sample size (often almost exponential) and choosing B that large is not feasible. We will discuss specific type of post-selection confidence regions in later sections and for these sets, it is relatively easy to prove a sharper bound between the empirical and the true bootstrap distribution. \square

5.2 Application to Linear Regression

In this section, we verify assumptions required for Theorems 13 and 14 for the case of linear regression. Following the discussion in Chapter 3, we verify these assumptions in the unified framework there by just assuming independent observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^p \times \mathbb{R}$. The discussion below is mostly based on [Kuchibhotla et al. \(2018\)](#); [Kuchibhotla et al. \(2019\)](#). In this section, we will restrict to the case of linear regression and note that similar results hold for the case of generalized linear models ([Kuchibhotla, 2018](#)). Define the least squares estimator and the target as

$$\begin{aligned}\hat{\beta}_{\mathcal{M}} &:= \arg \min_{\theta \in \mathbb{R}^{|\mathcal{M}|}} \frac{1}{n} \sum_{i=1}^n (Y_i - X_{i,\mathcal{M}}^\top \theta)^2, \\ \beta_{\mathcal{M}} &:= \arg \min_{\theta \in \mathbb{R}^{|\mathcal{M}|}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(Y_i - X_{i,\mathcal{M}}^\top \theta)^2].\end{aligned}$$

We work with the following assumptions. We write S^{d-1} to denote the unit sphere

in \mathbb{R}^d . Define matrices

$$\Sigma_M := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_{i,M} X_{i,M}^\top], \quad \text{and} \quad V_M := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_{i,M} X_{i,M}^\top (Y_i - X_{i,M}^\top \beta_M)^2].$$

(DGP) The observations $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \leq i \leq n$ are independent.

(E)(q) There exists some $q \geq 2$ and a constant $K_q \in (0, \infty)$ such that

$$\left(\mathbb{E}[|Y_i|^q] \right)^{1/q} \leq K_q < \infty, \quad \text{for all } 1 \leq i \leq n.$$

(X-SW) There exists a constant $K_x \in (0, \infty)$ such that

$$\mathbb{E} \left[\exp \left(\frac{|u^\top \Sigma_M^{-1/2} X_{i,M}|^\beta}{K_x^\beta} \right) \right] \leq 2, \quad \text{for all } 1 \leq i \leq n \text{ and } u \in S^{|\mathbb{M}|-1}.$$

(Σ-V) There exist constants $0 < \underline{\lambda} \leq \bar{\lambda} < \infty$ such that

$$\underline{\lambda} \leq \lambda_{\min}(\Sigma_M^{1/2} V_M^{-1} \Sigma_M^{1/2}) \leq \lambda_{\max}(\Sigma_M^{1/2} V_M^{-1} \Sigma_M^{1/2}) \leq \bar{\lambda}.$$

We now provide some comments on these assumptions. Condition **(DGP)** requires observations to be independent but for some of the results we do not even need this assumption. Condition **(E)(q)** requires the existence of q -th order moment of the response Y_i . Condition **(X-SW)** is a rewording of K_x -sub-Weibull property of X_1, \dots, X_n and the condition necessarily requires $K_x \geq 1$. Condition **(Σ-V)** requires Σ_M and V_M to be of the “same order”. Note that

$$V_M = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_{i,M} X_{i,M}^\top (Y_i - X_{i,M}^\top \beta_M)^2] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_{i,M} X_{i,M}^\top \mathbb{E}[(Y_i - X_{i,M}^\top \beta_M)^2 | X_i]].$$

Hence condition $(\Sigma\text{-}V)$ is satisfied if

$$\bar{\lambda}^{-1} \leq \inf_x \mathbb{E}[(Y_i - X_{i,M}^\top \beta_M)^2 | X_i = x] \leq \sup_x \mathbb{E}[(Y_i - X_{i,M}^\top \beta_M)^2 | X_i = x] \leq \underline{\lambda}^{-1}.$$

Here \inf_x and \sup_x should be taken as essential infimum and supremum with respect to the distribution of X_i .

The following theorem provides a control of $R_{j,M}$ in Assumption $(B3)$. Interestingly, we only need Assumption $(\Sigma\text{-}V)$ and do not require the independence assumption. For $j \in M$, let $\sigma_{j,M}^2$ denotes the diagonal entry of $\Sigma_M^{-1} V_M \Sigma_M^{-1}$ corresponding to the j -th covariate. For instance, if $M = \{2, 3\}$, then $\sigma_{2,M}^2$ denotes the first diagonal entry of $\Sigma_M^{-1} V_M \Sigma_M^{-1}$ and $\sigma_{3,M}$ denotes the second diagonal entry. Further define

$$\mathcal{D}_M^\Sigma := \|\Sigma_M^{-1/2}(\hat{\Sigma}_M - \Sigma_M)\Sigma_M^{-1/2}\|_{op}.$$

Theorem 15. *Under Assumption $(\Sigma\text{-}V)$, simultaneously for all $M \subseteq \{1, 2, \dots, p\}$,*

$$n^{1/2}(\hat{\beta}_{j,M} - \beta_{j,M}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j,M}(Z_i) + R_{j,M},$$

such that

$$\max_{j \in M} \left| \frac{R_{j,M}}{\sigma_{j,M}} \right| \leq \sqrt{\frac{\bar{\lambda}}{\underline{\lambda}}} \frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)_+} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n V_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) \right\|_2,$$

where

$$\psi_M(Z_i) := \Sigma_M^{-1} X_{i,M} (Y_i - X_{i,M}^\top \beta_M),$$

and $\psi_{j,M}(\cdot)$ denotes the coordinate of $\psi_M(\cdot)$ corresponding to the covariate j in M .

Proof. See Appendix 5.C for a proof. This result appeared in [Kuchibhotla et al. \(2019\)](#) albeit with a minor mistake. I thank Prof. Alessandro Rinaldo of CMU for

pointing this mistake. \square

Remark 5.2.1 (Comments on Theorem 15). As mentioned before, Theorem 15 is deterministic. It does not require any independence/dependence assumptions on the observations and holds for any realization of the data. To bound the error $R_{j,M}$ we need to control \mathcal{D}_M^Σ and the average of $V_M^{-1/2} X_{i,M}(Y_i - X_{i,M}^\top \beta_M)$ which we will under the assumptions listed above. We refer the reader to [Kuchibhotla et al. \(2018\)](#) for control of these terms under dependence. \diamond

In the following theorem, we control the terms from Theorem 15. Define the kurtosis and “regression variance” for model M as

$$\kappa_M^\Sigma := \max_{\theta \in \mathbb{R}^{|\mathbb{M}|}} \frac{1}{n} \sum_{i=1}^n \frac{\text{Var}((X_{i,M}^\top \theta)^2)}{\|\Sigma_M^{1/2} \theta\|^4} \quad \text{and} \quad \mathfrak{V}_M := \max_{\substack{\theta \in \mathbb{R}^{|\mathbb{M}|}, \\ \|\theta\|=1}} \frac{1}{n} \sum_{i=1}^n \text{Var}(\theta^\top \Sigma_M^{-1/2} X_{i,M} Y_i).$$

Theorem 16. Fix any $t \geq 0$. Under **(X-SW)**, we have with probability at least $1 - 3e^{-t}$, simultaneously for any $1 \leq s \leq p$, for any $\mathbb{M} \subseteq \{1, \dots, p\}$ with $|\mathbb{M}| = s$,

$$\mathcal{D}_M^\Sigma \leq 14 \sqrt{\frac{\kappa_M^\Sigma (t + s \log(9e^2 p/s))}{n}} + \frac{C_\beta K_x^2 (\log(2n))^{2/\beta} (t + s \log(9e^2 p/s))^{\max\{1, 2/\beta\}}}{n}. \quad (5.11)$$

If **(X-SW)** and **(E)(q)** hold true, then with probability at least $1 - 3e^{-t_1} - t_2^{-q+1}$, for any $1 \leq s \leq p$, for any model $\mathbb{M} \subseteq \{1, \dots, p\}$ with $|\mathbb{M}| = s$,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) \right\| &\leq 14 \sqrt{\frac{\mathfrak{V}_M (t_1 + s \log(5e^2 p/s))}{n}} + \mathcal{D}_M^\Sigma \left(\sum_{i=1}^n \mathbb{E}[Y_i^2]/n \right)^{1/2} \\ &\quad + \frac{C_\beta K_q K_x (\log(2n))^{1/\beta} (t_2 + s \log(5e^2 p/s))^{\max\{1, 1/\beta\}}}{n^{1-1/q}} \\ &\quad + \frac{t_2 C_{\beta,q} K_q K_x (s \log(5e^2 p/s) + \log n)^{1/\beta}}{n^{1-1/q}}, \quad (5.12) \end{aligned}$$

for some constants $C_\beta, C_{\beta,q} > 0$ depending only on β and (β, q) , respectively.

Proof. See Appendix 5.D for a proof. The result here is essentially Proposition 5.1 of Kuchibhotla et al. (2019) but with a change for the second part where we are now using t_1 and t_2 instead of the same t . The proof is almost verbatim. \square

It is worth mentioning that the rate bound in Theorem 16 can possibly be improved in the second order terms; see Theorem 3.1 of Guédon et al. (2015).

Remark 5.2.2 (Comments on Theorem 16 and Verification of (B1)). The discussion below will be assuming both (X-SW) and (E)(q). The probability of error for \mathcal{D}_M^Σ is exponential and hence taking $t = \log(n/3)$ yields with probability at least $1 - 1/n$,

$$\mathcal{D}_M^\Sigma \leq C \left[\sqrt{\frac{s \log(epn/s)}{n}} + \frac{(\log n)^{2/\beta} (s \log(epn/s))^{\max\{1, 2/\beta\}}}{n} \right],$$

taking the quantities like κ_M^Σ and K_x into C . On the other hand, the probability of error for the second quantity only decreases polynomially. Hence, assuming $q \geq 2$ and taking $t_1 = \log(n/3)$ and $t_2 = n^{1/q}$, we get with probability at least $1 - n^{-1} - n^{-1+1/q}$,

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) \right\| &\leq C \sqrt{\frac{s \log(epn/s)}{n}} + \mathcal{D}_M^\Sigma \left(\sum_{i=1}^n \mathbb{E}[Y_i^2]/n \right)^{1/2} \\ &\quad + C \frac{(\log n)^{1/\beta} (s \log(epn/s))^{\max\{1, 1/\beta\}}}{n^{1-1/q}} \\ &\quad + C \frac{(s \log(epn/s))^{1/\beta}}{n^{1-2/q}}. \end{aligned}$$

Essentially the last term is the most dominating and for it converge to zero, we need q to be strictly larger than 2. To control Δ_M^{Var} , we will require $q > 4$ and in this case, the last term can be bounded by $(s \log(epn/s))^{1/\beta}/n^{1/2}$. Combining these two results along with Theorem 15, we get that with probability at least $1 - 6n^{-1} - n^{-1+1/q}$,

simultaneously over all $M \subseteq \{1, 2, \dots, p\}$,

$$\max_{j \in \mathcal{M}} |R_{j \cdot M}| \leq C \frac{(|M| \log(epn/|M|))^{1/2+1/\beta}}{n^{1/2}}.$$

This implies that we can take $\delta_1 = C \max_{M \in \mathcal{M}} (|M| \log(epn/|M|))^{1/2+1/\beta} / n^{1/2}$ and get

$$\mathbb{P}(\Delta_{\mathcal{M}}^{\text{ULR}} > \delta_1) \leq \frac{6}{n} + \frac{1}{n^{1-1/q}}.$$

◇

Having settled Assumption (B1), we now proceed to Assumptions (B2) and (B3). Following the discussion surrounding (5.9), we proceed to verifying Assumption (B3). From Theorem 15, recall that

$$\psi_M(Z_i) = \Sigma_M^{-1} X_{i \cdot M} (Y_i - X_{i \cdot M}^\top \beta_M).$$

A natural estimator of $\psi_{j \cdot M}(\cdot)$ is given by

$$\hat{\psi}_M(Z_i) := \hat{\Sigma}_M^{-1} X_{i \cdot M} (Y_i - X_{i \cdot M}^\top \hat{\beta}_M).$$

Based on this, we consider

$$\hat{\sigma}_{j \cdot M}^2 := \frac{1}{n} \sum_{i=1}^n \hat{\psi}_{j \cdot M}^2(Z_i),$$

where $\psi_{j \cdot M}(\cdot)$ is the coordinate of $\psi_M(\cdot)$ corresponding to the covariate j . Note that

$$\sigma_{j \cdot M}^2 = \text{Var} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j \cdot M}(Z_i) \right) = \frac{1}{n} \sum_{i=1}^n \text{Var}(\psi_{j \cdot M}(Z_i)) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\psi_{j \cdot M}^2(Z_i)] = (\sigma_{j \cdot M}^*)^2,$$

the right hand side of which is the target of $\hat{\sigma}_{j \cdot M}^2$. The following lemma provides a

bound on the accuracy of influence function estimator all models M .

Lemma 10. *Under Assumption $(\Sigma-V)$, for any $M \subseteq \{1, 2, \dots, p\}$,*

$$\begin{aligned} \max_{j \in M} \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}} \right|^2 &\leq \mathcal{D}_M^\Sigma \sqrt{\frac{\bar{\lambda}}{n} \sum_{i=1}^n \|\Sigma_M^{1/2} \psi_M(Z_i)\|^2} \\ &+ (1 + \mathcal{D}_M^\Sigma) \|\hat{\beta}_M - \beta_M\|_{\Sigma_M} \sqrt{\frac{\bar{\lambda}}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^4}. \end{aligned}$$

Proof. See Appendix 5.E for a proof. □

Similar to Theorem 15, Lemma 10 is also deterministic and does not require any independence/dependence assumptions on Z_1, \dots, Z_n . Theorem 16 already bounds \mathcal{D}_M^Σ under the independence assumption. Using Theorem 15 allows us to bound $\|\hat{\beta}_M - \beta_M\|_{\Sigma_M}$:

$$\|\hat{\beta}_M - \beta_M\|_{\Sigma_M} \leq \frac{1}{1 - \mathcal{D}_M^\Sigma} \left\| \frac{1}{n} \sum_{i=1}^n \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) \right\|_2.$$

See (5.35) in the proof of Theorem 15 for details. Both the terms on the right hand side are bounded from Theorem 16. We only need to bound

$$\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2} \psi_M(Z_i)\|^2 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^4. \quad (5.13)$$

To control these terms we use the independence as well as the tail assumptions.

Firstly, we note the following inequality:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2} \psi_M(Z_i)\|^2 &= \frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M)\|^2 \\
&\leq \frac{2}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2 + \frac{2}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M} X_{i,M}^\top \Sigma_M^{-1/2} \Sigma_M^{1/2} \beta_M\|^2 \\
&\leq \frac{2}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2 + \frac{2}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^4 \|\Sigma_M^{1/2} \beta_M\|^2.
\end{aligned}$$

Because $\|\Sigma_M^{1/2} \beta_M\|^2 \leq n^{-1} \sum_{i=1}^n \mathbb{E}[Y_i^2]$ (see (4.32) in Chapter 4), the second term is controlled by a constant multiple of (5.13).

Theorem 17. *Under Assumptions (DGP), (E)(q), (X-SW), we have with probability at least $1 - 2/n$*

$$\max_{\substack{1 \leq s \leq p, \\ |M|=s}} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2}{|M| \lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))} \leq CK_x^2 K_q^2 \left[1 + \frac{(\log(ep))^{\frac{2}{\beta}} \log n}{n^{1-4/q}} \left(\frac{\log(pn)}{n^{2/q}} + 1 \right) \right].$$

With probability at least $1 - 1/(pn)$,

$$\max_{\substack{1 \leq s \leq p, \\ |M|=s}} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^4}{|M|^2 \lambda_{\max}^2(\text{Cor}(\Sigma_M^{-1}))} \leq CK_x^4 \left[1 + \frac{(\log(3pn))^{1+4/\beta} (\log(en))^{4/\beta}}{n} \right].$$

With probability at least $1 - 1/n - 1/\sqrt{n}$,

$$\Delta_{\mathcal{M}}^{\text{Boot}} \leq C \sqrt{\frac{\mathfrak{B}_{\mathcal{M}} \log(n|\mathcal{Q}|)}{n}} + C \frac{\bar{\lambda} K_x^4 K_q^2 (\log(|\mathcal{Q}|))^{2/\beta} (\log(n|\mathcal{Q}|))^{2/\beta}}{n^{1-3/q}} \left[1 + \frac{(\log(|\mathcal{Q}|))^{1-2/q}}{n^{1/q}} \right],$$

where

$$\mathfrak{B}_{\mathcal{M}} := \max_{\substack{j \in M \in \mathcal{M}, \\ j' \in M' \in \mathcal{M}}} \frac{1}{n} \sum_{i=1}^n \text{Var} \left(\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} \right).$$

Proof. See Appendix 5.F for a proof. □

With this result, we have verified all the assumptions required to show the validity of bootstrap for valid post-selection inference. The main conclusion is that no linear model or distributional assumptions are necessary for validity, we only need light tail assumptions on covariates as well as q -th moment assumptions on the response. The number of moments of the response can be as small as 4. Of course, the tail assumptions change the dimension requirements for convergence guarantees; the smallest exponent $\gamma > 0$ such that $\log^\gamma(|\mathcal{Q}|) = o(n)$ for validity.

5.3 On the Shape of Intervals for Valid Post-selection Inference

Having completed the question of approximating the true distribution of the vector of t -statistics, we now discuss the question of the choosing $(a_{j,\mathcal{M}})$ and $(b_{j,\mathcal{M}})$. As we mentioned, there are multiple choices that guarantee the coverage of $(1 - \alpha)$. In this section, we will discuss the classical choice based on $\max\text{-}|t|$ and then discuss a few desiderata for the confidence intervals leading to powerful inference.

In previous sections, we have not assumed any special structure on \mathcal{M} and this was intentional because of applicability for interaction models as well as experimental designs. In this section, we will consider some specific examples of \mathcal{M} :

$$\begin{aligned} \mathcal{M}_p(k) &:= \{M \subseteq \{1, 2, \dots, p\} : 1 \leq |M| \leq k\}, \\ \mathcal{M}_p(k; 1) &:= \{M \subseteq \{1, 2, \dots, p\} : 1 \leq |M| \leq k, 1 \in M\}. \end{aligned} \tag{5.14}$$

The first collection represents the set of all k -sparse models (that is, the set of all subsets with at most k variables). The second collection represents the set of all k -sparse models that contain the first covariate; this is interesting in the context of causal/treatment effect.

Berk et al. (2013); Bachoc et al. (2016) construct valid PoSI regions $\widehat{\mathcal{R}}_{\mathcal{M}}$ satisfying (5.5) based on quantiles of the “max-t” statistic

$$\max_{\mathcal{M} \in \mathcal{M}} \max_{j \in \mathcal{M}} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot \mathcal{M}} - \beta_{j \cdot \mathcal{M}})}{\widehat{\sigma}_{j \cdot \mathcal{M}}} \right|. \quad (5.15)$$

Equivalently this yields

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\beta_{j \cdot \widehat{\mathcal{M}}} \in \left[\widehat{\beta}_{j \cdot \widehat{\mathcal{M}}} \pm K_{\alpha} \frac{\widehat{\sigma}_{j \cdot \widehat{\mathcal{M}}}}{n^{1/2}} \right] \text{ for all } j \in \widehat{\mathcal{M}} \right) \geq 1 - \alpha, \quad (5.16)$$

where $[a \pm b] = [a - b, a + b]$ for any two reals a, b .

The definition (5.15) of the “max-t” statistic shows that it gives equal importance to models of all sizes and all covariates of all models irrespective of their sizes. For instance, models of size 1 are given the same importance of models of size k which is not preferable. We will discuss several disadvantages of the “max-t” statistic and propose an alternative based on the ideas of pre pivoting as well as balanced simultaneous confidence intervals of Beran (1987, 1988).

5.3.1 Disadvantages of the “max-t” Statistic

The “max-t” statistic (5.15) is a natural generalization of inference for a single model to simultaneous inference over a collection of models. The maximum statistic would be the right thing to do if we are concerned with simultaneous inference for p parameters (all of which are of same order) but this is not the case with OLS under variable selection. It is intuitively expected that models with more number of covariates would have larger width intervals. For this reason by taking the maximum over the collection \mathcal{M} of models, one is ignoring the smaller models and the fact that small models have smaller width confidence intervals. To be concrete, if \mathcal{M} is $\mathcal{M}_p(k)$

it follows from the results of Berk et al. (2013) that

$$\max_{M \in \mathcal{M}_p(k)} \max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| = O_p(\sqrt{k \log(ep/k)}), \quad (5.17)$$

and in the worst case this rate can be attained. But if $k = 40$ (for example) but the selected model \hat{M} happened to have only two covariates, then the confidence interval is (unnecessarily) wider by a factor of $\sqrt{20}$. Allowing for a model dependent quantile $K_{\hat{M}}(\alpha)$ (instead of K_α independent of \hat{M}) can tighten the confidence intervals appropriately. To illustrate this disadvantage in practice, we consider the telomere length example that we described in Section 1.2.4 of Chapter 1. In this data, there are 21 total covariates (including the pairwise interactions). For illustration, we select one covariate that is most correlated with the response and require inference for the slope in this simple linear regression. For this selection, it is enough to take $k = 1$ but if we were to set k larger than 1, then the quantile of the $\max\text{-}|t|$ changes as shown in Figure 5.1.

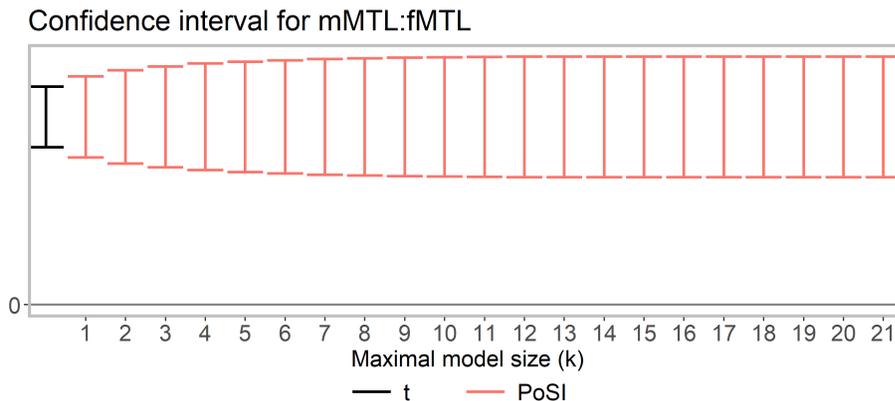


Figure 5.1: Illustrating the dependence on k of the $\max\text{-}|t|$ statistic: Telomere length analysis. The most correlated covariate in this data is the interaction between the telomere lengths of the parents.

For this particular disadvantage, it is enough to have $K_{\hat{M}}(\alpha)$ depend on \hat{M} only

through $|\widehat{M}|$, its size.

There is a second disadvantage of the maximum statistic that requires dependence of $K_{\widehat{M}}(\alpha)$ on the covariates in \widehat{M} . To describe this second disadvantage we look at the conditions under which worst case rate in (5.17) is attained when $k = p$. Berk et al. (2013, Section 6.2) shows that if the covariates are non-stochastic, and

$$\frac{1}{n} \sum_{i=1}^n X_i X_i^\top := \begin{bmatrix} I_{p-1} & c \mathbf{1}_{p-1} \\ c \mathbf{1}_{p-1}^\top & 1 \end{bmatrix}, \text{ for some } c^2 < 1/(p-1), \quad (5.18)$$

then there exists a constant $\mathfrak{C} > 0$, such that with high probability

$$\max_{M \in \mathcal{M}_p(p)} \max_{j \in M} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\widehat{\sigma}_{j \cdot M}} \right| \geq \mathfrak{C} \sqrt{p}. \quad (5.19)$$

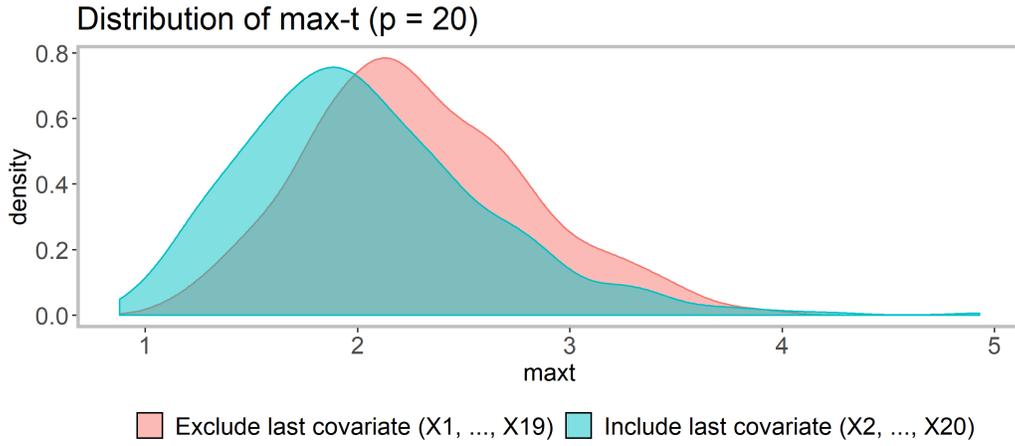
Now define $\mathcal{M} = \{M \subseteq \{1, \dots, p\} : p \notin M\}$, that is, \mathcal{M} is the collection of models that only contain the first $p - 1$ covariates. It now follows from (Berk et al., 2013, Section 6.1) that

$$\max_{M \in \mathcal{M}} \max_{j \in M} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\widehat{\sigma}_{j \cdot M}} \right| \asymp O_p(1) \sqrt{\log(ep)}. \quad (5.20)$$

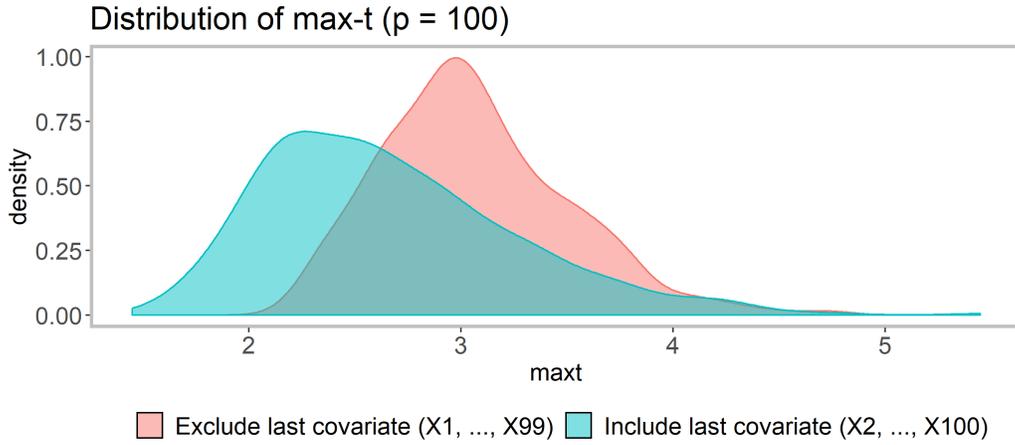
Comparing (5.19) and (5.20), it is clear that the inclusion of the last covariate increases the order of the maximum statistic from $\sqrt{\log(ep)}$ to \sqrt{p} ; this shift is because of increased collinearity. To illustrate this, we compare the maximum of the t -statistics within two models of same size: one including the last covariate and one excluding the last covariate. Figure 5.2 shows the distribution of the $\max\text{-}|t|$ in each model, that is, we are plotting the distributions of

$$\max_{j \in M} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\widehat{\sigma}_{j \cdot M}} \right|,$$

for M including and excluding the last covariate.



(a) $p = 20$



(b) $p = 100$

Figure 5.2: Distribution of max- t for one model that includes the last covariate and one that excludes that last covariate for $p = 20, 100$ under Setting (5.18).

This means that if in the selection procedure we allow all models but end up choosing the model that only contains the first $p - 1$ covariates, we pay of lot more price than necessary. Note that if p increases with n , this increase (in rate) could hurt more. Once again allowing for $K_{\hat{M}}(\alpha)$ a model dependent quantile for maximum (over $j \in \hat{M}$) in that model resolves this disadvantage.

In conclusion, we want to construct confidence regions $\widehat{\mathcal{R}}_M, M \in \mathcal{M}$ of the form:

$$\widehat{\mathcal{R}}_M := \left\{ \theta \in \mathbb{R}^{|\mathcal{M}|} : \max_{j \in \mathcal{M}} \left| \frac{n^{1/2}(\widehat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\widehat{\sigma}_{j \cdot M}} \right| \leq K_M(\alpha) \right\}, \quad (5.21)$$

for some constant $K_M(\alpha)$ that depends on the model M and satisfying the simultaneous inference guarantee (5.5). At this point, we emphasize that we are discussing the generic case where there is no specific variable of interest. If the analyst, however, is interested in the effect of one variable, say the first one, then the collection of models \mathcal{M} is $\mathcal{M}_p(k; 1)$ defined in (5.14) and $\widehat{\mathcal{R}}_M$ will be changed to $\{\theta \in \mathbb{R} : |n^{1/2}(\widehat{\beta}_{1 \cdot M} - \theta)| \leq K_M(\alpha)\widehat{\sigma}_{1 \cdot M}\}$ that lead to inference only for the coefficient of X_1 in model M .

We now proceed to construct a few desirable properties for valid simultaneous confidence intervals and propose a particular way of choosing $K_M(\alpha)$ in (5.21).

5.3.2 A New Statistic for PoSI

Suppose $\widehat{\mathcal{C}}_j, 1 \leq j \leq q$ (for some $q \geq 1$) denote a collection of valid simultaneous confidence regions for a collection of functionals $\theta_j, 1 \leq j \leq q$, that is,

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{j=1}^q \{\theta_j \in \widehat{\mathcal{C}}_j\} \right) \geq 1 - \alpha.$$

is satisfied. These confidence regions are said to be *balanced* in the sense of [Beran \(1988\)](#) if

$$\mathbb{P} \left(\theta_j \in \widehat{\mathcal{C}}_j \right) \text{ stays constant across } 1 \leq j \leq q.$$

From (5.16), it is clear that the “max-t” statistic provides balanced confidence intervals for the collection of functionals $\{\beta_{j \cdot M} : M \in \mathcal{M}, j \in M\}$. In particular,

$$\mathbb{P} \left(\max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \leq K_\alpha \right) \text{ does not stay constant across } M \in \mathcal{M},$$

where the quantile K_α is the quantile of the “max-t” statistic as defined in (5.16). Balancedness at the level of models is important for post-selection inference because model selection has a hierarchical structure of first choosing a model or a subset of variables and then think about variables in the model. There is one more level of hierarchy where analysts usually prefer a model of smaller size to a bigger model given that they have similar performance.

Based on this discussion, we desire to construct simultaneous confidence intervals $\hat{\mathcal{R}}_{j \cdot M}, M \in \mathcal{M}_p(k), j \in M$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{\substack{M \in \mathcal{M}_p(k), \\ j \in M}} \{ \beta_{j \cdot M} \in \hat{\mathcal{R}}_{j \cdot M} \} \right) \geq 1 - \alpha,$$

and the confidence intervals $\hat{\mathcal{R}}_{j \cdot M}$ satisfy:

- **Property 1.** For each $M \in \mathcal{M}$, $\mathbb{P}(\beta_{j \cdot M} \in \hat{\mathcal{R}}_{j \cdot M})$ stays constant across $j \in M$.
- **Property 2.** For each model size $1 \leq s \leq k$,

$$\mathbb{P} \left(\beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right) \text{ stays constant across all models } M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1).$$

Note that $\mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)$ is the set of all models of size exactly s .

- **Property 3.** Finally, the simultaneous coverage on all models of size s increases

as s increases, that is,

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} \left\{ \beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right\} \right) \text{ increases as } s \text{ increases from } 1 \text{ to } k.$$

In the properties above $\bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M}$ represents the Cartesian product of the confidence intervals $\hat{\mathcal{R}}_{j \cdot M}, j \in M$.

Now we give an development of confidence intervals $\hat{\mathcal{R}}_{j \cdot M}, M \in \mathcal{M}_p(k), j \in M$ (in installments) satisfying the properties above. We will start with confidence intervals

$$\hat{\mathcal{R}}_{j \cdot M} := \left\{ \theta \in \mathbb{R} : \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \theta)}{\hat{\sigma}_{j \cdot M}} \right| \leq K_M(\alpha) \right\}, \quad (5.22)$$

for some $\alpha \mapsto K_M(\alpha)$ to be determined below. Note that $K_M(\alpha)$ does not depend on j (that is, critical value is the same for all $j \in M$).

Satisfying Property 1. Noting that $n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})/\hat{\sigma}_{j \cdot M}$ is asymptotically close in distribution to a standard normal distribution, confidence intervals $\hat{\mathcal{R}}_{j \cdot M}$ in (5.22) satisfy property 1. Formally, we have

$$\mathbb{P} \left(\beta_{j \cdot M} \in \hat{\mathcal{R}}_{j \cdot M} \right) = \mathbb{P} \left(\left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \leq K_M(\alpha) \right) \approx \mathbb{P} (|Z| \leq K_M(\alpha)),$$

where $Z \sim N(0, 1)$ and hence, the quantity on the right end is constant across $j \in M$.

Satisfying Property 2. For the confidence intervals $\hat{\mathcal{R}}_{j \cdot M}$ in (5.22), we have

$$\begin{aligned} \mathbb{P} \left(\beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right) &= \mathbb{P} \left(\max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \leq K_M(\alpha) \right) \\ &= \mathbb{P} (K_M^{-1}(T_M) \leq \alpha), \end{aligned}$$

where we assume $\alpha \mapsto K_M(\alpha)$ is invertible and write

$$T_M := \max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right|.$$

Property 2 requires $\mathbb{P}(K_M^{-1}(T_M) \leq \alpha)$ to be a constant across $M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)$.

Without loss of generality, we can take

$$K_M^{-1}(t) := b_s \times \text{Cumulative distribution function of } T_M,$$

for some constant b_s depending on the model size s .

Define the cumulative distribution function of T_M as

$$H_M(t) := \mathbb{P}(T_M \leq t) = \mathbb{P} \left(\max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \leq t \right).$$

It is clear that $\mathbb{P}(H_M(T_M) \leq \delta) \leq \delta$ for all $\delta \in [0, 1]$; this becomes an equality if T_M is a continuous random variable. Now define the refined confidence intervals $\hat{\mathcal{R}}_{j \cdot M}$ as

$$\hat{\mathcal{R}}_{j \cdot M} := \left\{ \theta \in \mathbb{R}^M : \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \leq H_M^{-1}(b_{|M|}(\alpha)) \right\}. \quad (5.23)$$

It now follows that for models $M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)$,

$$\begin{aligned} \mathbb{P} \left(\beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right) &= \mathbb{P} \left(H_M \left(\max_{j \in M} \left| \frac{n^{1/2}(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})}{\hat{\sigma}_{j \cdot M}} \right| \right) \leq b_s(\alpha) \right) \\ &= \mathbb{P}(H_M(T_M) \leq b_s(\alpha)) \leq b_s(\alpha), \end{aligned}$$

which is constant across all models $M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)$ because $b_s(\alpha)$ depends only on s . Note that the confidence region (5.23) is not actionable in practice because $H_M(\cdot)$ is unknown but it can be estimated through bootstrap the details of which will

be given later.

Satisfying Property 3. With the refined confidence interval (5.23), we proved that

$$\mathbb{P} \left(\beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right) \leq b_s(\alpha).$$

The simultaneous coverage for all models $M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)$ is given by

$$\mathbb{P} \left(\bigcap_{M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} \left\{ \beta_M \in \bigotimes_{j \in M} \hat{\mathcal{R}}_{j \cdot M} \right\} \right) = \mathbb{P} \left(\max_{M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} H_M(T_M) \leq b_s(\alpha) \right).$$

For property 3, we need to choose $b_s(\alpha)$ so that the right hand side probability increases as s increases from 1 to k . For notational ease, define

$$T_s := \max_{M \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} H_M(T_M).$$

We want to choose $b_s(\alpha)$ so that

$$\mathbb{P}(T_s \leq b_s(\alpha)) \text{ increases with } s.$$

There are many such choices that can be constructed by first transforming T_s to a uniform random variable by using its cumulative distribution function (CDF). We follow another approach that does not use the CDF of individual T_s , $1 \leq s \leq k$.

For this approach, we recall a basic fact about a sequence of (possibly dependent) standard Gaussian random variables Z_1, Z_2, \dots that

$$\max_{j \geq 1} \frac{|Z_j|}{\sqrt{\log j}} = O_p(1). \tag{5.24}$$

See, for example, the discussion following Proposition 4.3.1 of [de la Peña and Giné](#)

(1999). If Z_1, Z_2, \dots are independent, this fact can be written as

$$\max_{j \geq 1} \frac{|Z_j|}{G_j^{-1}(1/2)} = O_p(1),$$

where G_j represents the CDF of $\max_{1 \leq i \leq j} |Z_i|$. We can now formulate an alternative fact that

$$\max_{j \geq 1} G_j(|Z_j|) = O_p(1), \quad \text{where} \quad G_j(t) := \mathbb{P} \left(\max_{1 \leq i \leq j} |Z_i| \leq t \right). \quad (5.25)$$

This fact provide a more accurate information than (5.24). For instance, if Z_1, Z_2, \dots are perfectly dependent in that $Z_1 = Z_2 = \dots$, then $|Z_j| = |Z_1| = O_p(1)$ simultaneously for all $j \geq 1$. In this case

$$G_j(t) = \mathbb{P} \left(\max_{1 \leq i \leq j} |Z_i| \leq t \right) = \mathbb{P}(|Z_1| \leq t)$$

which is independent of $j \geq 1$ and hence (5.25) implies $|Z_j| = O_p(1)$ simultaneously for all $j \geq 1$ which closely describes the truth while (5.24) only implies $|Z_j| = O_p(\sqrt{\log j})$ for all $j \geq 1$ simultaneously which, even though correct, does not describe the true behavior. In general, $G_j^{-1}(t), j \geq 1$ is an increasing sequence and because $|Z_j|$ share a common distribution, (5.25) implies that $|Z_j| = O_p(C_j)$ simultaneously for all $j \geq 1$ for constants C_j increasing with j .

We will use this approach with $|Z_j|, j \geq 1$ replaced by $H_s(T_s), 1 \leq s \leq k$, where $H_s(\cdot)$ is the cumulative distribution of T_s ; the transformation $H_s(T_s)$ is chosen so that they share a common distribution. Define

$$\overline{H}_s(t) := \mathbb{P} \left(\max_{1 \leq q \leq s} H_q(T_q) \leq t \right), \quad (5.26)$$

as the cumulative distribution function of $\max_{1 \leq q \leq s} H_q(T_q)$. Because $\max_{1 \leq q \leq s} H_q(T_q)$ is increasing with s , we get that $\overline{H}_s^{-1}(t) \geq \overline{H}_{s-1}^{-1}(t)$ for all t and $1 \leq s \leq k$. Based on (5.26), we take $b_s(\alpha) = H_s^{-1}(\overline{H}_s^{-1}(C_\alpha))$ for some constant C_α and refine the confidence intervals (5.23) by defining

$$\hat{\mathcal{R}}_{j, \mathbf{M}} := \left\{ \theta \in \mathbb{R} : \left| \frac{n^{1/2}(\hat{\beta}_{j, \mathbf{M}} - \theta)}{\hat{\sigma}_{j, \mathbf{M}}} \right| \leq H_{\mathbf{M}}^{-1}(H_{|\mathbf{M}|}^{-1}(\overline{H}_{|\mathbf{M}|}^{-1}(C_\alpha))) \right\}. \quad (5.27)$$

To verify Property 3 for this refinement, note that

$$\begin{aligned} \mathbb{P} \left(\bigcap_{\mathbf{M} \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} \left\{ \beta_{\mathbf{M}} \in \bigotimes_{j \in \mathbf{M}} \hat{\mathcal{R}}_{j, \mathbf{M}} \right\} \right) &= \mathbb{P} \left(\max_{\mathbf{M} \in \mathcal{M}_p(s) \setminus \mathcal{M}_p(s-1)} H_{\mathbf{M}}(T_{\mathbf{M}}) \leq H_s^{-1} \overline{H}_s^{-1}(C_\alpha) \right) \\ &= \mathbb{P} \left(T_s \leq H_s^{-1}(\overline{H}_s^{-1}(C_\alpha)) \right) \\ &= \mathbb{P}(H_s(T_s) \leq \overline{H}_s^{-1}(C_\alpha)) \leq \overline{H}_s^{-1}(C_\alpha), \end{aligned}$$

with the last term increasing as s increases.

Putting it all together. Combining all the refinements, the final confidence regions are given by (5.27). Now the requirement of simultaneous coverage requires

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\bigcap_{\mathbf{M} \in \mathcal{M}_p(k), j \in \mathbf{M}} \{ \beta_{j, \mathbf{M}} \in \hat{\mathcal{R}}_{j, \mathbf{M}} \} \right) \geq 1 - \alpha.$$

The probability on the left hand side is exactly same as

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(\max_{1 \leq s \leq k} \overline{H}_s \circ H_s \left(\max_{\mathbf{M}: |\mathbf{M}|=s} H_{\mathbf{M}}(T_{\mathbf{M}}) \right) \leq C_\alpha \right) \geq 1 - \alpha.$$

Hence the final confidence region is based on the quantile of the statistic

$$\max_{1 \leq s \leq k} \overline{H}_s \circ H_s \left(\max_{\mathbf{M}: |\mathbf{M}|=s} H_{\mathbf{M}}(T_{\mathbf{M}}) \right). \quad (5.28)$$

Comparing this statistic to the “max-t” statistic in (5.15), we notice that at different levels the maximum statistic is transformed to a uniform random variable by its CDF.

Bootstrap Implementation of the New Statistic

This statistic is not actionable in practice because the CDFs involved: H_M, H_s, \overline{H}_s are not known and needs to be estimated. The statistic used in practice is

$$\max_{1 \leq s \leq k} \widehat{H}_s \circ \widehat{H}_s \left(\max_{M:|M|=s} \widehat{H}_M(T_M) \right),$$

with $\widehat{H}_s, \widehat{H}_s, \widehat{H}_M$ estimated through bootstrap as explained below.

The following represents the main steps in the bootstrap implementation:

1. Generate independent and identically distributed standard normal random variables $e_i^{(b)}, 1 \leq i \leq n, 1 \leq b \leq B$. (B represents the number of bootstrap samples and n represents the sample size).
2. Compute bootstrap version of T_M for all $M \in \mathcal{M}_p(k)$ as $T_M^{(b)}, 1 \leq b \leq B, M \in \mathcal{M}_p(k)$, where

$$T_M^{(b)} := \max_{j \in M} \left| \frac{1}{\sqrt{n} \widehat{\sigma}_{j \cdot M}} \sum_{i=1}^n e_i^{(b)} \widehat{\psi}_{j \cdot M}(X_{i,M}, Y_i) \right|,$$

with $\widehat{\psi}_{j \cdot M}(\cdot, \cdot)$ representing the estimated influence function. In the context of linear regression, $\widehat{\sigma}_{j \cdot M}^2 := n^{-1} \sum_{i=1}^n \widehat{\psi}_{j \cdot M}^2(X_{i,M}, Y_i)$ and

$$\left(\widehat{\psi}_{j \cdot M}(X_{i,M}, Y_i) \right)_{j \in M} := \widehat{\Sigma}_M^{-1} X_{i,M} (Y_i - X_{i,M}^\top \widehat{\beta}_M).$$

Here $\widehat{\Sigma}_M$ is the sample Gram matrix for the model M .

3. Estimate $H_M(\cdot)$ the CDF of T_M by

$$\hat{H}_M(t) := \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{T_M^{(b)} \leq t\}.$$

Now the bootstrapped version of $H_M(T_M)$ can be taken as $\hat{H}_M(T_M^{(b)})$, $1 \leq b \leq B$ and the bootstrapped version of $\max_{|M|=s} H_M(T_M)$ can be taken as

$$T_s^{(b)} := \max_{M:|M|=s} \hat{H}_M(T_M^{(b)}), \quad 1 \leq b \leq B.$$

The estimators of H_s , \bar{H}_s are obtained as

$$\hat{H}_s(t) := \frac{1}{B} \sum_{b=1}^B \mathbb{1}\{T_s^{(b)} \leq t\} \quad \text{and} \quad \hat{\bar{H}}_s(t) := \frac{1}{B} \sum_{b=1}^B \mathbb{1}\left\{\max_{1 \leq q \leq s} \hat{H}_q(T_q^{(b)}) \leq t\right\}.$$

4. Finally, the bootstrapped values of the statistic (5.28) are

$$\max_{1 \leq s \leq k} \hat{\bar{H}}_s \circ \hat{H}_s \left(\max_{M:|M|=s} \hat{H}_M(T_M^{(b)}) \right), \quad 1 \leq b \leq B, \quad (5.29)$$

and required quantile of (5.28) is estimated as the $(1 - \alpha)$ -th quantile of these B numbers.

If $\hat{\pi}_\alpha$ represents the $(1 - \alpha)$ -th quantile of (5.29) then for a data-driven model \hat{M} , the confidence regions are given by

$$\hat{\mathcal{R}}_{j,\hat{M}} := \left\{ \theta \in \mathbb{R} : \left| \frac{n^{1/2}(\hat{\beta}_{j,\hat{M}} - \theta)}{\hat{\sigma}_{j,\hat{M}}} \right| \leq \hat{H}_{\hat{M}}^{-1}(\hat{H}_{|\hat{M}|}^{-1}(\hat{H}_{|\hat{M}|}^{-1}(\hat{\pi}_\alpha))) \right\}. \quad (5.30)$$

Hence the new statistic also leads to intervals centered around the ordinary least squares estimator except with a multiplier that depends on the model \hat{M} in a more intricate manner.

We call the confidence intervals (5.30) as “Hierarchical PoSI Intervals” (HPoSI Intervals). The theoretical validity of HPoSI follows straightforwardly from the results in previous sections. We will see in the next chapter the increased power of HPoSI compared to the $\max\text{-}|t|$ confidence intervals.

5.4 Simulations illustrating the Power of HPoSI

Following the data generating model (4.29) and under setting (5.18), we select one covariate that is most correlated with the response in models of different dimensions (p) and construct PoSI and HPoSI considering different maximal model sizes (k) for the slope. Figure 5.3 shows the ratio of widths of PoSI and the unadjusted interval as well as that of HPoSI and the unadjusted interval for many k 's. It is clear that the width of the interval obtained from HPoSI is much smaller than the width of the corresponding interval obtained from PoSI.

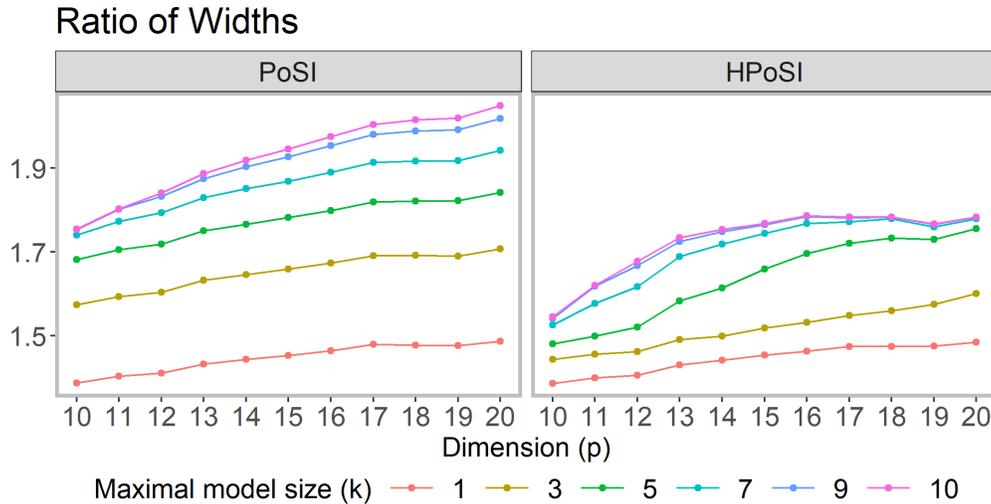


Figure 5.3: Average ratio of widths PoSI vs HPoSI of 1,000 simulations with different maximal model size k , i.e., $|M| \leq k, k = 1, 3, 5, 7, 9, 10$ for $p = 11, 12, \dots, 20$ under Setting (5.18).

For the telomere length data, we have shown the width of PoSI intervals when k increases but the selection only selects a single covariate in Figure 5.1. Figure 5.4

shows the same setup but now includes the HPoSI intervals for each k .

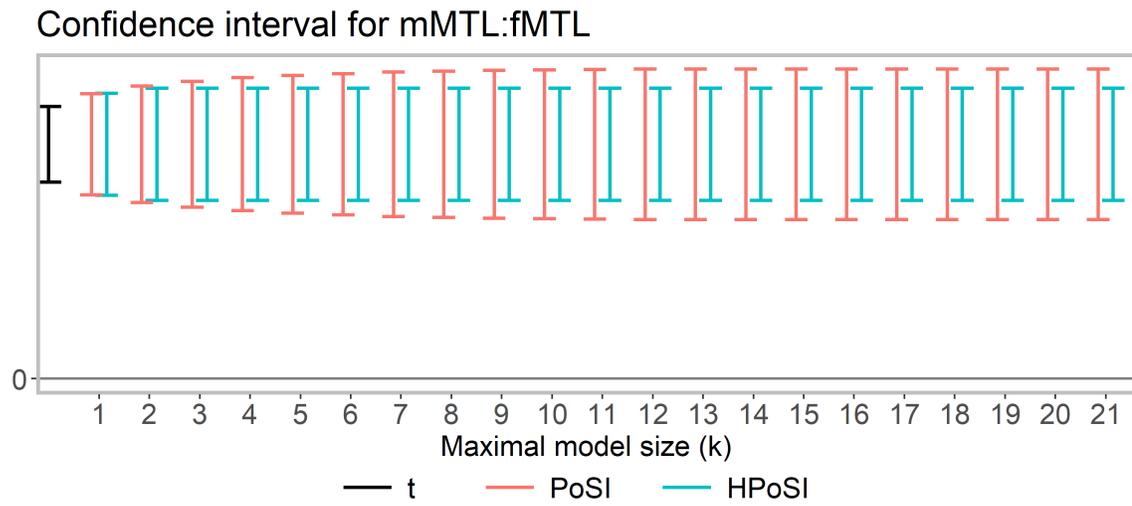


Figure 5.4: Illustrating the dependence on k : Telomere length analysis. Comparison of PoSI and HPoSI.

APPENDIX

5.A Proof of Theorem 13

The definition of $\Delta_{\mathcal{M}}^{\text{CLT}}$ proves

$$\begin{aligned} & \mathbb{P} \left(\max_{(j, \mathcal{M}) \in \mathcal{Q}} \left| \frac{1}{\sqrt{n}\sigma_{j, \mathcal{M}}} \sum_{i=1}^n \psi_{j, \mathcal{M}}(Z_i) \right| \geq 3\sqrt{\log(|\mathcal{Q}|)} + \frac{\pi}{\sqrt{2}} \sqrt{\log(1/\Delta_{\mathcal{M}}^{\text{CLT}})} \right) \\ & \leq \Delta_{\mathcal{M}}^{\text{CLT}} + \mathbb{P} \left(\max_{(j, \mathcal{M}) \in \mathcal{Q}} |G_{j, \mathcal{M}}| \geq 3\sqrt{\log(|\mathcal{Q}|)} + \frac{\pi}{\sqrt{2}} \sqrt{\log(1/\Delta_{\mathcal{M}}^{\text{CLT}})} \right) \\ & \leq \Delta_{\mathcal{M}}^{\text{CLT}} + 2\Delta_{\mathcal{M}}^{\text{CLT}} = 3\Delta_{\mathcal{M}}^{\text{CLT}}. \end{aligned}$$

The last inequality here follows from Equations (3.2) and (3.13) of [Ledoux and Talagrand \(1991\)](#). Consider the event

$$\mathcal{E} := \left\{ \Delta_{\mathcal{M}}^{\text{ULR}} \leq \delta_1, \Delta_{\mathcal{M}}^{\text{Var}} \leq \delta_2 \right\} \cap \left\{ \max_{(j, \mathcal{M}) \in \mathcal{Q}} \left| \sum_{i=1}^n \frac{\psi_{j, \mathcal{M}}(Z_i)}{\sqrt{n}\sigma_{j, \mathcal{M}}} \right| \leq 6\sqrt{\log\left(\frac{|\mathcal{Q}|}{\Delta_{\mathcal{M}}^{\text{CLT}}}\right)} \right\}, \quad (5.31)$$

for some constants $\delta_1, \delta_2 \in (0, 1)$.

Combining assumptions [\(B1\)](#) and [\(B2\)](#) yields

$$\begin{aligned} t_{j, \mathcal{M}} &= \frac{n^{1/2}(\hat{\beta}_{j, \mathcal{M}} - \beta_{j, \mathcal{M}})}{\hat{\sigma}_{j, \mathcal{M}}} = \frac{\sigma_{j, \mathcal{M}}^*}{\hat{\sigma}_{j, \mathcal{M}}} \times \frac{\sigma_{j, \mathcal{M}}}{\sigma_{j, \mathcal{M}}^*} \times \frac{n^{1/2}(\hat{\beta}_{j, \mathcal{M}} - \beta_{j, \mathcal{M}})}{\sigma_{j, \mathcal{M}}} \\ &= \frac{\sigma_{j, \mathcal{M}}^*}{\hat{\sigma}_{j, \mathcal{M}}} \times \frac{\sigma_{j, \mathcal{M}}}{\sigma_{j, \mathcal{M}}^*} \times \left[\frac{1}{\sqrt{n}\sigma_{j, \mathcal{M}}} \sum_{i=1}^n \psi_{j, \mathcal{M}}(Z_i) + \frac{R_{j, \mathcal{M}}}{\sigma_{j, \mathcal{M}}} \right]. \end{aligned}$$

Now using the notation $\Delta_{\mathcal{M}}^{\text{ULR}}$ and $\Delta_{\mathcal{M}}^{\text{Var}}$ implies

$$\begin{aligned}
& \left| t_{j \cdot \mathcal{M}} - \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \right| \\
& \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\hat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| \times \left| \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \right| + \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \left[\left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\hat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| + 1 \right] |R_{j \cdot \mathcal{M}}| \\
& \leq \left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\hat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| \times \left| \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \right| + \left[\left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\hat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| + 1 \right] |R_{j \cdot \mathcal{M}}|.
\end{aligned} \tag{5.32}$$

The last inequality here follows from (5.8). On the event (5.31), we have

$$\left| \frac{\hat{\sigma}_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} - 1 \right| \leq \delta_2 < 1,$$

and hence

$$\left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\hat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| = \frac{|\sigma_{j \cdot \mathcal{M}}^* - \hat{\sigma}_{j \cdot \mathcal{M}}|}{\hat{\sigma}_{j \cdot \mathcal{M}}} \leq \frac{|\sigma_{j \cdot \mathcal{M}}^* - \hat{\sigma}_{j \cdot \mathcal{M}}|}{\sigma_{j \cdot \mathcal{M}}^* - |\hat{\sigma}_{j \cdot \mathcal{M}} - \sigma_{j \cdot \mathcal{M}}^*|} \leq \frac{\delta_2}{1 - \delta_2}.$$

Substituting this inequality in (5.32) concludes, on event \mathcal{E} ,

$$\left| t_{j \cdot \mathcal{M}} - \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \right| \leq \frac{6\delta_2}{1 - \delta_2} \sqrt{\log \left(\frac{|\mathcal{Q}|}{\Delta_{\mathcal{M}}^{\text{CLT}}} \right)} + \frac{\delta_1}{(1 - \delta_2)}.$$

Observe now that for any $(a_{j \cdot \mathcal{M}}), (b_{j \cdot \mathcal{M}})$,

$$\begin{aligned}
\mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \right) & \leq \mathbb{P} \left(\mathcal{E} \cap \bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \right) \\
& + \mathbb{P}(\mathcal{E}^c).
\end{aligned} \tag{5.33}$$

From the definition (5.31) of \mathcal{E} ,

$$\mathcal{E} \cap \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \Rightarrow \left\{ a_{j \cdot \mathcal{M}} - \Delta \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \leq b_{j \cdot \mathcal{M}} + \Delta \right\},$$

where $\Delta := (6\delta_2 \sqrt{\log(|\mathcal{Q}|/\Delta_{\mathcal{M}}^{\text{CLT}})} + \delta_1)/(1 - \delta_2)$. This implies that

$$\begin{aligned} & \mathbb{P} \left(\mathcal{E} \cap \bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \right) \\ & \leq \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} - \Delta \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \leq b_{j \cdot \mathcal{M}} + \Delta \right\} \right) \\ & \leq \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} - \Delta \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} G_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}} + \Delta \right\} \right) + \Delta_{\mathcal{M}}^{\text{CLT}} \\ & \leq \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} G_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}} \right\} \right) + \Delta_{\mathcal{M}}^{\text{CLT}} + \frac{2\Delta}{\underline{\sigma}} \left(\sqrt{2 \log(|\mathcal{Q}|)} + 2 \right). \end{aligned}$$

The last inequality above follows from Theorem 1 of Chernozhukov et al. (2017b) and assumption (5.8). Substituting this in (5.33) concludes

$$\begin{aligned} \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \right) & \leq \mathbb{P} \left(\bigcap_{(j, \mathcal{M}) \in \mathcal{Q}} \left\{ a_{j \cdot \mathcal{M}} \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} G_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}} \right\} \right) \\ & \quad + \Delta_{\mathcal{M}}^{\text{CLT}} + \frac{2\Delta}{\underline{\sigma}} (\sqrt{2 \log(|\mathcal{Q}|)} + 2) + \mathbb{P}(\mathcal{E}^c). \end{aligned}$$

This proves one part of the result. To prove the remaining part, we note that on the event \mathcal{E} ,

$$\mathcal{E} \cap \{a_{j \cdot \mathcal{M}} \leq t_{j \cdot \mathcal{M}} \leq b_{j \cdot \mathcal{M}}\} \Leftarrow \mathcal{E} \cap \left\{ a_{j \cdot \mathcal{M}} + \Delta \leq \frac{\sigma_{j \cdot \mathcal{M}}}{\sigma_{j \cdot \mathcal{M}}^*} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}(Z_i)}{\sqrt{n} \sigma_{j \cdot \mathcal{M}}} \leq b_{j \cdot \mathcal{M}} - \Delta \right\}.$$

Following the same calculations as above yields

$$\begin{aligned} \mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \{a_{j \cdot M} \leq t_{j \cdot M} \leq b_{j \cdot M}\} \right) &\geq \mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \left\{ a_{j \cdot M} \leq \frac{\sigma_{j \cdot M}}{\hat{\sigma}_{j \cdot M}^*} G_{j \cdot M} \leq b_{j \cdot M} \right\} \right) \\ &\quad - \Delta_{\mathcal{M}}^{\text{CLT}} - \frac{2\Delta}{\underline{\sigma}} (\sqrt{2 \log(|\mathcal{Q}|)} + 2) - \mathbb{P}(\mathcal{E}^c). \end{aligned}$$

The result now follows from the fact that

$$\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}(\Delta_{\mathcal{M}}^{\text{ULR}} > \delta_1) + \mathbb{P}(\Delta_{\mathcal{M}}^{\text{Var}} > \delta_2) + 3\Delta_{\mathcal{M}}^{\text{CLT}}.$$

5.B Proof of Theorem 14

Let the left hand side quantity be denoted by ρ_n^{MB} . Because e_1^1, \dots, e_n^1 (conditional on \mathcal{D}_n) are independent standard Gaussian random variables,

$$\left(t_{j \cdot M}^{(1)} \right)_{(j,M) \in \mathcal{Q}} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n e_i^1 \times \frac{\hat{\psi}_{j \cdot M}(Z_i)}{\hat{\sigma}_{j \cdot M}} \right)_{(j,M) \in \mathcal{Q}},$$

is a Gaussian random vector with mean zero. Remark 4.1 of [Chernozhukov et al. \(2017a\)](#) yields

$$\rho_n^{\text{MB}} \leq C(\Delta_n^{\text{MB}})^{1/3} (\log(|\mathcal{Q}|))^{2/3},$$

where

$$\Delta_n^{\text{MB}} := \max_{\substack{(j,M), \\ (j',M') \in \mathcal{Q}}} \left| \text{Cov}(t_{j \cdot M}^{(1)}, t_{j' \cdot M'}^{(1)} | \mathcal{D}_n) - \text{Cov}(G_{j \cdot M}, G_{j' \cdot M'}) \right|.$$

Conditional on \mathcal{D}_n , the covariance between $t_{j \cdot M}^{(1)}$ and $t_{j' \cdot M'}^{(1)}$ is given by

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{\psi}_{j \cdot M}(Z_i) \hat{\psi}_{j' \cdot M'}(Z_i)}{\hat{\sigma}_{j \cdot M} \hat{\sigma}_{j' \cdot M'}}.$$

The definition of $\Delta_{\mathcal{M}}^{\text{Var}}$ implies

$$\left| \frac{\sigma_{j \cdot \mathcal{M}}^*}{\widehat{\sigma}_{j \cdot \mathcal{M}}} - 1 \right| \leq \frac{\Delta_{\mathcal{M}}^{\text{Var}}}{(1 - \Delta_{\mathcal{M}}^{\text{Var}})_+} \quad \text{for all } (j, \mathcal{M}) \in \mathcal{Q}.$$

This implies, using $|ab - 1| \leq |a - 1||b - 1| + |a - 1| + |b - 1|$, that

$$\left| \frac{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*}{\widehat{\sigma}_{j \cdot \mathcal{M}} \widehat{\sigma}_{j' \cdot \mathcal{M}'}} - 1 \right| \leq \frac{2\Delta_{\mathcal{M}}^{\text{Var}}}{(1 - \Delta_{\mathcal{M}}^{\text{Var}})_+} + \left(\frac{\Delta_{\mathcal{M}}^{\text{Var}}}{(1 - \Delta_{\mathcal{M}}^{\text{Var}})_+} \right)^2.$$

Hence, on the event $\Delta_{\mathcal{M}}^{\text{Var}} \leq 1/2$, we get

$$\left| \frac{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*}{\widehat{\sigma}_{j \cdot \mathcal{M}} \widehat{\sigma}_{j' \cdot \mathcal{M}'}} - 1 \right| \leq 6\Delta_{\mathcal{M}}^{\text{Var}}.$$

Therefore, on the event $\Delta_{\mathcal{M}}^{\text{Var}} \leq 1/2$,

$$\left| \text{Cov}(t_{j \cdot \mathcal{M}}^{(1)}, t_{j' \cdot \mathcal{M}'}^{(1)} | \mathcal{D}_n) - \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\psi}_{j \cdot \mathcal{M}}(Z_i) \widehat{\psi}_{j' \cdot \mathcal{M}'}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*} \right| \leq 6\Delta_{\mathcal{M}}^{\text{Var}} \left| \frac{1}{n} \sum_{i=1}^n \frac{\widehat{\psi}_{j \cdot \mathcal{M}}(Z_i) \widehat{\psi}_{j' \cdot \mathcal{M}'}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*} \right|.$$

Cauchy-Schwarz inequality concludes

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n \left\{ \frac{\widehat{\psi}_{j \cdot \mathcal{M}}(Z_i) \widehat{\psi}_{j' \cdot \mathcal{M}'}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*} - \frac{\psi_{j \cdot \mathcal{M}}(Z_i) \psi_{j' \cdot \mathcal{M}'}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^* \sigma_{j' \cdot \mathcal{M}'}^*} \right\} \right| \\ & \leq 2 \max_{(j, \mathcal{M}) \in \mathcal{Q}} \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{\widehat{\psi}_{j \cdot \mathcal{M}}(Z_i) - \psi_{j \cdot \mathcal{M}}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^*} \right)^2} \max_{(j, \mathcal{M}) \in \mathcal{Q}} \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}^2(Z_i)}{(\sigma_{j \cdot \mathcal{M}}^*)^2}} \\ & \quad + \max_{(j, \mathcal{M}) \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^n \left(\frac{\widehat{\psi}_{j \cdot \mathcal{M}}(Z_i) - \psi_{j \cdot \mathcal{M}}(Z_i)}{\sigma_{j \cdot \mathcal{M}}^*} \right)^2 \\ & \leq 2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} \max_{(j, \mathcal{M}) \in \mathcal{Q}} \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot \mathcal{M}}^2(Z_i)}{\sigma_{j \cdot \mathcal{M}}^2}} + \Delta_{\mathcal{M}}^{\text{Inf}} \\ & \leq 2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} (1 + \Delta_{\mathcal{M}}^{\text{Boot}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Inf}}. \end{aligned}$$

The last inequality above follows because under independence $\sigma_{j \cdot M}^2 := \sum_{i=1}^n \mathbb{E}[\psi_{j \cdot M}^2(X_i)]/n$.

This concludes that

$$\begin{aligned}
& \left| \text{Cov}(t_{j \cdot M}^{(1)}, t_{j' \cdot M'}^{(1)} | \mathcal{D}_n) - \frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right| \\
& \leq 6\Delta_{\mathcal{M}}^{\text{Var}} \left| \frac{1}{n} \sum_{i=1}^n \frac{\hat{\psi}_{j \cdot M}(Z_i) \hat{\psi}_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right| \\
& \quad + 2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} (1 + \Delta_{\mathcal{M}}^{\text{Boot}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Inf}} \\
& \leq 6\Delta_{\mathcal{M}}^{\text{Var}} \left| \frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right| \\
& \quad + (6\Delta_{\mathcal{M}}^{\text{Var}} + 1) \left[2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} (1 + \Delta_{\mathcal{M}}^{\text{Boot}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Inf}} \right] \\
& \leq 6\Delta_{\mathcal{M}}^{\text{Var}} [1 + \Delta_{\mathcal{M}}^{\text{Boot}}] + (6\Delta_{\mathcal{M}}^{\text{Var}} + 1) \left[2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} (1 + \Delta_{\mathcal{M}}^{\text{Boot}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Inf}} \right].
\end{aligned}$$

Therefore, on the event $\mathcal{E}^{\text{Boot}} := \{\Delta_{\mathcal{M}}^{\text{Var}} \leq 1/2\} \cap \{\Delta_{\mathcal{M}}^{\text{Inf}} \leq 1\} \cap \{\Delta_{\mathcal{M}}^{\text{Boot}} \leq 1\}$, we get for all $(j, M) \in \mathcal{Q}$

$$\begin{aligned}
& \left| \text{Cov}(t_{j \cdot M}^{(1)}, t_{j' \cdot M'}^{(1)} | \mathcal{D}_n) - \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right] \right| \\
& \leq \Delta_{\mathcal{M}}^{\text{Boot}} + 6\Delta_{\mathcal{M}}^{\text{Var}} [1 + \Delta_{\mathcal{M}}^{\text{Boot}}] + (6\Delta_{\mathcal{M}}^{\text{Var}} + 1) \left[2 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} (1 + \Delta_{\mathcal{M}}^{\text{Boot}})^{1/2} + \Delta_{\mathcal{M}}^{\text{Inf}} \right] \\
& \leq \Delta_{\mathcal{M}}^{\text{Boot}} + 12\Delta_{\mathcal{M}}^{\text{Var}} + 4 \left[2\sqrt{2} (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} + (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2} \right] \\
& \leq \Delta_{\mathcal{M}}^{\text{Boot}} + 12\Delta_{\mathcal{M}}^{\text{Var}} + 16 (\Delta_{\mathcal{M}}^{\text{Inf}})^{1/2}.
\end{aligned}$$

This proves the first part of the result because under the independence assumption

$$\text{Cov}(G_{j \cdot M}^{\text{Boot}}, G_{j' \cdot M'}^{\text{Boot}}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M}^* \sigma_{j' \cdot M'}^*} \right].$$

To prove the second part, note that the covariance matrix of $(G_{j \cdot M}^{\text{Boot}})$ is lower bounded (in the matrix sense) by the covariance matrix of $(\sigma_{j \cdot M} G_{j \cdot M} / \sigma_{j \cdot M}^*)$ and hence by An-

derson's lemma,

$$\mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \{-b_{j \cdot M} \leq G_{j \cdot M}^{\text{Boot}} \leq b_{j \cdot M}\} \right) \leq \mathbb{P} \left(\bigcap_{(j,M) \in \mathcal{Q}} \left\{ -b_{j \cdot M} \leq \frac{\sigma_{j \cdot M} G_{j \cdot M}}{\sigma_{j \cdot M}^*} \leq b_{j \cdot M} \right\} \right).$$

This implies the second part of the result.

5.C Proof of Theorem 15

Because the result is deterministic, it is enough to prove the result for any fixed $M \subseteq \{1, 2, \dots, p\}$. Fix any $M \subseteq \{1, 2, \dots, p\}$. The result is trivially true if $\mathcal{D}_M^\Sigma \geq 1$. Hence assume that $\mathcal{D}_M^\Sigma < 1$. For notational convenience, let

$$\hat{\Sigma}_M := \frac{1}{n} \sum_{i=1}^n X_{i,M} X_{i,M}^\top \quad \text{and} \quad \hat{\Gamma}_M := \frac{1}{n} \sum_{i=1}^n X_{i,M} Y_i.$$

From the definition of $\hat{\beta}_M$, we have the normal equations $\hat{\Sigma}_M \hat{\beta} = \hat{\Gamma}_M$. Subtracting $\hat{\Sigma}_M \beta_M \in \mathbb{R}^d$ from both sides, we get $\hat{\Sigma}_M (\hat{\beta}_M - \beta_M) = \hat{\Gamma}_M - \hat{\Sigma}_M \beta_M$, which is equivalent to

$$(\Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2}) \Sigma_M^{1/2} (\hat{\beta}_M - \beta_M) = \Sigma_M^{-1/2} (\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M),$$

since Σ_M is invertible. Adding and subtracting $\Sigma_M^{-1/2} (\hat{\beta}_M - \beta_M)$ on both sides further yields the identity

$$\Sigma_M^{1/2} \left[\hat{\beta}_M - \beta_M - \Sigma_n^{-1} (\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M) \right] = (I_d - \Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2}) \Sigma_M^{1/2} (\hat{\beta}_M - \beta_M).$$

Applying the Euclidean norm, we get that

$$\begin{aligned}
\|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M} &\leq \|(I_d - \Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2}) \Sigma_M^{1/2} (\hat{\beta}_M - \beta_M)\| \\
&\leq \|I_d - \Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2}\|_{op} \|\hat{\beta}_M - \beta_M\|_{\Sigma_M} \quad (5.34) \\
&= \mathcal{D}_M^\Sigma \|\hat{\beta}_M - \beta_M\|_{\Sigma_M}.
\end{aligned}$$

This inequality along with the triangle inequality implies that

$$\begin{aligned}
\|\hat{\beta}_M - \beta_M\|_{\Sigma_M} &\leq \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M} + \|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M} \\
&\leq \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M} + \mathcal{D}_M^\Sigma \|\hat{\beta}_M - \beta_M\|_{\Sigma_M},
\end{aligned}$$

and hence (using $\mathcal{D}_M^\Sigma < 1$) yields

$$\|\hat{\beta} - \beta\|_{\Sigma_M} \leq \frac{1}{(1 - \mathcal{D}_n^\Sigma)} \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M}. \quad (5.35)$$

Combining (5.35) and (5.34) concludes

$$\|\hat{\beta} - \beta - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta)\|_{\Sigma_M} \leq \frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)} \|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|_{\Sigma_M}. \quad (5.36)$$

To replace the norms with respect to Σ_M to those with $\Sigma_M V_n^{-1} \Sigma_M$, we note that for any $\theta \in \mathbb{R}^d$,

$$\|\theta\|_{\Sigma_M V_n^{-1} \Sigma_M} = \|V_n^{-1/2} \Sigma_M \theta\| = \sqrt{\theta^\top \Sigma_M V_n^{-1} \Sigma_M \theta} \leq \|\theta\|_{\Sigma_M} \sqrt{\lambda},$$

and

$$\begin{aligned}\|\theta\|_{\Sigma_M} &= \|\Sigma_M^{1/2}\theta\| = \|\Sigma_M^{-1/2}V_M^{1/2}V_M^{-1/2}\Sigma_M\theta\| \\ &\leq \sqrt{\lambda_{\max}(\Sigma_M^{-1/2}V_M\Sigma_M^{-1/2})}\|\theta\|_{\Sigma_M V_M^{-1}\Sigma_M} \leq \frac{1}{\sqrt{\underline{\lambda}}}\|\theta\|_{\Sigma_M V_M^{-1}\Sigma_M}.\end{aligned}$$

Substituting these inequalities in (5.36) yields

$$\|\hat{\beta}_M - \beta_M - \Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M} \leq \sqrt{\frac{\bar{\lambda}}{\underline{\lambda}}}\frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)}\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}, \quad (5.37)$$

Observe now that for any $x \in \mathbb{R}^d$ and any invertible matrix A ,

$$\|x\|_A = \|A^{1/2}x\| = \max_{\theta \in \mathbb{R}^d} \frac{\theta^\top x}{\sqrt{\theta^\top A^{-1}\theta}} \geq \max_{\substack{\theta = \pm e_j \\ 1 \leq j \leq d}} \frac{|\theta^\top x|}{\sqrt{\theta^\top A^{-1}\theta}} = \max_{1 \leq j \leq d} \frac{|x_j|}{\sqrt{(A^{-1})_{jj}}}. \quad (5.38)$$

Applying this inequality for the left hand side of (5.37) concludes

$$\max_{j \in M} \left| \frac{\hat{\beta}_{j \cdot M} - \beta_{j \cdot M}}{\sigma_{j \cdot M}} - \frac{1}{n} \sum_{i=1}^n \frac{\psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}} \right| \leq \sqrt{\frac{\bar{\lambda}}{\underline{\lambda}}}\frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)}\|\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M)\|_{\Sigma_M V_M^{-1}\Sigma_M}.$$

Because

$$\left(V_M^{-1/2}\Sigma_M^{-1}\right)\Sigma_M^{-1}(\hat{\Gamma}_M - \hat{\Sigma}_M\beta_M) = \frac{1}{n} \sum_{i=1}^n V_M^{-1/2}X_{i,M}(Y_i - X_{i,M}^\top\beta_M),$$

the result follows.

5.D Proof of Theorem 16

Observe that

$$\mathcal{D}_M^\Sigma = \|\Sigma_M^{-1/2}\hat{\Sigma}_M\Sigma_M^{-1/2} - I_{|M|}\|_{op} \leq 2 \sup_{\nu \in \mathcal{N}_{|M|}^{1/4}} \left| \frac{1}{n} \sum_{i=1}^n (\nu^\top \Sigma_M^{-1/2}X_{i,M})^2 - 1 \right|, \quad (5.39)$$

where $\mathcal{N}_{|M|}^{1/4}$ represents the 1/4-net of $\{\theta \in \mathbb{R}^{|M|} : \|\theta\| = 1\}$; see Lemma 2.2 of [Vershynin \(2012\)](#). Note that $|\mathcal{N}_{|M|}^{1/4}| \leq 9^{|M|}$. Therefore the right hand side of (5.39) is a maximum over a finite number of mean zero averages with summands satisfying

$$\mathbb{E} \left[\exp \left(\mathfrak{K}_\beta^{-\beta} |\nu^\top \Sigma_M^{-1/2} X_{i,M}|^\beta \right) \right] \leq 2, \text{ for all } \nu \in \mathcal{N}_{|M|}^{1/4} \text{ and } M \subseteq \{1, 2, \dots, d\}.$$

Applying Theorem 3.4 of [Kuchibhotla and Chakraborty \(2018\)](#), we get for any $t \geq 0$ that with probability $1 - 3e^{-t}$,

$$\mathcal{D}_M^\Sigma \leq 14 \sqrt{\frac{\kappa_M^\Sigma (t + |M| \log(9))}{n}} + \frac{C_\beta \mathfrak{K}_\beta^2 (\log(2n))^{2/\beta} (t + |M| \log(9))^{\max\{1, 2/\beta\}}}{n},$$

for some constant $C_\beta > 0$ depending only β . Since there are $\binom{d}{s} \leq (ed/s)^s$ models of size s , taking $t = s \log(ed/s) + u$ (for any $u \geq 0$) and applying union bound over all models of size s , we get that with probability $1 - 3e^{-u}$, simultaneously for all $M \subseteq \{1, 2, \dots, d\}$ with $|M| = s$,

$$\mathcal{D}_M^\Sigma \leq 14 \sqrt{\frac{\kappa_M^\Sigma (u + s \log(9ed/s))}{n}} + \frac{C_\beta \mathfrak{K}_\beta^2 (\log(2n))^{2/\beta} (u + s \log(9ed/s))^{\max\{1, 2/\beta\}}}{n}.$$

To prove the result simultaneously over all $1 \leq s \leq d$, take $u = v + \log(\pi^2 s^2/6)$ and apply union bound over $1 \leq s \leq d$ to get with probability $1 - 3e^{-v}$ simultaneously over all $M \subseteq \{1, 2, \dots, d\}$ with $|M| = s$ for some $1 \leq s \leq d$,

$$\begin{aligned} \mathcal{D}_M^\Sigma &\leq 14 \sqrt{\frac{\kappa_M^\Sigma (v + \log(\pi^2 s^2/6) + s \log(9ed/s))}{n}} \\ &\quad + \frac{C_\beta \mathfrak{K}_\beta^2 (\log(2n))^{2/\beta} (v + \log(\pi^2 s^2/6) + s \log(9ed/s))^{\max\{1, 2/\beta\}}}{n}. \end{aligned}$$

Since $s^{-1} \log(\pi^2 s^2/6) \leq (2\pi/\sqrt{6}) \sup_{x \geq \pi/\sqrt{6}} \exp(-x)x \leq 1$, we get with probability $1 - 3e^{-v}$ simultaneously for any $1 \leq s \leq d$ and for any model $M \subseteq \{1, 2, \dots, d\}$ with

$|M| = s,$

$$\mathcal{D}_M^\Sigma \leq 14 \sqrt{\frac{\kappa_M^\Sigma (v + s \log(9e^2 d/s))}{n}} + \frac{C_\beta \mathfrak{K}_\beta^2 (\log(2n))^{2/\beta} (v + s \log(9e^2 d/s))^{\max\{1, 2/\beta\}}}{n}.$$

This proves (5.11).

We now bound $\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\|$ simultaneously over all M . Observe from the definition of β_M that

$$0 \leq \sum_{i=1}^n \mathbb{E}[(Y_i - X_{i,M}^\top \beta_M)^2] = \sum_{i=1}^n \mathbb{E}[Y_i^2] - \sum_{i=1}^n \mathbb{E}[(X_{i,M}^\top \beta_M)^2],$$

and hence $\|\Sigma_M^{1/2} \beta_M\| \leq (\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2}$. Now note that since $\mathbb{E}[\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M] = 0$ (from the definition of β_M), we have

$$\begin{aligned} \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\| &= \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M) - \Sigma_M^{-1/2}(\hat{\Sigma}_M - \Sigma_M)\beta_M\| \\ &\leq \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| + \|\Sigma_M^{-1/2}(\hat{\Sigma}_M - \Sigma_M)\Sigma_M^{-1/2}\|_{op} \|\Sigma_M^{1/2} \beta_M\| \\ &\leq \|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| + \mathcal{D}_M^\Sigma (\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2}. \end{aligned} \quad (5.40)$$

We have already controlled \mathcal{D}_M^Σ uniformly over all models $M \subseteq \{1, 2, \dots, d\}$ and hence it is enough to control $\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\|$. As before, observe that

$$\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \mathbb{E}\hat{\Gamma}_M)\| \leq 2 \max_{\nu \in \mathcal{N}_{|M|}^{1/2}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \nu^\top \tilde{X}_{i,M} Y_i - \mathbb{E}[\nu^\top \tilde{X}_{i,M} Y_i] \right\} \right| =: 2\mathcal{E}_M,$$

where $\tilde{X}_{i,M} := \Sigma_M^{-1/2} X_{i,M}$. To control \mathcal{E}_M we split Y_i in to two parts depending on whether $\{|Y_i| \leq B\}$ or $\{|Y_i| > B\}$ (for a B to be chosen later). Define $Y_{i,1} = Y_i \mathbb{1}_{\{|Y_i| \leq B\}}$

$B\}$, $Y_{i,2} = Y_i - Y_{i,1}$ and for $\ell = 1, 2$,

$$\mathcal{E}_{M,\ell} := \max_{\nu \in \mathcal{N}_{|M|}^{1/2}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \nu^\top \tilde{X}_{i,M} Y_{i,\ell} - \mathbb{E}[\nu^\top \tilde{X}_{i,M} Y_{i,\ell}] \right\} \right|.$$

Using this notation, from (5.40), we get

$$\|\Sigma_M^{-1/2}(\hat{\Gamma}_M - \hat{\Sigma}_M \beta_M)\| \leq 2\mathcal{E}_{M,1} + 2\mathcal{E}_{M,2} + \mathcal{D}_M^\Sigma (\sum_{i=1}^n \mathbb{E}[Y_i^2]/n)^{1/2}. \quad (5.41)$$

Since $|Y_{i,1}| \leq B$, we have for any $\nu \in \mathcal{N}_{|M|}^{1/2}$ and $M \subseteq \{1, 2, \dots, d\}$ that

$$\mathbb{E} \left[\exp \left(\frac{|\nu^\top \tilde{X}_{i,M} Y_{i,1}|^\beta}{(B\mathfrak{K}_\beta)^\beta} \right) \right] \leq 2.$$

Hence we get by Theorem 3.4 of [Kuchibhotla and Chakraborty \(2018\)](#) that for any $t \geq 0$, with probability $1 - 3e^{-t}$

$$\mathcal{E}_{M,1} \leq 7\sqrt{\frac{\mathfrak{V}_M(t + |M| \log(5))}{n}} + \frac{C_\beta B \mathfrak{K}_\beta (\log(2n))^{1/\beta} (t + |M| \log(5))^{\max\{1, 1/\beta\}}}{n}.$$

Now following same approach as used for \mathcal{D}_M^Σ , we get with probability $1 - 3e^{-u}$, for any $1 \leq s \leq d$, for any model $M \subseteq \{1, 2, \dots, d\}$ such that $|M| = s$,

$$\mathcal{E}_{M,1} \leq 7\sqrt{\frac{\mathfrak{V}_M(v + s \log(5e^2 d/s))}{n}} + \frac{C_\beta B \mathfrak{K}_\beta (\log(2n))^{1/\beta} (v + s \log(5e^2 d/s))^{\max\{1, 1/\beta\}}}{n}. \quad (5.42)$$

To bound $\mathcal{E}_{M,2}$ simultaneously over all M , we take

$$B := 8\mathbb{E} \left[\max_{1 \leq i \leq n} |Y_i| \right] \leq 8n^{1/r} \max_{1 \leq i \leq n} (\mathbb{E}[|Y_i|^r])^{1/r} = 8n^{1/r} K_{n,r},$$

which is motivated by Proposition 6.8 of [Ledoux and Talagrand \(1991\)](#). Now consider

the normalized process

$$\mathcal{E}_{2,\text{Norm}} := \max_{1 \leq s \leq d} \max_{|M|=s} \frac{n^{1/2} \mathcal{E}_{M,2}}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}}.$$

Observe first that $\mathcal{E}_{2,\text{Norm}} \leq \mathcal{E}^{(1)} + \mathbb{E}[\mathcal{E}^{(1)}]$, where

$$\mathcal{E}^{(1)} = \frac{1}{n} \sum_{i=1}^n \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{1/2} |\nu^\top \tilde{X}_{i,M} Y_{i,2}|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}}.$$

Note that $\mathcal{E}^{(1)}$ is an average of non-negative random variables and hence by the choice of B above and Proposition 6.8 of [Ledoux and Talagrand \(1991\)](#), we get

$$\begin{aligned} \mathbb{E}[\mathcal{E}^{(1)}] &\leq 8\mathbb{E} \left[\frac{1}{n} \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{1/2} |\nu^\top \tilde{X}_{i,M} Y_{i,2}|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right] \\ &\leq 8\mathbb{E} \left[\max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{-1/2} |\nu^\top \tilde{X}_{i,M} Y_i|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right] \quad (5.43) \\ &\leq 8 \left\| \max_{1 \leq i \leq n} \frac{|Y_i|}{K_{n,r} n^{1/r}} \right\|_2 \left\| \max_{1 \leq s \leq d} \max_{\substack{1 \leq i \leq n, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_2. \end{aligned}$$

Here we use $\|W\|_2$ for a random variable W to denote $(\mathbb{E}[W^2])^{1/2}$. In the second factor, the number of items in the maximum for any fixed s is given by $n \binom{d}{s} 5^s \leq n(5ed/s)^s$ and hence from [\(X-SW\)](#), we get

$$\mathbb{P} \left(\max_{\substack{1 \leq i \leq n, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} |\nu^\top \tilde{X}_{i,M}| \geq \mathfrak{K}_\beta(t + s \log(5ed/s) + \log(n))^{1/\beta} \right) \leq 2e^{-t},$$

and an application of union bound over $1 \leq s \leq d$ yields

$$\mathbb{P} \left(\bigcup_{1 \leq s \leq d} \left\{ \max_{\substack{1 \leq i \leq n, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} |\nu^\top \tilde{X}_{i,M}| \geq \mathfrak{K}_\beta(t + \log(\pi^2 s^2/6) + s \log(5ed/s) + \log(n))^{1/\beta} \right\} \right) \leq 2e^{-t},$$

which implies

$$\mathbb{P} \left(\bigcup_{1 \leq s \leq d} \left\{ \max_{\substack{1 \leq i \leq n, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} |\nu^\top \tilde{X}_{i,M}| \geq \mathfrak{K}_\beta (t + s \log(5e^2 d/s) + \log(n))^{1/\beta} \right\} \right) \leq 2e^{-t}. \quad (5.44)$$

Hence for a constant $C_\beta > 0$ (depending only on β),

$$\left\| \max_{1 \leq s \leq d} \max_{\substack{1 \leq i \leq n, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_2 \leq C_\beta. \quad (5.45)$$

For the first factor in (5.43), note that (since $r \geq 2$)

$$\left\| \max_{1 \leq i \leq n} \frac{|Y_i|}{K_{n,r} n^{1/r}} \right\|_2 \leq \left\| \max_{1 \leq i \leq n} \frac{|Y_i|}{K_{n,r} n^{1/r}} \right\|_r \leq \left(\sum_{i=1}^n \mathbb{E} \left[\frac{|Y_i|^r}{K_{n,r}^r n} \right] \right)^{1/r} \leq 1. \quad (5.46)$$

Substituting the bounds (5.46) and (5.45) in (5.43) yields

$$\mathbb{E}[\mathcal{E}_{2,\text{Norm}}] \leq 2\mathbb{E}[\mathcal{E}^{(1)}] \leq C_\beta, \quad (5.47)$$

for a constant $C_\beta > 0$ (which is different from the one in (5.45)). Applying Theorem 8 of [Boucheron et al. \(2005\)](#) now yields for every $q \geq 1$

$$\|\mathcal{E}^{(1)}\|_q \leq 2\mathbb{E}[\mathcal{E}^{(1)}] + Cq \left\| \frac{1}{n} \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{1/2} |\nu^\top \tilde{X}_{i,M} Y_{i,2}|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_q,$$

for some (other) absolute constant $C > 0$. This implies (using (5.47)) that

$$\|\mathcal{E}_{2,\text{Norm}}\|_q \leq 3C_\beta + Cq \left\| \frac{1}{n} \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{1/2} |\nu^\top \tilde{X}_{i,M} Y_{i,2}|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_q.$$

As before, we have

$$\begin{aligned}
& \left\| \frac{1}{n} \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{n^{1/2} |\nu^\top \tilde{X}_{i,M} Y_{i,2}|}{n^{-1/2+1/r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_q \\
& \leq \left\| \max_{1 \leq i \leq n} \frac{|Y_i|}{K_{n,r} n^{1/r}} \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_q \\
& \leq \left\| \max_{1 \leq i \leq n} \frac{|Y_i|}{K_{n,r} n^{1/r}} \right\|_r \left\| \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_{rq/(r-q)}.
\end{aligned}$$

where the last inequality holds for any $q < r$ by Hölder's inequality. We already have that the first factor is bounded by 1. From (5.44), we have

$$\left\| \max_{1 \leq i \leq n} \max_{\substack{1 \leq s \leq d, \\ |M|=s}} \max_{\nu \in \mathcal{N}_s^{1/2}} \frac{|\nu^\top \tilde{X}_{i,M}|}{\mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}} \right\|_{rq/(r-q)} \leq C_\beta \left(\frac{rq}{r-q} \right)^{1/\beta}.$$

Therefore taking $q = r - 1$, we get

$$\|\mathcal{E}_{2,\text{Norm}}\|_{r-1} \leq 3C_\beta + CC_\beta(r-1)(r(r-1))^{1/\beta} =: C_{\beta,r}.$$

Hence by Markov's inequality, we get with probability at least $1 - 1/t^{r-1}$, for any $1 \leq s \leq d$, for any model $M \subseteq \{1, 2, \dots, d\}$ such that $|M| = s$,

$$\mathcal{E}_{M,2} \leq \frac{tC_{\beta,r} K_{n,r} \mathfrak{K}_\beta(s \log(5e^2 d/s) + \log n)^{1/\beta}}{n^{1-1/r}}. \tag{5.48}$$

Combining the bounds (5.42) and (5.48) yields: with probability at least $1 - 3e^{-t_1} -$

t_2^{-r+1} , for any $1 \leq s \leq d$, for any model $M \subseteq \{1, 2, \dots, d\}$ such that $|M| = s$,

$$\begin{aligned} \mathcal{E}_M &\leq 7\sqrt{\frac{\mathfrak{V}_M(t_1 + s \log(5e^2 d/s))}{n}} + \frac{C_\beta K_{n,r} \mathfrak{K}_\beta (\log(2n))^{1/\beta} (t_1 + s \log(5e^2 d/s))^{\max\{1, 1/\beta\}}}{n^{1-1/r}} \\ &\quad + \frac{t_2 C_{\beta,r} K_{n,r} \mathfrak{K}_\beta (s \log(5e^2 d/s) + \log n)^{1/\beta}}{n^{1-1/r}}. \end{aligned}$$

Combining this inequality with (5.41) completes the proof of (5.12).

5.E Proof of Lemma 10

Recall that

$$\begin{aligned} \hat{\psi}_M(Z_i) &= \hat{\Sigma}_M^{-1} X_{i,M} (Y_i - X_{i,M}^\top \hat{\beta}_M) \\ &= \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \hat{\beta}_M) \\ &= \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) + \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} \Sigma_M^{-1/2} X_{i,M} X_{i,M}^\top (\hat{\beta}_M - \beta_M). \end{aligned}$$

This implies that

$$\begin{aligned} \Sigma_M^{1/2} \left(\hat{\psi}_M(Z_i) - \psi_M(Z_i) \right) &= \left[\Sigma_M^{1/2} \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} - I_{|M|} \right] \Sigma_M^{-1/2} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) \\ &\quad + \Sigma_M^{1/2} \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} \Sigma_M^{-1/2} X_{i,M} X_{i,M}^\top (\hat{\beta}_M - \beta_M) \\ &= \left[\Sigma_M^{1/2} \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} - I_{|M|} \right] \Sigma_M^{1/2} \psi_M(Z_i) \\ &\quad + \left[\Sigma_M^{1/2} \hat{\Sigma}_M^{-1} \Sigma_M^{1/2} - I_{|M|} \right] \Sigma_M^{-1/2} X_{i,M} X_{i,M}^\top (\hat{\beta}_M - \beta_M) \\ &\quad + \Sigma_M^{-1/2} X_{i,M} X_{i,M}^\top (\hat{\beta}_M - \beta_M). \end{aligned}$$

Applying the Euclidean norm yields

$$\begin{aligned} \|\Sigma_M^{1/2}\{\hat{\psi}_M(Z_i) - \psi_M(Z_i)\}\| &\leq \mathcal{D}_M^\Sigma \|\Sigma_M^{1/2}\psi_M(Z_i)\| + \mathcal{D}_M^\Sigma \|\Sigma_M^{-1/2}X_{i,M}\|^2 \|\hat{\beta}_M - \beta_M\|_{\Sigma_M} \\ &\quad + \|\Sigma_M^{-1/2}X_{i,M}\|^2 \|\hat{\beta}_M - \beta_M\|_{\Sigma_M}. \end{aligned}$$

Summing over $1 \leq i \leq n$ concludes

$$\begin{aligned} \sqrt{\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2}\{\hat{\psi}_M(Z_i) - \psi_M(Z_i)\}\|^2} &\leq \mathcal{D}_M^\Sigma \sqrt{\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2}\psi_M(Z_i)\|^2} \\ &\quad + (1 + \mathcal{D}_M^\Sigma) \|\hat{\beta}_M - \beta_M\|_{\Sigma_M} \sqrt{\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2}X_{i,M}\|^4}. \end{aligned}$$

Note that (using (5.38))

$$\|\Sigma_M^{1/2}\{\hat{\psi}_M(Z_i) - \psi_M(Z_i)\}\| \geq \max_{j \in M} \frac{|\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}(Z_i)|}{\sqrt{(\Sigma_M^{-1})_{jj}}},$$

and hence,

$$\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2}\{\hat{\psi}_M(Z_i) - \psi_M(Z_i)\}\|^2 \geq \max_{j \in M} \frac{1}{n(\Sigma_M^{-1})_{jj}} \sum_{i=1}^n |\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}(Z_i)|^2.$$

Therefore,

$$\max_{j \in M} \frac{1}{n} \sum_{i=1}^n \left(\frac{\hat{\psi}_{j \cdot M}(Z_i) - \psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}} \right)^2 \leq \max_{j \in M} \frac{(\Sigma_M^{-1})_{jj}}{\sigma_{j \cdot M}} \sqrt{\frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{1/2}\{\hat{\psi}_M(Z_i) - \psi_M(Z_i)\}\|^2}.$$

Finally, we observe

$$\max_{j \in M} \frac{(\Sigma_M^{-1})_{jj}}{\sigma_{j \cdot M}} \leq \sup_{\theta \in \mathbb{R}^{|M|}} \sqrt{\frac{\theta^\top \Sigma_M^{-1} \theta}{\theta^\top \Sigma_M^{-1} V_M \Sigma_M^{-1} \theta}} \leq \sqrt{\bar{\lambda}}.$$

This completes the proof.

5.F Proof of Theorem 17

It is clear that $\|\Sigma_M^{-1/2} X_{i,M}\| \leq \lambda_{\max}(\text{Cor}(\Sigma_M^{-1})) \|\tilde{X}_{i,M}\|$ where $\tilde{X}_{i,M}$ is $X_{i,M}$ scaled by their standard deviations. This implies that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2 &\leq \frac{\lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))}{n} \sum_{i=1}^n \|\tilde{X}_{i,M}\|^2 Y_i^2 \\ &= \frac{\lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))}{n} \sum_{i=1}^n \sum_{j \in M} \tilde{X}_{i,j}^2 Y_i^2. \end{aligned}$$

Hence the maximum over all $M \subseteq \{1, 2, \dots, p\}$ with $|M| = s$ is upper bounded by

$$\max_{|M|=s} \lambda_{\max}(\text{Cor}(\Sigma_M^{-1})) \times \sum_{j=1}^s \left(\frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(j)} Y_i^2 \right),$$

where

$$\frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(1)} Y_i^2 \geq \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(2)} Y_i^2 \geq \dots \geq \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(p)} Y_i^2,$$

is a reordering of $n^{-1} \sum_{i=1}^n \tilde{X}_{i,j}^2 Y_i^2$, $1 \leq j \leq p$. Define, for convenience,

$$\mathcal{E}_s := \max_{|M|=s} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2}{s \lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))}.$$

Note that

$$\begin{aligned} 8\mathbb{E} \left[\max_{1 \leq i \leq n} \|\tilde{X}_i\|_{\infty}^2 |Y_i|^2 \right] &\leq 8n^{2/(q-\eta)} \max_{1 \leq i \leq n} \left(\mathbb{E}[\|\tilde{X}_i\|_{\infty}^{q-\eta} |Y_i|^{q-\eta}] \right)^{2/(q-\eta)} \\ &\leq 8n^{2/(q-\eta)} \max_{1 \leq i \leq n} \left\| \max_{1 \leq j \leq p} |\tilde{X}_{i,j}| \right\|_{q(q-\eta)/\eta}^2 \|Y_i\|_q^2 \quad (5.49) \\ &\leq Cn^{2/(q-\eta)} K_x^2 (\log(ep))^{2/\beta} (q(q-\eta)/\eta)^{2/\beta} K_q^2. \end{aligned}$$

This bound holds for every $\eta > 0$ and taking $\eta = 2/(\beta \log n)$ yields

$$8\mathbb{E} \left[\max_{1 \leq i \leq n} \|\tilde{X}_i\|_\infty^2 |Y_i|^2 \right] \leq C_{q,\beta} K_x^2 K_q^2 n^{2/q} \log n (\log(ep))^{2/\beta} =: B, \quad (5.50)$$

for a constant $C_{q,\beta} > 0$ depending only on q, β . Using B , we write

$$\mathcal{E}_s \leq \mathcal{E}_{s,1} + \mathcal{E}_{s,2}, \quad (5.51)$$

where

$$\begin{aligned} \mathcal{E}_{s,1} &:= \max_{|M|=s} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2}{s \lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))} \mathbb{1}\{\|\tilde{X}_i\|_\infty^2 |Y_i|^2 \leq B\}, \\ &\leq \frac{1}{s} \sum_{j=1}^s \left(\frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(j)}^2 Y_i^2 \mathbb{1}\{\|\tilde{X}_i\|_\infty^2 Y_i^2 \leq B\} \right), \\ \mathcal{E}_{s,2} &:= \max_{|M|=s} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^2 Y_i^2}{s \lambda_{\max}(\text{Cor}(\Sigma_M^{-1}))} \mathbb{1}\{\|\tilde{X}_i\|_\infty^2 |Y_i|^2 > B\} \\ &\leq \frac{1}{s} \sum_{j=1}^s \left(\frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,(j)}^2 Y_i^2 \mathbb{1}\{\|\tilde{X}_i\|_\infty^2 Y_i^2 > B\} \right). \end{aligned}$$

The bound for $\mathcal{E}_{s,2}$ is similar to the technique used to bound $\mathcal{E}_{2,\text{Norm}}$ in the proof of Theorem 16. From the definition of B , it follows that

$$\mathbb{P} \left(\bigcup_{s=1}^p \{\mathcal{E}_{s,2} \neq 0\} \right) \leq \mathbb{P} \left(\max_{1 \leq i \leq n} \|\tilde{X}_i\|_\infty^2 Y_i^2 > B \right) \leq \frac{1}{8}.$$

Hence Proposition 6.8 of [Ledoux and Talagrand \(1991\)](#) yields

$$\mathbb{E} \left[\max_{1 \leq s \leq p} \mathcal{E}_{s,2} \right] \leq \frac{8}{n} \mathbb{E} \left[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |\tilde{X}_{i,j}|^2 Y_i^2 \right] \leq \frac{8B}{n}.$$

We now apply Theorem 8 of [Boucheron et al. \(2005\)](#) to conclude

$$\left\| \max_{1 \leq s \leq p} \mathcal{E}_{s,2} \right\|_{(q-\eta)/2} \leq 2\mathbb{E} \left[\max_{1 \leq s \leq p} \mathcal{E}_{s,2} \right] + \frac{C(q-\eta)}{n} \left\| \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |\tilde{X}_{i,j}|^2 |Y_i|^2 \right\|_{(q-\eta)/2}.$$

From the calculation (5.49) (by taking $\eta = 2/(\beta \log n)$), we get

$$\frac{C(q-\eta)}{n} \left\| \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |\tilde{X}_{i,j}|^2 |Y_i|^2 \right\|_{(q-\eta)/2} \leq \frac{C_{q,\beta} K_x^2 K_q^2 \log n (\log(ep))^{2/\beta}}{n^{1-2/q}}.$$

Hence

$$\left\| \max_{1 \leq s \leq p} \mathcal{E}_{s,2} \right\|_{q/2-1/(\beta \log n)} \leq \frac{C_{q,\beta} K_x^2 K_q^2 \log n (\log(ep))^{2/\beta}}{n^{1-2/q}}.$$

By Markov's inequality,

$$\mathbb{P} \left(\max_{1 \leq s \leq p} \mathcal{E}_{s,2} \geq t \left\| \max_{1 \leq s \leq p} \mathcal{E}_{s,2} \right\|_{q/2-1/(\beta \log n)} \right) \leq \frac{1}{t^{q/2-1/(\beta \log n)}}.$$

With $t = n^{-1/(q/2-1/(\beta \log n))}$, we get

$$\mathbb{P} \left(\max_{1 \leq s \leq p} \mathcal{E}_{s,2} \geq \frac{C_{q,\beta} K_x^2 K_q^2 (\log(ep))^{2/\beta} \log n}{n^{1-4/q}} \right) \leq \frac{1}{n}.$$

This proves

$$\mathbb{P} \left(\max_{1 \leq s \leq p} \mathcal{E}_{s,2} \geq \frac{CK_x^2 K_q^2 (\log(ep))^{2/\beta} (\log n)}{n^{1-4/q}} \right) \leq \frac{1}{n}. \quad (5.52)$$

To bound $\mathcal{E}_{s,1}$ in (5.51), define

$$U_j := \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,j}^2 Y_i^2 \mathbb{1}_{\{\|\tilde{X}_i\|_\infty^2 |Y_i|^2 \leq B\}}.$$

An application of Theorem 8 of [Boucheron et al. \(2005\)](#) (with $\theta = 1$) yields for every

$r \geq 1$

$$\|U_j\|_r \leq \frac{2}{n} \sum_{i=1}^n \mathbb{E}[\tilde{X}_{i,j}^2 Y_i^2] + \frac{2r}{n} \left\| \max_{1 \leq i \leq n} \|\tilde{X}_i\|_\infty^2 Y_i^2 \right\|_r \leq \frac{2}{n} \sum_{i=1}^n \mathbb{E}[\tilde{X}_{i,j}^2 Y_i^2] + \frac{2r}{n} B.$$

By Markov's inequality, this proves

$$\mathbb{P} \left(U_j \geq \frac{2}{n} \sum_{i=1}^n \mathbb{E}[\tilde{X}_{i,j}^2 Y_i^2] + \frac{2tB}{n} \right) \leq 3e^{-t} \quad \text{for all } t \geq 0. \quad (5.53)$$

Note that

$$\mathbb{E}[\tilde{X}_{i,j}^2 Y_i^2] \leq \|\tilde{X}_{i,j}\|_4^2 \|Y_i\|_4^2 \leq CK_x^2 K_q^2 \quad \Rightarrow \quad \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\tilde{X}_{i,j}^2 Y_i^2] \leq CK_x^2 K_q^2,$$

for a constant $C > 0$ if $q \geq 4$. Taking $t = \log(3p^2 n/j)$ in (5.53), we get

$$\mathbb{P} \left(U_j \geq 2CK_x^2 K_q^2 + \frac{2B \log(3p^2 n/j)}{n} \right) \leq 3e^{-\log(6p^2 n/j)} = \frac{3j}{3p^2 n}.$$

Now employing the union bound yields

$$\mathbb{P} \left(\bigcup_{j=1}^p \left\{ U_j \geq 2CK_x^2 K_q^2 + \frac{2B \log(3p^2 n/j)}{n} \right\} \right) \leq \frac{1}{np^2} \sum_{j=1}^p j = \frac{p(p+1)}{2np^2} \leq \frac{1}{n}.$$

Hence with probability at least $1 - 1/n$, for all $1 \leq j \leq p$,

$$U_j \leq 2CK_x^2 K_q^2 + \frac{4B \log(3pn/j)}{n}.$$

On this event, we have for all $1 \leq s \leq p$,

$$\frac{1}{s} \sum_{j=1}^s U_{(j)} \leq 2CK_x^2 K_q^2 + \frac{4B}{ns} \sum_{j=1}^s \log(3pn/j) \leq 2CK_x^2 K_q^2 + \frac{CB}{n} \log(epn/s),$$

where the last inequality follows from the fact that $\sum_{j=1}^s \log(1/j) = -\log(s!) \leq -s \log(s/e)$ (because $e^s \geq s^s/s!$). From the definition (5.51) of $\mathcal{E}_{s,1}$ and (5.50), we conclude with probability at least $1 - 1/n$, for all $1 \leq s \leq p$,

$$\mathcal{E}_{s,1} \leq CK_x^2 K_q^2 \left[1 + \frac{\log(epn/s) \log n (\log(ep))^{2/\beta}}{n^{1-2/q}} \right].$$

Combining this inequality with (5.52) and (5.51) proves the first result.

To prove the second result, note that

$$\begin{aligned} \|\Sigma_M^{-1/2} X_{i,M}\|^4 &\leq |M|^2 \lambda_{\max}(\text{Cor}(\Sigma_M^{-1})) \left(\frac{1}{|M|} \sum_{j \in M} \tilde{X}_{i,j}^2 \right)^2 \\ &\leq |M|^2 \lambda_{\max}^2(\text{Cor}(\Sigma_M^{-1})) \left(\frac{1}{|M|} \sum_{j \in M} \tilde{X}_{i,j}^4 \right). \end{aligned}$$

Therefore,

$$\max_{\substack{1 \leq s \leq p, \\ |M|=s}} \frac{1}{n} \sum_{i=1}^n \frac{\|\Sigma_M^{-1/2} X_{i,M}\|^4}{|M|^2 \lambda_{\max}^2(\text{Cor}(\Sigma_M^{-1}))} \leq \max_{1 \leq j \leq p} \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,j}^4.$$

For each j , Theorem 8 of [Boucheron et al. \(2005\)](#) yields

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,j}^4 \geq CK_x^4 + \frac{CK_x^4 t^{1+4/\beta} (\log(en))^{4/\beta}}{n} \right) \leq 3e^{-t}, \quad \text{for all } t \geq 0.$$

Union bound proves

$$\mathbb{P} \left(\max_{1 \leq j \leq p} \frac{1}{n} \sum_{i=1}^n \tilde{X}_{i,j}^4 \geq CK_x^4 \left[1 + \frac{(\log(3pn))^{1+4/\beta} (\log(en))^{4/\beta}}{n} \right] \right) \leq \frac{1}{pn}.$$

This completes the proof of the second result.

To prove the third result, observe that

$$\frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} = \frac{e_j^\top \Sigma_M^{-1} X_{i,M} (Y_i - X_{i,M}^\top \beta_M) (Y_i - X_{i,M'}^\top \beta_{M'}) e_{j'}^\top \Sigma_{M'}^{-1} X_{i,M'}}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}}.$$

Assumption **(X-SW)** and **(Σ -V)** yields

$$\mathbb{E} \left[\exp \left(\frac{|e_j^\top \Sigma_M^{-1} X_{i,M}|^\beta}{(\sqrt{\lambda} \sigma_{j \cdot M} K_x)^\beta} \right) \right] \leq 2.$$

Proposition B.1 of [Kuchibhotla and Patra \(2019\)](#) yields

$$\begin{aligned} \mathbb{E} [\Delta_M^{\text{Boot}}] &\leq C \mathfrak{B}_M^{1/2} \sqrt{\frac{\log(|\mathcal{Q}|)}{n}} \\ &\quad + C \left(\frac{\log(|\mathcal{Q}|)}{n} \right)^{1-2/(q-\eta)} \max_{1 \leq i \leq n} \left(\mathbb{E} \left[\max_{M, M', j, j'} \left| \frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} \right|^{(q-\eta)/2} \right] \right)^{2/(q-\eta)}. \end{aligned}$$

This holds for any $\eta > 0$ and we take $\eta = 1/\log n$. We now bound the second term on the right hand side as follows:

$$\begin{aligned} \max_{M, M', j, j'} \left| \frac{\psi_{j \cdot M}(Z_i) \psi_{j' \cdot M'}(Z_i)}{\sigma_{j \cdot M} \sigma_{j' \cdot M'}} \right| &\leq \max_{M, j} \left| \frac{\psi_{j \cdot M}(Z_i)}{\sigma_{j \cdot M}} \right|^2 \\ &\leq \max_{M, j} \left| \frac{e_j^\top \Sigma_M^{-1} X_{i,M}}{\sigma_{j \cdot M}} \right|^2 Y_i^2 + \max_{M, j} \left| \frac{e_j^\top \Sigma_M^{-1} X_{i,M} X_{i,M}^\top \beta_M}{\sigma_{j \cdot M}} \right|^2. \end{aligned}$$

Hölder's inequality yields

$$\begin{aligned} \left\| \max_{M, j} \left| \frac{e_j^\top \Sigma_M^{-1} X_{i,M}}{\sigma_{j \cdot M}} \right|^2 Y_i^2 \right\|_{(q-\eta)/2} &\leq \left\| \max_{M, j} \left| \frac{e_j^\top \Sigma_M^{-1} X_{i,M}}{\sigma_{j \cdot M}} \right| \right\|_{q(q-\eta)/\eta}^2 \|Y_i\|_q^2 \\ &\leq C \bar{\lambda} K_q^2 K_x^2 (\log(|\mathcal{Q}|))^{2/\beta} (q^2/\eta)^{2/\beta} \\ &= C \bar{\lambda} K_q^2 K_x^2 (\log(|\mathcal{Q}|))^{2/\beta} (q^2 \log n)^{2/\beta}. \end{aligned} \tag{5.54}$$

Further

$$\left\| \max_{M, j} \left| \frac{e_j^\top \Sigma_M^{-1} X_{i,M} X_{i,M}^\top \beta_M}{\sigma_{j \cdot M}} \right| \right\|_{q-\eta}^2 \leq C \bar{\lambda} K_x^4 K_q^2 (\log(|\mathcal{Q}|))^{4/\beta} (q-\eta)^{4/\beta}. \tag{5.55}$$

Therefore,

$$\mathbb{E} [\Delta_{\mathcal{M}}^{\text{Boot}}] \leq C \mathfrak{B}_{\mathcal{M}}^{1/2} \sqrt{\frac{\log(|\mathcal{Q}|)}{n}} + C \bar{\lambda} K_x^4 K_q^2 \left(\frac{\log(|\mathcal{Q}|)}{n} \right)^{1-2/q} (\log(|\mathcal{Q}|))^{2/\beta} (\log(n|\mathcal{Q}|))^{2/\beta}.$$

The tail bound can now be obtained from [Einmahl and Li \(2008, Theorem 3.1\)](#). This result implies

$$\begin{aligned} & \mathbb{P} (\Delta_{\mathcal{M}}^{\text{Boot}} \geq 2\mathbb{E} [\Delta_{\mathcal{M}}^{\text{Boot}}] + t) \\ & \leq \exp \left(-\frac{nt^2}{3\mathfrak{B}_{\mathcal{M}}} \right) + Cn \max_{1 \leq i \leq n} \left(\frac{\| \max_{\mathcal{M},j} |\psi_{j \cdot \mathcal{M}}(Z_i) / \sigma_{j \cdot \mathcal{M}} \|_{q-1/\log n}^2}{nt} \right)^{(q-1/\log n)/2} \\ & \leq \exp \left(-\frac{nt^2}{3\mathfrak{B}_{\mathcal{M}}} \right) + Cn \left(\frac{C \bar{\lambda} K_x^4 K_q^2 (\log(|\mathcal{Q}|))^{2/\beta} (\log(n|\mathcal{Q}|))^{2/\beta}}{nt} \right)^{(q-1/\log n)/2}, \end{aligned}$$

where the second inequality follows from [\(5.54\)](#) and [\(5.55\)](#). Hence taking

$$t = \sqrt{\frac{3\mathfrak{B}_{\mathcal{M}} \log(n)}{n}} + C \frac{\bar{\lambda} K_x^4 K_q^2 (\log(|\mathcal{Q}|))^{2/\beta} (\log(n|\mathcal{Q}|))^{2/\beta}}{n^{1-3/q}},$$

proves that with probability at least $1 - 1/n - 1/\sqrt{n}$,

$$\Delta_{\mathcal{M}}^{\text{Boot}} \leq 2\mathbb{E} [\Delta_{\mathcal{M}}^{\text{Boot}}] + \sqrt{\frac{3\mathfrak{B}_{\mathcal{M}} \log(n)}{n}} + C \frac{\bar{\lambda} K_x^4 K_q^2 (\log(|\mathcal{Q}|))^{2/\beta} (\log(n|\mathcal{Q}|))^{2/\beta}}{n^{1-3/q}}.$$

This proves the last result.

End of Chapter 5.

Chapter 6

Real Data Examples

In this chapter, we provide two real data examples to illustrate the practical worth of the valid post-selection confidence regions discussed in this thesis: Boston housing data (where the target is inference after variable transformation) and Telomere length data (where the target is inference after covariate selection). Both of these analysis can be found at

1. Boston Housing Data: https://github.com/post-selection-inference/R/tree/master/case_study/boston_housing and
2. Telomere Length Data: https://github.com/post-selection-inference/R/tree/master/case_study/MTL.

6.1 Boston housing data

This section demonstrates PoSI on the Boston Housing data from [Harrison Jr and Rubinfeld \(1978\)](#) with amendments by [Gilley et al. \(1996\)](#). The original study intends to investigate the willingness to pay for clean by the “hedonic housing price equation” incorporating the level of air pollution, specifically the concentration of nitrogen oxides (NOX), as one attribute and concludes that “the valuation placed on a marginal

improvement in air quality is quite sensitive to the specification of the hedonic housing value equation”. In other words, NOX is a significant factor to the housing price controlling other covariates. The final housing value equation is chosen as the best fit after comparing models with transformations of both the response (Median value of house prices) and covariates, which invalidates the classical inference. After our post-selection adjustment to the inference, the conclusion that NOX is a significant factor at 0.05 level to the housing price controlling for other covariates still holds, but the Charles River dummy and racial diversity become insignificant. See Table 6.1 for the description of the data.

Variable	Description
CRIM	crime rate
ZN	proportion of residential land zoned for lots over 25,000 sq. ft
INDUS	proportion of non-retail business acres
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	nitrogen oxides concentration, pphm
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centers
RAD	index of accessibility to radial highways
TAX	full-value property tax rate per \$10,000
PTRATIO	pupil teacher ratio
B	$1000 \times (Bk - 0.63)^2$ where Bk is the proportion of blacks
LSTAT	percent lower status population
MEDV	median value of owner occupied homes in \$1000's

Table 6.1: Variables in the Boston housing data by [Harrison Jr and Rubinfeld \(1978\)](#).

We provide PoSI adjustment considering the following transformations mentioned in [Harrison Jr and Rubinfeld \(1978\)](#). After grid search for Box-Cox transformation on NOX, they use NOX^2 because “the statistical fit is the best”.

- Log transformation of the response MEDV;
- Grid search of Box-Cox transformation of the variable of interest NOX;
- Log, linear, quadratic transformation of RM.

After HPoSI or HPoSI1 adjustment, RIVER and BLK change from significant at 0.05 level to insignificant. The conclusion for the variable of interest NOX does not change. See [Table 6.3](#), [6.4](#) and [6.5](#) for summary tables for unadjusted confidence interval, HPoSI and HPoSI1.

Variable	Unadjusted	Adjusted
NOX	✓	✓
RM	✓	✓
AGE	×	×
CRIM	✓	✓
RES	×	×
INDUS	×	×
RIVER	✓	×
TAX	✓	✓
PTR	✓	✓
BLK	✓	×
LSTAT	✓	✓
DEMPC	✓	✓
DRADH	✓	✓

Variable Unadjusted Adjusted

Table 6.2: Significance at 0.05 level of variables in the final model with and without adjustment.

Variable	Lower	Upper	Significance	K	<i>p</i> -value
NOX	-0.86	-0.42	TRUE	1.96	0.00
RM	0.00	0.01	TRUE	1.96	0.00
AGE	-0.00	0.00	FALSE	1.96	0.89
CRIM	-0.01	-0.01	TRUE	1.96	0.00
RES	-0.00	0.00	FALSE	1.96	0.85
INDUS	-0.00	0.00	FALSE	1.96	0.94
RIVER	0.03	0.16	TRUE	1.96	0.01
TAX	-0.00	-0.00	TRUE	1.96	0.00
PTR	-0.04	-0.02	TRUE	1.96	0.00
BLK	0.00	0.00	TRUE	1.96	0.00
LSTAT	-0.42	-0.33	TRUE	1.96	0.00
DEMPC	-0.26	-0.13	TRUE	1.96	0.00
DRADH	0.05	0.13	TRUE	1.96	0.00

Table 6.3: Unadjusted confidence interval for the final model.

Variable	Lower	Upper	Significance	K	<i>p</i> -value
NOX	-1.00	-0.28	TRUE	3.00	0.00
RM	0.00	0.01	TRUE	3.00	0.03
AGE	-0.00	0.00	FALSE	3.00	1.00
CRIM	-0.02	-0.00	TRUE	3.00	0.00
RES	-0.00	0.00	FALSE	3.00	1.00
INDUS	-0.01	0.01	FALSE	3.00	1.00
RIVER	-0.01	0.20	FALSE	3.00	0.16
TAX	-0.00	-0.00	TRUE	3.00	0.00
PTR	-0.04	-0.02	TRUE	3.00	0.00
BLK	-0.00	0.00	FALSE	3.00	0.20
LSTAT	-0.48	-0.27	TRUE	3.00	0.00
DEMPC	-0.31	-0.08	TRUE	3.00	0.00
DRADH	0.04	0.14	TRUE	3.00	0.00

Table 6.4: HPoSI (simultaneous) for the final model.

Variable	Lower	Upper	Significance	K	p -value
NOX	-1.00	-0.28	TRUE	2.90	0.00
RM	0.00	0.01	TRUE	2.98	0.02
AGE	-0.00	0.00	FALSE	2.99	1.00
CRIM	-0.02	-0.01	TRUE	2.96	0.00
RES	-0.00	0.00	FALSE	3.01	1.00
INDUS	-0.00	0.01	FALSE	2.98	1.00
RIVER	-0.01	0.20	FALSE	2.99	0.14
TAX	-0.00	-0.00	TRUE	2.98	0.00
PTR	-0.04	-0.02	TRUE	2.95	0.00
BLK	-0.00	0.00	FALSE	2.96	0.16
LSTAT	-0.48	-0.27	TRUE	2.99	0.00
DEMPC	-0.31	-0.08	TRUE	2.92	0.00
DRADH	0.04	0.14	TRUE	2.95	0.00

Table 6.5: HPoSI (marginal) for the final model.

6.2 Telomere length example

This section demonstrates PoSI on the Telomere Length (TL) analysis by [Nersisyan et al. \(2019\)](#). The goal of the analysis is to study the inheritance patterns and associated genetic factors. The data and scripts of their analyses are available on [Github](#). Here we focus on the first part of the multiple linear regression (MLR) analysis to understand the TL inheritance patterns. The article concludes from the analysis that TL is a mostly heritable trait, more from mother's and less from father's. TL is also age-related and linked to mother's age at conception. The conclusion is based on regression analysis but with covariate selection that invalidates the classical

inference. After our post-selection adjustment to the inference, only the inheritance factors are significant controlling for other covariates.

There are in total of 250 family trios in the original data. According to the article, two of the families with missing data for the mother were removed, and two families with discordant age differences at the time of data collection and at conception were also discarded. Hence there are 246 samples left in the `child.df` data. See Table 6.6 for the description of the data.

Variable	Description
GoNL_Sample_ID	id of Genome of the Netherlands (GoNL) project
family	family ID
member	family member (here are all children, c)
Sex	sex
Age	age
MTL	mean telomere length
ageC	NA
MAC	mother's age at conception
PAC	father's age at conception
mMTL	mother's mean telomere length
fMTL	father's mean telomere length
mAge	mother's age
fAge	father's age

Table 6.6: Variables in the Telomere Length (TL) analysis by [Nersisyan et al. \(2019\)](#).

In their MLR analysis, they consider Sex, Age, mMTL, fMTL, MAC, and PAC along with all the pairwise interactions. They land on the final model with Age,

mMTL, fMTL, MAC, and PAC after model selection based on adjusted R^2 value and parsimony and claim significance at 0.05 level for Age, mMTL, fMTL and MAC based on the classical inference on the final model. They conclude that TL is a mostly heritable trait, but also age-related and linked to mother's age at conception.

After HPoSI or HPoSI1 adjustment, Age and MAC change from significant at 0.05 level to insignificant. The adjustment changes the conclusion of the study to TL is a heritable trait only. The heritable nature of telomeres, as the authors claim, echoes prior studies. Age and parents' ages at conception, however, are not significant factors. The authors claim there is a lack of evidence on the effect of MAC in previous studies and their MLR analysis (with covariate selection) confirms the association of MAC with offspring MTL, suggesting that further investigation. Nevertheless, after PoSI adjustment for selection, MAC is not significant anymore. See Table 6.8, 6.9 and 6.10 for summary tables for unadjusted confidence interval, HPoSI and HPoSI1.

Variable	Unadjusted	Adjusted
Age	✓	×
mMTL	✓	✓
fMTL	✓	✓
MAC	✓	×
PAC	×	×

Table 6.7: Significance at 0.05 level of variables in the final model with and without adjustment.

variable	lower	upper	Significance	K	<i>p</i> -value
Age	-40.81	-4.97	TRUE	1.96	0.01
mMTL	0.33	0.57	TRUE	1.96	0.00
fMTL	0.16	0.43	TRUE	1.96	0.00
MAC	33.12	155.35	TRUE	1.96	0.00
PAC	-85.89	28.35	FALSE	1.96	0.32

Table 6.8: Unadjusted confidence interval for the final model.

variable	lower	upper	Significance	K	<i>p</i> -value
Age	-52.83	7.05	FALSE	3.80	0.97
mMTL	0.21	0.68	TRUE	3.80	0.00
fMTL	0.02	0.57	TRUE	3.80	0.00
MAC	-45.97	234.44	FALSE	3.80	1.00
PAC	-163.31	105.78	FALSE	3.80	1.00

Table 6.9: HPoSI (simultaneous) for the final model.

variable	lower	upper	Significance	K	<i>p</i> -value
Age	-48.06	2.27	FALSE	3.19	0.08
mMTL	0.26	0.64	TRUE	3.05	0.00
fMTL	0.06	0.53	TRUE	3.19	0.00
MAC	-23.31	211.78	FALSE	3.19	0.23
PAC	-141.57	84.03	FALSE	3.19	1.00

Table 6.10: HPoSI (marginal) for the final model.

End of Chapter 6.

Chapter 7

Conclusions

This thesis is motivated by the use of invalid statistical inference tools that prevails in the practice of statistics in almost all of empirical sciences. We have provided various different methods of drawing (asymptotically) valid statistical conclusions after performing quite arbitrary exploration of the data. In attaining this goal, we have relaxed many of the existing assumptions on the data generating process such as sparsity, linearity, Gaussianity. In particular, our results provide validity only making tail/moment assumptions on the observations and some weak dependence assumptions. This is an important feature, as in real data analyses, analysts often do not know much about the data distribution.

The main contribution of this work is in providing a unified framework in solving the problem of valid post-selection inference. Most of the existing works in this area concentrate on a particular kind of exploration, if not a particular method of exploration. For instance, [Berk et al. \(2013\)](#) focuses on covariate selection and [Lee et al. \(2016\)](#) focuses on covariate selection by lasso specifically. Unlike these papers, we provide a unified framework which allows for an infinite universe to select from, such as selecting a transformation for the response or the covariates.

Regarding the theoretical analysis, an interesting aspect of the work is that most of it is based on deterministic inequalities. In all of the results, we have reduced the problem to the case of bounding averages of random variables/vectors using only deterministic inequalities (which hold for all realizations of the datasets alike, without any independence/dependence assumptions). Once this is available, all one needs to do is bound the averages under the independence/dependence assumptions. This is a very robust way of deriving results for asymptotic guarantees in statistical inference.

One important component of valid post-selection inference that we did not discuss is computation. Except for Approach 1 discussed in Chapter 4, the procedures PoSI and HPOSI are not computationally efficient. A direct implementation of these methods required complete enumeration of all models of size at most k , which, for p, k large, is NP-hard. An alternative would be to provide an approximate calculation of the maximum statistic, and this can be done through many randomized algorithms such as simulated annealing or greedy approximation. At present, we do not know of any guarantees for these randomized algorithms, which leads to an interesting open problem for further research exploration. One randomized algorithm with guarantees can be constructed as follows: consider the problem of computing the maximum of a sequence $a_1, \dots, a_N \geq 0$ with N large; think of $N = \binom{p}{k} = O(p^k)$. Each a_i corresponds to the absolute t -statistic, $|t_{j,M}|$, of a particular covariate in a particular model. The idea is based on the well-known inequality

$$\max_{1 \leq j \leq N} a_j \leq \left(\sum_{j=1}^N a_j^q \right)^{1/q} \leq N^{1/q} \max_{1 \leq j \leq N} a_j.$$

If we choose $q = \log(N)/\varepsilon$, then

$$\max_{1 \leq j \leq N} a_j \leq \left(\sum_{j=1}^N a_j^q \right)^{1/q} \leq e^\varepsilon \max_{1 \leq j \leq N} a_j.$$

Hence by computing the ℓ_q norm of the sequence, we can approximate the maximum up to a factor of e^ε . Of course, computing the ℓ_q -norm itself is hard because it is a sum over N quantities, which is as hard as computing the maximum. However, we note that

$$\left(\sum_{j=1}^N a_j^q \right)^{1/q} = N^{1/q} \left(\frac{1}{N} \sum_{j=1}^N a_j^q \right)^{1/q} = N^{1/q} (\mathbb{E}[a_J^q])^{1/q},$$

where the expectation is with respect to the random variable J which is uniformly distributed on the set $\{1, 2, \dots, N\}$. Hence computing the ℓ_q -norm is same as computing an expectation of a random variable. Using the idea of Monte Carlo integration, we can draw samples J_1, J_2, \dots, J_m from the uniform distribution on $\{1, 2, \dots, N\}$ and approximate the expectation $\mathbb{E}[a_J^q]$ by

$$\frac{1}{m} \sum_{i=1}^m a_{J_i}^q.$$

An issue is that this average approximates the expectation but does not upper bound or lower bound the expectation. But from the central limit theorem or concentration inequalities, we can provide a confidence interval for the true expectation based on the estimate (the average). Suppose we construct a quantity \hat{C}_δ such that

$$\mathbb{P} \left(\mathbb{E}[a_J^q] \leq \frac{1}{m} \sum_{i=1}^m a_{J_i}^q + \hat{C}_\delta \right) \geq 1 - \delta,$$

then we can use the upper bound to construct an upper bound on the maximum of a_1, \dots, a_N that is a valid upper bound with probability at least $1 - \delta$. See [Huber \(2019\)](#) for references on how to construct upper bounds for the true expectation based on the samples. Preliminary investigations in this direction show some promise, and this is currently under investigation in collaboration with Junhui Cai, my technical advisor.

Given the importance of valid statistical conclusions in practice, we hope this work draws the attention of practitioners to the respective issues as well as the solutions.

Practice Safe Statistics.

End of Chapter 7.

Bibliography

- Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *J. Amer. Statist. Assoc.*, 109(508):1601–1614.
- Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.*, 6:170–176.
- Arcones, M. A. (2005). Convergence of the optimal M -estimator over a parametric family of M -estimators. *Test*, 14(1):281–315.
- Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016). Uniformly valid confidence intervals post-model-selection. *ArXiv e-prints*.
- Bellec, P. C. (2016). Aggregation of supports along the lasso path. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 488–529, Columbia University, New York, New York, USA. PMLR.
- Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv:1806.01888*.

- Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.
- Belloni, A., Rosenbaum, M., and Tsybakov, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):939–956.
- Bentkus, V. (2004). A Lyapunov type bound in \mathbf{R}^d . *Teor. Veroyatn. Primen.*, 49(2):400–410.
- Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468.
- Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686.
- Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., and Zhao, L. (2014). Misspecified mean function regression: making good use of regression models that are wrong. *Sociol. Methods Res.*, 43(3):422–451.
- Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.*, 41(2):802–837.
- Bolt, S., Eadie, T., Yorkston, K., Baylor, C., and Amtmann, D. (2016). Variables associated with communicative participation after head and neck cancer. *JAMA Otolaryngology–Head & Neck Surgery*, 142(12):1145–1151.
- Bolthausen, E., Perkins, E., and van der Vaart, A. (2002). *Lectures on probability theory and statistics*, volume 1781 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 29th Summer School on Probability Theory held in Saint-Flour, July 8–24, 1999, Edited by Pierre Bernard.

- Borwein, J. M. and Chan, O.-Y. (2009). Uniform bounds for the complementary incomplete gamma function. *Math. Inequal. Appl.*, 12(1):115–121.
- Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560.
- Buehler, R. J. and Feddersen, A. P. (1963). Note on a conditional property of Student’s t . *Ann. Math. Statist.*, 34:1098–1100.
- Buja, A., Berk, R., Brown, L., George, E., Kuchibhotla, A. K., and Zhao, L. (2016). Models as Approximations — Part II: A General Theory of Model-Robust Regression. *ArXiv e-prints*.
- Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and Zhao, L. (2014). Models as Approximations, Part I: A Conspiracy of Nonlinearity and Random Regressors in Linear Regression. *ArXiv e-prints*.
- Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351.
- Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28, pages 774–782.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2017a). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, 45(4):2309–2352.
- Chernozhukov, V., Chetverikov, D., and Kato, K. (2017b). Detailed proof of nazarov’s inequality. *arXiv preprint arXiv:1711.10696*.

- Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2019). Improved central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:1912.10529*.
- Chowdhury, M. Z. I. and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1).
- Claeskens, G. and Carroll, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 94(2):249–265.
- Cui, Y., Leng, C., and Sun, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Comput. Statist. Data Anal.*, 93:390–403.
- DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer Texts in Statistics. Springer, New York.
- de la Peña, V. H. and Giné, E. (1999). *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York. From dependence to independence, Randomly stopped processes. U -statistics and processes. Martingales and beyond.
- Deng, H. and Zhang, C.-H. (2017). Beyond gaussian approximation: Bootstrap for maxima of sums of independent random vectors. *arXiv preprint arXiv:1705.09528*.
- Devroye, L. (1982). Bounds for the uniform deviation of empirical measures. *J. Multivariate Anal.*, 12(1):72–79.
- Dodge, Y. and Jurevckova, J. (2000). *Adaptive regression*. Springer-Verlag, New York.

- Durrett, R. (2010). *Probability: theory and examples*, volume 31 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, fourth edition.
- Einmahl, U. and Li, D. (2008). Characterization of lil behavior in banach space. *Transactions of the American Mathematical Society*, 360(12):6677–6693.
- Elker, J., Pollard, D., and Stute, W. (1979). Glivenko-Cantelli theorems for classes of convex sets. *Adv. in Appl. Probab.*, 11(4):820–833.
- Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, 21(4):487–502.
- Fang, X. and Koike, Y. (2020). High-dimensional central limit theorems by stein’s method. *arXiv preprint arXiv:2001.10917*.
- Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.
- Freedman, D. A. (1981). Bootstrapping regression models. *Ann. Statist.*, 9(6):1218–1228.
- Freedman, D. A. (1983). A note on screening regression equations. *Amer. Statist.*, 37(2):152–155.
- Gallant, A. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. B. Blackwell.
- Gilley, O. W., Pace, R. K., et al. (1996). On the harrison and rubinfeld data. *Journal of Environmental Economics and Management*, 31(3):403–405.

- Guédon, O., Litvak, A. E., Pajor, A., and Tomczak-Jaegermann, N. (2015). On the interval of fluctuation of the singular values of random matrices. *arXiv preprint arXiv:1509.02322*.
- Harrison Jr, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.:(United States)*, 5(1).
- Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.*, 98(464):879–899.
- Hörmann, S. (2009). Berry-Esseen bounds for econometric time series. *ALEA Lat. Am. J. Probab. Math. Stat.*, 6:377–397.
- Hu, F. and Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canad. J. Statist.*, 28(3):449–499. With discussion and rejoinder by the authors.
- Huber, M. (2019). An optimal (ϵ, δ) -randomized approximation scheme for the mean of random variables with bounded relative variance. *Random Structures & Algorithms*, 55(2):356–370.
- Koike, Y. (2019). Notes on the dimension dependence in high-dimensional central limit theorems for hyperrectangles. *arXiv preprint arXiv:1911.00160*.
- Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. Springer, New York.
- Kuchibhotla, A. K. (2018). Deterministic inequalities for smooth m-estimators. *arXiv preprint arXiv:1809.05172*.
- Kuchibhotla, A. K., Brown, L. D., and Buja, A. (2018). Assumption-lean linear regression. *Unpublished Manuscript*, pages 1–30.

- Kuchibhotla, A. K., Brown, L. D., Buja, A., and Cai, J. (2019). All of linear regression. *arXiv preprint arXiv:1910.06386*.
- Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2018). A Model Free Perspective for Linear Regression: Uniform-in-model Bounds for Post Selection Inference. *ArXiv e-prints*.
- Kuchibhotla, A. K., Brown, L. D., Buja, A., George, E. I., and Zhao, L. (2020). Valid post-selection inference in model-free linear regression. *Annals of Statistics (to appear)*. arXiv:1806.04119.
- Kuchibhotla, A. K. and Chakraborty, A. (2018). Moving Beyond Sub-Gaussianity in High-Dimensional Statistics: Applications in Covariance Estimation and Linear Regression. *ArXiv e-prints:1804.02605*.
- Kuchibhotla, A. K., Mukherjee, S., and Banerjee, D. (2018). High-dimensional ckt: Improvements, non-uniform extensions and large deviations. *arXiv preprint arXiv:1806.06153*.
- Kuchibhotla, A. K. and Patra, R. K. (2019). On least squares estimation under heteroscedastic and heavy-tailed errors. *arXiv preprint arXiv:1909.02088*.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.
- Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.
- Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

- Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591.
- Levit, B. Y. (1976). On the efficiency of a class of non-parametric estimates. *Theory of Probability & Its Applications*, 20(4):723–740.
- Liu, R. Y. and Singh, K. (1995). Using i.i.d. bootstrap inference for general non-i.i.d. models. *J. Statist. Plann. Inference*, 43(1-2):67–75.
- Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Economic Theory*, 26(4):1218–1245.
- Liu, W., Xiao, H., and Wu, W. B. (2013). Probability and moment inequalities under dependence. *Statist. Sinica*, 23(3):1257–1272.
- Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664.
- Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.
- Lunde, R. (2019). Sample splitting and weak assumption inference for time series. *arXiv preprint arXiv:1902.07425*.
- Lydersen, S. (2014). Statistical review: frequently given comments. *Annals of the rheumatic diseases*, pages annrheumdis–2014.
- McNeney, W. B. (1998). *Asymptotic efficiency in semiparametric models with non-i.i.d. data*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Washington.

- Moore, D. and McCabe, G. (1998). *Introduction to the Practice of Statistics*. W. H. Freeman.
- Nersisyan, L., Nikoghosyan, M., and Arakelyan, A. (2019). Wgs-based telomere length analysis in dutch family trios implicates stronger maternal inheritance and a role for rrm1 gene. *Scientific Reports*, 9(1):1–9.
- Newey, W. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135.
- Olshen, R. A. (1973). The conditional level of the F -test. *J. Amer. Statist. Assoc.*, 68:692–698.
- Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, 16(2).
- Paulauskas, V. and Rackauskas, A. (1989). *Approximation theory in the central limit theorem*, volume 32 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht. Exact results in Banach spaces, Translated from the Russian by B. Svecevičius and Paulauskas.
- Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York.
- Pötscher, B. M. (2002). Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots, and estimation of long memory parameters. *Econometrica*, 70(3):1035–1065.
- Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic nonlinear econometric models*. Springer-Verlag, Berlin. Asymptotic theory.

- Rencher, A. C. and Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, 22(1):49–53.
- Rinaldo, A., Wasserman, L., G'Sell, M., Lei, J., and Tibshirani, R. (2016). Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Free Inference. *ArXiv e-prints*.
- Rio, E. (2009). Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1):146–163.
- Romano, J. P. and Wolf, M. (2000). A more general central limit theorem for m -dependent random variables with unbounded m . *Statist. Probab. Lett.*, 47(2):115–124.
- Rosenbaum, M. and Tsybakov, A. B. (2010). Sparse recovery under matrix uncertainty. *Ann. Statist.*, 38(5):2620–2651.
- Shao, Q.-M. (2000). A comparison theorem on moment inequalities between negatively associated and independent random variables. *J. Theoret. Probab.*, 13(2):343–356.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.
- Statulevicius, V., editor (2000). *Limit theorems of probability theory*. Springer-Verlag, Berlin. Translation of it Probability theory. 6 (Russian), Itogi Nauki i Tekhniki, Sovrem. Probl. Mat. Fund. Naprav., 81, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Inform. (VINITI), Moscow, 1991 [MR1157205 (92k:60001)], Translation edited by Yu. V. Prokhorov and V. Statulevičius.

- Stine, R. and Foster, D. (2013). *Statistics for Business: Decision Making and Analysis*. Always learning. Pearson Education.
- Tian, X., Bi, N., and Taylor, J. (2016). MAGIC: a general, powerful and tractable method for selective inference. *ArXiv e-prints*.
- Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scand. J. Stat.*, 44(2):480–499.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.
- Tibshirani, R. J., Rinaldo, A., Tibshirani, R., Wasserman, L., et al. (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287.
- Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.*, 111(514):600–620.
- Tullock, G. (2001). A comment on daniel klein’s” a plea to economists who favor liberty”. *Eastern Economic Journal*, 27(2):203–207.
- van de Geer, S. and Muro, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8(2):3031–3061.
- Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *J. Theoret. Probab.*, 25(3):655–686.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

- White, H. (2001). *Asymptotic Theory for Econometricians*. Economic theory, econometrics, and mathematical economics. Academic Press.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of animal ecology*, 75(5):1182–1189.
- Wiens, M., Kumbakumba, E., Larson, C., Ansermino, J., Singer, J., Kissoon, N., Wong, H., Ndamira, A., Kabakyenga, J., Kiwanuka, J., et al. (2015). Postdischarge mortality in children with acute infectious diseases: derivation of postdischarge mortality prediction models. *BMJ open*, 5(11):e009449.
- Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natl. Acad. Sci. USA*, 102(40):14150–14154.
- Wu, W. B. and Mielniczuk, J. (2010). A new look at measuring dependence. In *Dependence in probability and statistics*, volume 200 of *Lect. Notes Stat.*, pages 123–142. Springer, Berlin.
- Wu, W.-B. and Wu, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.*, 10(1):352–379.
- Yuan, K.-H. and Jennrich, R. I. (1998). Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.*, 65(2):245–260.
- Zhang, D. and Wu, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.*, 45(5):1895–1919.
- Zhang, K. (2012). *Valid Post-selection Inference*. PhD thesis, Uni-

versity of Pennsylvania, Publicly Accessible Penn Dissertations. 598.
<https://repository.upenn.edu/edissertations/598>.

Zhang, X. and Cheng, G. (2014). Bootstrapping High Dimensional Time Series.
ArXiv e-prints.

Zhang, X., Cheng, G., et al. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli*, 24(4A):2640–2675.