Vol. 0 (2020) 1–8

Optimal selection of sample-size dependent common subsets of covariates for multi-task regression prediction

David Azriel

Faculty of Industrial Engineering and Management,
Technion Israel Institute of Technology, Haifa, Israel
e-mail: davidazr@technion.ac.il

Yosef Rinott

Department of Statistics and Center for the Study of Rationality
The Hebrew University, Jerusalem, Israel
e-mail: yosef.rinott@mail.huji.ac.il

Abstract:

An analyst is given a training set consisting of regression datasets D_j of different sizes, which are distributed according to some G_j, $j = 1, \ldots, \mathcal{J}$, where the distributions G_j are assumed to form a random sample generated by some common source. In particular, the D_j's have a common set of covariates and they are all labeled. The training set is used by the analyst for selection of subsets of covariates denoted by $\mathcal{P}^*(n)$, whose role is described next.

The multi-task problem we consider is as follows: given a number of random labeled datasets (which may be in the training set or not) D_{J_k} of size n_k, $k = 1, \ldots, K$, estimate separately for each dataset the regression coefficients on the subset of covariates $\mathcal{P}^*(n_k)$ and then predict future dependent variables given their covariates.

Naturally, a large sample size n_k of D_{J_k} allows a larger subset of covariates, and the dependence of the size of the selected covariate subsets on n_k is needed in order to achieve good prediction and avoid overfitting. Subset selection is notoriously difficult and computationally demanding, and requires large samples; using all the regression datasets in the training set together amounts to borrowing strength toward better selection under suitable assumptions. Furthermore, using common subsets for all regressions having a given sample size standardizes and simplifies the data collection and avoids having to select and use a different subset for each prediction task. Our approach is efficient when the relevant covariates for prediction are common to the different regressions, while the models' coefficients may vary between different regressions.

Last but not least, we propose a simple and meaningful measure, GENO, that allows comparisons of the predictive value of different subsets of covariates by comparing the sample size they require in order to achieve the same prediction error.

MSC2020 subject classifications: 62J99.
Keywords and phrases: random covariates, model selection, Mallows C_p, equivalent number of observations (ENO), GENO, transfer learning, overfitting.

arXiv: 2012.05949

Contents

Appendix B: A table of notation1
1 Introduction . . . 2
1.1 A general description of the problem . . . 2
1.2 A formal setup . . . 5
1.3 GENO, a measure of usefulness . . . 7
2 Prediction error with random covariates: A single dataset . . . 8
2.1 Preliminaries . . . 9
2.2 Equally good sequences of models . . . 10
2.3 Versions of Mallows C_p for random covariates . . . 11
2.4 Discrete covariates . . . 14
2.5 Approximations and consistency . . . 14
3 Several datasets . . . 16
3.1 Model selection observing all regressions . . . 16
3.2 Consistency . . . 19
3.3 A population of distributions . . . 20
4 GENO . . . 22
4.1 Definition of GENO . . . 22
4.2 Estimation of GENO . . . 23
5 Simulations . . . 24
5.1 A single dataset . . . 25
5.2 Multiple datasets . . . 28
5.3 Comparisons to other approaches . . . 29
6 Prediction of durations of medical examinations . . . 30
6.1 Description of the data . . . 30
6.2 A regression model . . . 31
6.3 $\mathbb{C}^{(p)}$ and model selection . . . 32
6.4 Comparisons to other approaches . . . 34
References . . . 35
7 Appendix A: Proofs . . . 36
8 Appendix B: A table of notation . . . 48

1. Introduction

1.1. A general description of the problem

This paper concerns data consisting of a class of regression datasets, and a multi-task of predictions in different regressions. The emphasis is on selection of common subsets of covariates for prediction in the different regressions, which depend on the regression datasets' sample sizes. As documented in the classical model selection literature, the size or dimension of models for prediction should depend on the sample size, for example through a penalty function that depends on the dimension of the model and the sample size. References and further details will be provided following a description of our motivating problem.

In order to improve service, a hospital wants to develop a tool for predicting the actual duration of planned visits of any particular patient to any doctor in the hospital. Given a sample of size n of different patients' visits to any particular doctor (n can vary between doctors) with covariates such as the past durations of the patients' visits, the nature of the visits, the time scheduled etc., and a response variable, which is the actual duration, the goal is to predict the duration of the next visit of a given patient to the particular doctor. Our objective is to select an optimal subset of covariates, denoted by $\mathcal{P}^*(n)$, to be used for the prediction of a future visit's duration. We shall provide a procedure that selects the optimal set with high probability. The number of covariates in the set $\mathcal{P}^*(n)$ depends naturally on n, with a large n allowing more variables in the regression, taking account of the need to find the right balance between efficiency of models, and the pitfall of overfitting. For any given n, we want to select a standard set of covariates to be used for any doctor in the hospital for whom we have a regression dataset of size n. However, we allow different regression coefficients for different doctors since different doctors may be influenced differently by the patient's background. An intercept for each doctor represents her or his general tendency for longer or shorter visits.

Standardization is desirable for more than one reason, and will be discussed in more detail later. First, it obviously simplifies the data collection and maintenance. Second, performing separate model selection for each doctor may be computationally demanding. Third, model selection is notorious for being difficult and to require much data. The idea we present is to perform model selection on the basis of a sample of doctors as described below, and thus borrow strength from different datasets and obtain better subset selection.

In order to perform the subset selection we assume we have a training sample of $\mathcal{J}$ doctors each consisting of a dataset containing the covariates and the response (actual duration) of N_j visits, $j = 1, \ldots, \mathcal{J}$. Under certain assumptions, we use these data to select subsets of covariates for different value of n. We then use the selected subset for prediction for any doctor (who in general may not be in the training sample) on the basis of a sample of visits (of some size n) as described above. When predicting for a doctor in the sample, say doctor j, it is natural to take $n = N_j$. The subset selection procedure and its properties are the focus of this paper.

One may suggest to concatenate the whole training sample and perform a single regression with the same coefficients for all doctors, but allowing a different intercept for each doctor. In certain cases this may result in good subset selection. However, suppose, for example, that for about half of the doctors the covariate "duration of previous visit" has a positive coefficient in the regression and for the other half it is negative. It is easy to conceive of a justification for each possibility. In this case, this covariate may not enter the model if the regression is computed by concatenating the data into a single model. However, allowing different regression coefficients for different doctors, the variable may enter the model and contribute to the prediction with a different coefficient for different doctors. Thus, allowing the regression coefficients to vary between individual regressions adds flexibility to the model, and in particular it improves

the prediction in our dataset (verified by cross validation, see Section 6.4). Of course an informal screening of variables is often done, either at the stage of collecting the data, or before conducting formal variable selection and analysis. In particular, researchers may decide to avoid certain variables or interaction terms in order to keep the selection process feasible.

The classical theory of model selection in regression deals with the selection of a subset of covariates (or features) that are useful for prediction based on a single regression dataset. Numerous model selection methods have been suggested; AIC (Akaike [1]), Mallows C_p (Mallows [15]), and FIC (Claeskens and Hjort [8]) are prominent examples. These methods apply to a single regression dataset of a given size, for which a model is to be selected and then used for prediction. For a well-known Bayesian approach to model selection, see Schwarz [23]. A large body of literature emerged following these articles. In the setup of a single dataset, serious issues of optimality arise; see, e.g., Yang [29].

Breiman's celebrated paper [4] starts with a similar training set of regression datasets with similar assumptions, however, both the subset of covariates and the regression coefficients used for prediction are common to all regressions. A very closely related setup appears in Obozinski, Taskar and Jordan [16] which "addresses the problem of recovering a common set of covariates that are relevant simultaneously to several classification problems." The paper focuses on classification or discrimination problems, but regression is also mentioned. References cited in this paper, which deal with the same problem, are referred to as "transfer learning" or "multi-task learning" in the machine learning literature. They demonstrate that learning multiple related tasks from data simultaneously can be advantageous in terms of predictive performance relative to learning these tasks independently. In Obozinski, Taskar and Jordan [16] the goal is to decide which variables are "relevant to the overall class of prediction problems without making a commitment to a specific value of a parameter," that is, allowing different parameters for the different prediction tasks, and to "borrow strength across multiple estimation problems in order to support a decision that a covariate is to be selected." A large number of papers and review articles on multi-task learning have appeared, mostly in the past decade. For a recent survey containing numerous applications and references, see, for example, Zhang [31]. In the latter paper transfer learning refers (in our setup) to predicting for a single target file that may not be in the training sample, while multi-task learning refers to predicting for every dataset in the training sample. Here we consider both possibilities.

Our paper differs from Obozinski, Taskar and Jordan [16] and more generally from the multi-task literature in several ways: our emphasis is on regression, our focus is not on algorithms but rather on asymptotic consistency and optimality type results; however, the main difference is that in the spirit of model selection, the selected common covariate subsets, depend on the sample sizes of the different regression prediction tasks, thus avoiding under and overfitting. The issue of different sample sizes n appears in Zhang [31] in a context where face databases have different image sizes; however the proposed solution is to project these databases to a common subspace, which results in loss of information, rather

than taking the task size into account as we propose.

1.2. A formal setup

Our setup is formalized as follows. We assume we have a training sample $\mathcal{T}$ of regression datasets all having the same set of covariates. Thus $\mathcal{T} = \{D_j : j = 1, \ldots, \mathcal{J}\}$ with $D_j = \{(\mathbf{X}_{ij}, Y_{ij})\}$, $i = 1, \ldots, N_j$, $j = 1, \ldots, \mathcal{J}$, where $\mathbf{X}_{ij} \in \mathbb{R}^d$ is a column vector of d random covariate values of the ith subject in the jth dataset, and $Y_{ij} \in \mathbb{R}$ is a response variable. For each j, the N_j vectors $(\mathbf{X}_{ij}, Y_{ij})$ are iid from some distribution $G_j \in \mathcal{G}$, where $\mathcal{G}$ is a set (population) of distributions of size $|\mathcal{G}| = \mathcal{K}$. We assume that $\mathcal{J} \leq \mathcal{K} \leq \infty$, and that $\{G_j\}_{j=1}^{\mathcal{J}}$ is a random sample from $\mathcal{G}$. Now consider a new regression dataset of some size n, $D_J = \{(\mathbf{X}_{iJ}, Y_{iJ})\}_{i=1}^{n}$ distributed according to G_J, a random element of $\mathcal{G}$, which may but need not be in the training set $\mathcal{T}$. If D_J is in $\mathcal{T}$ then it is natural to assume that $n = N_J$.

We consider the following task: for $(\mathbf{X}, Y) \sim G_J$ independent of the above datasets, we want to predict Y from a given $\mathbf{X}$ using the sample D_J. It is natural to be interested in the multi-task of prediction for many random D_J's; however, it suffices to study the prediction error for one such D_J. Since G_J is random, we clearly need to consider different possible values of n, and random covariates. Our treatment of random covariates is based on a generalization of Mallows C_p to random covariates that was inspired by notes generously given to us by Larry Brown (see [5]).

As usual, the prediction model involves two components, the subset of variables to be used, and their regression coefficients. The regression coefficients will be estimated by standard least squares based on the sample D_J, and thus will vary between D_J's. For the subset selection, a task that is known to require large samples, we shall pool the whole training set. Such pooling can be efficient if the set of distributions $\mathcal{G}$, which may be finite or generated by a probability model (superpopulation model), is sufficiently homogeneous (to be discussed in Section 3) in a way that justifies a common model selection. Besides some technical conditions for such homogeneity, a user would have to apply common sense to decide if one can borrow strength and learn the subset selection from the pooled sample $\mathcal{T}$ rather than from the individual dataset D_J. As mentioned before, numerous examples appear in Zhang [31] and the references therein. Our goal is to select for each possible value of n, a subset of covariates based on the pooled training sample $\mathcal{T}$ and use it for prediction, using least squares estimates, computed for each regression dataset D_J separately. Thus, we select subsets for prediction that are common to all regressions having the same sample size, but we allow different parameters for the regressions.

Given a distribution $G_j \in \mathcal{G}$, let $m_j(\mathbf{X}_{ij}) := E_{G_j}(Y_{ij} \mid \mathbf{X}_{ij})$ be the conditional expectation under G_j. We do not assume a linear model or any particular model for m_j when we analyze our procedures, but for the sake of prediction we shall approximate $m_j(\mathbf{X}_{ij})$ by a linear function $\mathbf{X}'_{ij}\boldsymbol{\beta}_j$, where $\boldsymbol{\beta}_j$ is the vector of projection coefficients under G_j. We shall require minimal assumptions on G_j such as moment conditions, to be specified later.

When $m_j(\mathbf{X}_{ij})$ is not linear then $\mathbf{X}_{ij}$ is not ancillary, and its marginal distribution matters; see, e.g, Buja et al. [6]. In this case, conditioning on $\mathbf{X}$ or considering it as nonrandom leads to loss of information. For a recent discussion on fixed versus random $\mathbf{X}$ in the context of model selection see Rosset and Tibshirani [20]. When $m_j(\mathbf{X}_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta}_j$, allowing linear models with different coefficients in different regressions is called *heterogeneous regression ANCOVA*; see, e.g., Rutherford [21], Chapter 8, and the references therein. Related models appear under titles such as repeated measure regression (see, e.g., Vonesh and Chinchilli [26]), often with mixed effects.

Given datasets $\{\mathbf{X}_{ij}, Y_{ij}\}$ from G_j, consider the subset of covariates $\mathcal{P}$ of size $p \leq d$. We may sometimes refer to $\mathcal{P}$ as a *model*. Let $\mathbf{X}_{ij}^{(\mathcal{P})}$ denote the subvector of $\mathbf{X}_{ij}$ consisting of the covariates in $\mathcal{P}$. Let $\boldsymbol{\beta}_j^{(\mathcal{P})}$ denote the linear projection coefficient vector and let $\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$ be its least squares estimator based on n observations, where we assume that $n > p$. In Section 2.4 we discuss the case of discrete covariates in which exact (or perfect) multicollinearity may occur with a positive probability, and the least squares estimators are not unique.

For now we focus on the case that $\mathcal{J} = \mathcal{K}$, that is we observe all datasets in $\mathcal{G}$. (In Section 3.3 we consider prediction of an out of (the training) sample dataset, in the spirit of transfer learning.) Consider prediction for a regression dataset of some size n, often referred to as the *task size*, which will be taken to equal N_j for the task of predicting for the dataset D_j in the multi-task problem of prediction for datasets in the training sample. The linear prediction of a response Y, based on n observations from a random $G_J \in \mathcal{G}$, and when the subset $\mathcal{P}$ is used is given by $(\mathbf{X}^{(\mathcal{P})})' \widehat{\boldsymbol{\beta}}_{J,n}^{(\mathcal{P})}$. In order to select a subset for regression tasks of size n we make the counterfactual assumption that all datasets in the training sample are of size n. Then the corresponding expected prediction error or risk is given by

$$\mathbf{R}(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} R_j(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} E_{G_j}\big(Y - (\mathbf{X}^{(\mathcal{P})})' \widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}\big)^2, \tag{1.1}$$

where $(\mathbf{X}, Y) \sim G_j$ independently of $\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$, and the expectation on the right-hand side of (1.1) applies to both $(\mathbf{X}, Y)$ and $\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$. This expression has the alternative interpretation where instead of predicting for a random D_J we predict for all D_j, $j = 1, \ldots, \mathcal{J}$, assuming that there is a common sample size n, and now $\mathbf{R}(n, \mathcal{P})$ represents the average prediction error. With either interpretation, our goal is to estimate $\mathbf{R}(n, \mathcal{P})$ and related quantities, in order to select (with high probability) an optimal common subsets $\mathcal{P}^*(n)$ for prediction for any dataset in the training set (taking $n = N_j$) and also for out of sample datasets when we later consider the case that $\mathcal{J} < \mathcal{K}$, on the basis of n observations, where the subset selection is based on the pooled training set $\mathcal{T}$.

It is natural to choose common subsets for prediction if the different regression datasets arise from a common source; besides efficiency in subset selection due to pooling, common subsets lead to computational efficiency. However, we

assert that in a variety of situations (but obviously not always) it is advantageous to choose standard common sets of covariates to be used for prediction even if the regression datasets do not arise from a homogeneous source. In this case we are trying to select compromise subsets that can be used for the different regressions (and may not be optimal for some or any of them). For example, a large health organization with $\mathcal{K}$ clinics often recommends a common standard set of tests for the purpose of certain diagnoses, thereby simplifying the instructions to participating clinics and doctors. In our notation, the set of tests is based on a sample of size $\mathcal{J}$, which is in general $\leq \mathcal{K}$. The regression coefficients used for prediction based on this common set of tests may differ between communities or doctors, who may attach different weights to different tests. Concerning economics models, consider the OECD, where $\mathcal{J} = \mathcal{K} = 37$ (as of 2021) since all countries are sampled and economic prediction are made in all of them. The OECD attempts to standardize sets of common economic indicators to be used for economic predictions (e.g., forecast of GDP growth) for its member countries, which are to be estimated by their bureaus of statistics by the same methodology. In general, it makes sense to assume that in different countries, economic variables may have different weights in economic predictions. For example, oil prices must weigh differently for economic predictions between oil importing and exporting countries.

1.3. GENO, a measure of usefulness

In order to compare the quality of different models, we introduce a new measure, GENO, which is inspired by the measure ENO (equivalent number of observations) of Erev, Roth, Slonim, and Barron [10]. To describe ENO in the context of experimental economics, consider an experiment where a game is played by a sample of subjects in order to study the average behavior of players, and predict future play. ENO is based on a comparison between the empirical statistics of past actions of the players, and a given model for predicting players' actions. The more subjects who have already played the game, the better the estimate that past play will give of the mean behavior of the subject population on this game. ENO measures the usefulness of the prediction of a particular model by asking how many prior observations of subjects playing the game, say m, would be needed to make the empirical statistics as accurate as the prediction by the model. ENO of a model is this number m.

While ENO compares a given model to the relevant empirical model, GENO generalizes ENO to comparing any two data-based models. Thus, let now $\mathbf{R}(n, \mathcal{P})$ denote the prediction error of some model $\mathcal{P}$ in a very general setup. For our present purposes, one can have regression models in mind, with $\mathbf{R}(n, \mathcal{P})$ defined above; however, the definition of GENO below is more general. Given two models $\mathcal{P}$ and $\mathcal{Q}$, define $\mathrm{GENO}(n; \mathcal{P}, \mathcal{Q})$ to be the value of m satisfying $\mathbf{R}(m, \mathcal{Q}) = \mathbf{R}(n, \mathcal{P})$. In words, $\mathrm{GENO}(n; \mathcal{P}, \mathcal{Q})$ is the number of observations required in order for a model based on the covariates in $\mathcal{Q}$ to predict equally well as a model based on the covariates in $\mathcal{P}$, when the parameters of the latter

model are estimated on the basis of n observations. In Section 4 we shall use an approximation to $\mathbf{R}(n, \mathcal{P})$ to formally define and estimate GENO. Such a measure allows us to decide between a set of covariates that may be good for prediction but costly to obtain, and another set of more accessible covariates that we may consider using, even if their predictive value is lower and therefore may require more observations. See the recent paper - Andrade et al. [2] and the references therein for a formal Bayesian approach to minimizing cost of classification in the presence of costly covariates. A comparison in terms of the sample size required by one model (for prediction, testing or estimation) to be as good as another with a given sample size is closely akin to the notion of Pitman efficiency; see, e.g., Zacks [30]. Our approach to quantifying the value of a model is close in spirit, but not in detail, to the work of Lindsay and Liu [14] who define a "model credibility index" as the sample size N^*, where data from the model and from the true generating process are indistinguishable in the sense that for a given goodness of fit test of the model with N^* observations, the probability of rejection under the model is, say, 50%.

A different approach to measuring the usefulness of a model is by Akaike weights, which are defined by the likelihood function of each model evaluated at the MLE, standardized by their sum; see Anderson and Burnham [3] (Page 75), where these weights are referred to informally as "the weight of evidence in favor of model." The AIC weights are sometimes [e.g., 27] interpreted as probabilities of a model to be the best in terms of the AIC criterion. With a uniform prior on the set of models this interpretation could be meaningful if we believe that one of the models is true. Otherwise, the weights are still informative, but their interpretation is less clear. GENO, on the other hand, is measured in units of number of observations, which are easy to grasp. Another advantage of GENO is that it accounts for the number of observations in the data to which the model is applied. This makes sense as the usefulness of a model for a given dataset is also a function of the size of the data.

In Section 2 we restate the problem and provide some basic results and notation for a single regression dataset, as a preliminary to the main part, Section 3, where we consider the multi-task problem of model selection for several regression datasets. In Section 4 we discuss the GENO measure of the relative quality of models. In Section 5 we demonstrate the results by simulations, and in Section 6 we discuss an application to a medical management problem of predicting service times, that is, visit durations of patients in hospital. Section 7 is an appendix containing the proofs. Appendix B summarizes the notation used in the paper.

2. Prediction error with random covariates: A single dataset

We start with $|\mathcal{G}| = \mathcal{J} = 1$, that is, with selection of a model for prediction given a training set consisting of a single regression dataset. This case is treated in the standard model selection literature. Although our real interest is in results for large $\mathcal{J}$, we consider $\mathcal{J} = 1$ as a starting point which simplifies the notation while

allowing us to present some of the ideas used in the general case. For now our training set $\mathcal{T}$ consists of a single dataset $D_1 = \{(\mathbf{X}_i, Y_i)\}_1^N$ of $N := N_1$ iid pairs from some distribution $G := G_1$. The distinction between N and n may seem artificial in this case, but we shall make it and consider prediction based on any sample size n for later purposes. We use D_1 for selecting a subset of covariates for linear prediction of a future Y from $\mathbf{X}$ distributed by G, with parameters that will be estimated using a dataset $D = \{(\mathbf{X}_i, Y_i)\}_1^n$ of n observations from G.

We derive some results that will be needed for the general case $\mathcal{J} > 1$, to be discussed in Section 3. Subsets of the covariates are denoted by letters like $\mathcal{P}$, $\mathcal{Q}$, etc., and their sizes by p and q, etc. We refer to the associated linear model as model $\mathcal{P}$. For now we fix $\mathcal{P}$ and suppress it in most of our notation and instead of $\mathbf{X}^{(\mathcal{P})}$ we write $\mathbf{X}$ and assume it is in $\mathbb{R}^p$. The same holds for other vectors and matrices. Later we shall assume that $\mathbf{X} \in \mathbb{R}^d$, and consider different subsets of covariates.

2.1. Preliminaries

Consider a dataset $D = \{(\mathbf{X}_i, Y_i)\}_1^n$ of iid pairs from some distribution G, where $\mathbf{X}_i$ is a column vector in $\mathbb{R}^p$, $i = 1, \ldots, n$. Let $(\mathbf{X}, Y)$ without indexes denote one such "generic" observation, distributed independently of the dataset D as any $(\mathbf{X}_i, Y_i)$ according to G. The first entry of each $\mathbf{X}_i$ may be 1, so that the models may include an intercept term.

Set $\mathbb{Q} := E(\mathbf{X}\mathbf{X}')$ and let $\mathbf{Y}_n \in \mathbb{R}^n$ denote the n-column vector of the Y_i's, and set $m(\mathbf{X}) := E(Y|\mathbf{X})$ for some function m. Assuming that both $\mathbf{X}$ and Y have finite second moments and that $\mathbb{Q}$ is invertible, the best linear approximation of $m(\mathbf{X})$ is $\mathbf{X}'\boldsymbol{\beta}$, where

$$\boldsymbol{\beta} := \arg\min_{\mathbf{b} \in \mathbb{R}^p} E(m(\mathbf{X}) - \mathbf{X}'\mathbf{b})^2 = \mathbb{Q}^{-1} E(\mathbf{X}Y). \tag{2.1}$$

The same projection coefficient vector $\boldsymbol{\beta}$ also satisfies $\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E(Y - \mathbf{X}'\mathbf{b})^2$; hence $\mathbf{X}'\boldsymbol{\beta}$ is the best linear predictor of Y. Our assumptions imply that the minimizer $\boldsymbol{\beta}$ is unique. Set $e_i := Y_i - \mathbf{X}_i'\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ defined in (2.1). By (2.25) in Hansen [12], where most of our notation and the standard results we use can be found, we have $E(\mathbf{X}e) = \mathbf{0}$, where again $\mathbf{X}$ and e are "generic" $\mathbf{X}_i$ and e_i.

Define $\mathbb{X}_n$ to be the $n \times p$ matrix whose n rows are the row vectors $\mathbf{X}_i'$. In this common notation the standard linear model will be written as $\mathbb{X}_n\boldsymbol{\beta}$, whereas each of its rows as $\mathbf{X}_i'\boldsymbol{\beta}$, and $\mathbb{X}_n'\mathbb{X}_n = \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$. Under standard assumptions, the least squares estimator is

$$\widehat{\boldsymbol{\beta}}_n := \arg\min_{\mathbf{b} \in \mathbb{R}^p} ||\mathbf{Y}_n - \mathbb{X}_n\mathbf{b}||^2 = (\mathbb{X}_n'\mathbb{X}_n)^{-1}\mathbb{X}_n'\mathbf{Y}_n. \tag{2.2}$$

The assumption that $(\mathbb{X}_n'\mathbb{X}_n)^{-1}$ exists (with probability 1) holds if we assume that $\mathbf{X}$ has a continuous distribution. For the existence of certain moments required later we shall assume that the distribution of $\mathbf{X}$ is a mixture of normals.

See Hansen [12], pp. 102–3, for a discussion of the existence of $(\mathbb{X}_n'\mathbb{X}_n)^{-1}$ and its moments. Without assuming continuity, the assumption that $\mathbb{Q}$ is invertible implies that $(\mathbb{X}_n'\mathbb{X}_n)^{-1}$ exists with probability converging to 1 as $n \to \infty$; however, for discrete distributions this probability is smaller than one, and thus $\widehat{\boldsymbol{\beta}}_n$ may not exist, and has no finite moments, a "conundrum" in the words of Hansen. In Section 2.4 we extend our discussion to discrete covariates by conditioning on the existence of a bounded inverse, and showing that under simple conditions this amounts to neglecting a set having an exponentially small probability, thus providing some solution to the above conundrum.

We now assume that $(\mathbb{X}_n'\mathbb{X}_n)^{-1}$ exists and has sufficiently many moments so that expressions like (2.4) below are finite. If $\mathbf{X}$ and Y have finite fourth moments, then by Theorem 7.3 in Hansen [12]

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbb{Q}^{-1}\mathbb{W}\mathbb{Q}^{-1}), \tag{2.3}$$

where $\mathbb{W} := E(\mathbf{X}\mathbf{X}'e^2)$, a $p \times p$ matrix assumed to be positive definite. For a single distribution G and a dataset D as above, the prediction error incurred by a model $\mathcal{P}$ based on all p covariates with linear regression coefficients computed from a sample of size n is

$$R(n, \mathcal{P}) = E_G\big(Y - \mathbf{X}'\,\widehat{\boldsymbol{\beta}}_n\big)^2. \tag{2.4}$$

Later we assume that $\mathbf{X} \in \mathbb{R}^d$ and set $\mathbf{X}^{(\mathcal{P})} \in \mathbb{R}^p$ to be the vector consisting of the covariates of $\mathbf{X}$ in the subset of covariates $\mathcal{P}$ of size p. When we consider several models, we set, for example, $\mathbb{X}_n^{(\mathcal{P})}$ to be the $n \times p$ matrix whose n rows are the row vectors $\mathbf{X}_i^{'(\mathcal{P})}$, $\widehat{\boldsymbol{\beta}}_n^{(\mathcal{P})} := (\mathbb{X}_n^{'(\mathcal{P})}\mathbb{X}_n^{(\mathcal{P})})^{-1}\mathbb{X}_n^{'(\mathcal{P})}\mathbf{Y}_n$, $\mathbb{W}^{(\mathcal{P})} := E(\mathbf{X}^{(\mathcal{P})}\mathbf{X}^{(\mathcal{P})'}e^2)$, and likewise for $\mathbb{Q}$, etc. We then have

$$R(n, \mathcal{P}) = E_G\big(Y - \mathbf{X}^{'(\mathcal{P})}\,\widehat{\boldsymbol{\beta}}_n^{(\mathcal{P})}\big)^2. \tag{2.5}$$

2.2. Equally good sequences of models

When selecting the best model for a given n, that is, the subset of covariates that minimizes $R(n, \mathcal{P})$, we should take into account that different samples yield different estimators $\widehat{\boldsymbol{\beta}}_n$, leading to different prediction errors; thus, there is no gain in optimizing more precisely than the difference between such errors. Consider the prediction error conditioned on the estimated regression coefficients

$$R\big(n, \mathcal{P}; \widehat{\boldsymbol{\beta}}_n\big) := E\Big[\big(Y - \mathbf{X}'\,\widehat{\boldsymbol{\beta}}_n\big)^2 \Big| \widehat{\boldsymbol{\beta}}_n\Big].$$

Note that $R(n, \mathcal{P}) = E\{R(n, \mathcal{P}; \widehat{\boldsymbol{\beta}}_n)\}$. By using the relation $\big(Y - \mathbf{X}'\,\widehat{\boldsymbol{\beta}}_n\big)^2 = \big(Y - \mathbf{X}'\,\boldsymbol{\beta} + \mathbf{X}'\,\boldsymbol{\beta} - \mathbf{X}'\,\widehat{\boldsymbol{\beta}}_n\big)^2$, expanding the latter term, and taking conditional expectation noting that $E[(Y - \mathbf{X}'\boldsymbol{\beta})\mathbf{X}'|\widehat{\boldsymbol{\beta}}_n] = E[e\mathbf{X}'] = 0$, we obtain

$$R\big(n, \mathcal{P}; \widehat{\boldsymbol{\beta}}_n\big) = E\,(Y - \mathbf{X}'\,\boldsymbol{\beta})^2 + E\Big\{\Big[\mathbf{X}'\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})\Big]^2 \Big| \widehat{\boldsymbol{\beta}}_n\Big\}. \tag{2.6}$$

The first term in (2.6) is a constant and the second equals $(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'\mathbb{Q}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$, which is of order $O_p(1/n)$ since $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = O_p(1)$; see (2.3). This means that $R\big(n, \mathcal{P}; \widehat{\boldsymbol{\beta}}_n\big)$ varies between different $\widehat{\boldsymbol{\beta}}_n$ by a quantity of order $O_p(1/n)$. Hence, if two sequences of models $\mathcal{P}(n)$ and $\mathcal{Q}(n)$ satisfy

$$|R(n, \mathcal{P}(n)) - R(n, \mathcal{Q}(n))| = o(1/n), \text{ i.e, } \lim_{n\to\infty} n|R(n, \mathcal{P}(n)) - R(n, \mathcal{Q}(n))| = 0,$$

we consider them to be *equally good*. If $\mathcal{P}(n)$ is best in the sense of minimizing $R(n, \mathcal{P}(n))$ and $\mathcal{Q}(n)$ is equally good, we say that $\mathcal{Q}(n)$ is *adequate*, and rather than choose "best models" we settle for adequate models. See, e.g., Nevo and Ritov [17] for a related approach.

2.3. Versions of Mallows C_p for random covariates

Given a dataset $D_1 = \{(\mathbf{X}_i, Y_i)\}$ of size N (which constitutes the training set when $\mathcal{J} = 1$), we first estimate the prediction error (2.4) incurred if prediction is to be based on n observations. We shall consider two types of asymptotics: one when n is considered to be large, and the other when n is fixed, and N is large. For now $\mathcal{J} = 1$; asymptotics in $\mathcal{J}$ will be considered later.

We use the following notation: set $\widehat{\mathbb{Q}}_N := \frac{1}{N}\mathbb{X}_N'\mathbb{X}_N$, and let $\mathbf{Y}_N$ denote the N-vector of the Y_i's. Recalling the notation $e_i = Y_i - \mathbf{X}_i'\boldsymbol{\beta}$, let $\mathbf{e}_N$ denote the N-vector having components e_i. Set $\widehat{\mathbb{W}}_N := \frac{1}{N}\sum_{i=1}^N \mathbf{X}_i\mathbf{X}_i'\widehat{e}_i^2$ with $\widehat{e}_i = Y_i - \mathbf{X}_i'\widehat{\boldsymbol{\beta}}_N$, where $\widehat{\boldsymbol{\beta}}_N$ is given by (2.2) upon replacing n by N. Thus in (2.8) below, $\frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 = \frac{1}{N}\sum_{i=1}^N \widehat{e}_i^2$. Let $\mathbb{V} := \mathbb{W}\mathbb{Q}^{-1}$, and $\widehat{\mathbb{V}}_N := \widehat{\mathbb{W}}_N\widehat{\mathbb{Q}}_N^{-1}$. In addition we define $\mathbf{U}_N := \frac{1}{\sqrt{N}}\mathbb{X}_N'\mathbf{e}_N = \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{X}_i e_i$. Note that $\mathbf{U}_N$ is not a statistic and that $E(\mathbf{U}_N\mathbf{U}_N') = \mathbb{W}$ since $\mathrm{E}(\mathbf{X}e) = \mathbf{0}$ implies that the expectations of mixed terms vanish. In all the vectors and matrices above and below the index $\mathcal{P}$ was suppressed unless otherwise indicated.

Akin to (2.4), we define the approximate prediction error to be

$$AR(n, \mathcal{P}) := E\big(Y - \mathbf{X}'\boldsymbol{\beta}\big)^2 + \frac{1}{n}tr(\mathbb{V}), \tag{2.7}$$

where tr denotes trace, and equation (2.9) of Theorem 2.1 below shows that it is an approximation to the quantity $R(n, \mathcal{P})$ of (2.4). Clearly $E\big(Y - \mathbf{X}'\widehat{\boldsymbol{\beta}}_n\big)^2 \geq E\big(Y - \mathbf{X}'\boldsymbol{\beta}\big)^2$ and the trace is an approximation of the difference with precision of order $O(1/n^{3/2})$; see (2.9). Next we define the statistic $C^{(\mathcal{P})}(n, N)$ as an estimator of $AR(n, \mathcal{P})$ by

$$C^{(\mathcal{P})}(n, N) := \frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 + tr(\widehat{\mathbb{V}}_N)\left(\frac{1}{n} + \frac{1}{N}\right). \tag{2.8}$$

The new term $\frac{1}{N}tr(\widehat{\mathbb{V}}_N)$ is an approximately (up to $o_p(1/N)$) unbiased estimator of $\frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 - E\big(Y - \mathbf{X}'\boldsymbol{\beta}\big)^2$, as shown in (7.6) and (7.7). The fact that $tr(\widehat{\mathbb{V}}_N)$ is a biased estimator of $tr(\mathbb{V})$ entails a bias of order $1/n$ for the

estimator $C^{(\mathcal{P})}(n,N)$ as an estimator of $AR(n,\mathcal{P})$. We shall study the latter estimator, and when we use it, we shall apply a standard jackknife correction for its bias; see Efron [9], Equation (2.8). We denote the bias-corrected $C^{(\mathcal{P})}(n,N)$ by $\mathbb{C}^{(\mathcal{P})}(n,N)$. It suffices to bias-correct only $tr(\widehat{\mathbb{V}}_N)$ in (2.8) as explained in the fourth paragraph after Theorem 2.1.

The superscript $\mathcal{P}$ in the statistic $C^{(\mathcal{P})}$ refers to the set of covariates in $\mathcal{P}$ and for now we have $\mathbf{X}_i^{(\mathcal{P})} = \mathbf{X}_i \in \mathbb{R}^p$ and the subset $\mathcal{P}$ is fixed and suppressed. The statistic $C^{(\mathcal{P})}(n,N)$ is a counterpart of Mallows C_p, but here we consider random covariates. Furthermore, we distinguish between the number N of observations used for the choice of the model and the sample size n of observations used for estimating the model's parameters. The classic Mallows C_p concerns nonrandom covariates, where $n = N$, and the true model is assumed to be linear. To see the relation to Mallows C_p, assuming a homoskedastic linear model, we have that $e_i = Y_i - \mathbf{X}_i'\boldsymbol{\beta}$ is uncorrelated with the covariates, with variance σ^2, and $\mathbb{W} = \sigma^2\mathbb{Q}$, and therefore $\widehat{\mathbb{V}}_N = \widehat{\mathbb{W}}_N\big(\widehat{\mathbb{Q}}_N\big)^{-1}$ will converge to $\sigma^2 I_p$ and $tr(\widehat{\mathbb{V}}_N)$ to $\sigma^2 p$. If we use $\sigma^2 p$ as an approximation of $tr(\widehat{\mathbb{V}}_N)$ (and therefore we only have to estimate σ^2 rather than a trace), then $C^{(\mathcal{P})}$ in the case $N = n$ coincides with Mallows C_p.

The following theorem provides the rate of approximation of $AR(n,\mathcal{P})$ to $R(n,\mathcal{P})$, and then analyzes $C^{(\mathcal{P})}$ as an estimator of $AR(n,\mathcal{P})$; some of its conditions and implications are discussed below. All proofs are in the Appendix. Our proof shows that Assumption (i) below can be replaced by the assumption that $\mathbf{X}$ and Y have 24 finite moments, and a careful inspection of the proof shows that this number can be somewhat reduced.

Theorem 2.1. *Assume that*
(i) *The coordinates of* $\mathbf{X}$ *and* Y *have finite moments of all orders.*
(ii) *The entries of* $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ *have third moments that are bounded uniformly in* n*. Then*

$$|R(n,\mathcal{P}) - AR(n,\mathcal{P})| = O(1/n^{3/2}), \tag{2.9}$$

and

$$AR(n,\mathcal{P}) - C^{(\mathcal{P})}(n,N) = \mathcal{E}_N + \frac{1}{n}\Big\{tr(\mathbb{V}) - tr(\widehat{\mathbb{V}}_N)\Big\} + o_p(1/N), \tag{2.10}$$

where

$$\mathcal{E}_N = E(Y - \mathbf{X}'\boldsymbol{\beta})^2 - \frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\boldsymbol{\beta}||^2 + \frac{1}{N}\{tr(\mathbf{U}_N\mathbf{U}_N'\mathbb{Q}^{-1}) - tr(\mathbb{V})\}. \tag{2.11}$$

Furthermore,

$$(a)\quad \mathcal{E}_N = O_p(1/\sqrt{N}), \qquad (b)\quad tr(\mathbb{V}) - tr(\widehat{\mathbb{V}}_N) = O_p(1/\sqrt{N}), \tag{2.12}$$

and

$$\sqrt{N}\big(C^{(\mathcal{P})}(n,N) - AR(\mathcal{P},p)\big) \xrightarrow{\mathcal{D}} N(0,\tau^2) \tag{2.13}$$

for some asymptotic variance τ^2 *as* $N \to \infty$*, and* n *is fixed.*

Since there is only a finite number of models, the above terms O, O_p, and o_p do not depend on the subset of covariates $\mathcal{P}$. For example, we could replace (2.9) by $|R(n,\mathcal{P}) - AR(n,\mathcal{P})| \leq B/n^{3/2}$ for all n and $\mathcal{P}$, where B is a constant. Moreover, the term $o_p(1/N)$ in (2.10) does not depend on n.

Condition (i) of Theorem of 2.1 is standard, and Lemma 2.2 below shows that Condition (ii) is satisfied if $\mathbf{X}$ is distributed as a mixture of normals; see Sampson [22]. Such mixtures form a dense family of distributions with respect to weak convergence in the space of distribution on $\mathbb{R}^p$. As the distribution of $\mathbf{X}$ is never known exactly, it makes sense to assume, as an approximation, that the data satisfy such a condition. The case where $\mathbf{X}$ has discrete components is discussed in Section 2.4.

We shall later compare models consisting of different subsets of covariates. Equation (2.9) suggests that choosing a model by minimizing a good estimate of $AR(n,\mathcal{P})$ with respect to $\mathcal{P}$ can lead to a model for which $R(n,\mathcal{P})$ is within $o(1/n)$ of the best model, and thus $\mathcal{P}$ is an adequate model in the sense of Section 2.2. This is stated formally in Proposition 2.4.

In view of (2.10) we use $C^{(\mathcal{P})}(n,N)$ of (2.8) as an estimator of the approximate prediction error $AR(n,\mathcal{P})$ and hence of the prediction error $R(n,\mathcal{P})$. This is formalized in Propositions 2.5 and 2.6 below. We now briefly discuss Equations (2.10) and (2.11). First consider the bias of $C^{(\mathcal{P})}(n,N)$ as an estimator of $AR(n,\mathcal{P})$. It is easy to see that $E\mathcal{E}_N = 0$. By (2.12) (b), $tr(\mathbb{V}) - tr(\widehat{\mathbb{V}}_N) = O_p(1/\sqrt{N})$, and after dividing the latter term by n as in (2.10), it is of a smaller order than the term $tr(\widehat{\mathbb{V}}_N)\left(\frac{1}{n}+\frac{1}{N}\right)$ appearing in $C^{(\mathcal{P})}(n,N)$. This shows that the latter term contributes to reducing the bias of $C^{(\mathcal{P})}(n,N)$ as an estimator of $AR(n,\mathcal{P})$.

Our main interest is in the case of $\mathcal{J} > 1$ regressions, and in choosing a model that minimizes an average of $\mathcal{J}$ values of AR. Averaging (nearly) unbiased estimates can result in consistency in $\mathcal{J}$, which explains why we care about correcting the bias of $C^{(\mathcal{P})}(n,N)$. In this case, a further bias correction using the jackknife is useful (see Section 5.2). The above discussion implies that it suffices to bias-correct the estimator $tr(\widehat{\mathbb{V}}_N)$, which is what we do when using the jackknife.

Choosing a good model can be reduced to choosing between two models, say, $\mathcal{P}$ and $\mathcal{Q}$ at a time, by approximating the difference $AR(\mathcal{P}) - AR(\mathcal{Q})$ using $C^{(\mathcal{P})}(n,N) - C^{(\mathcal{Q})}(n,N)$. The leading terms in the latter expression will be the difference between the relevant values of $\mathcal{E}_N$ for the two models, and it is easy to see that the leading term of this difference is the difference between the values of $\frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\boldsymbol{\beta}||^2$ for the corresponding models, which is of order $O_p(1/\sqrt{N})$ by the central limit theorem. However, when two models having very similar prediction values are compared by differencing their corresponding values of $C^{(\mathcal{P})}(n,N)$, their leading terms will approximately cancel, and in this case the second term on the right-hand side of (2.8) plays a role. This holds also for Mallows C_p and the AIC, [1], and will be exploited formally in the Propositions 2.5 and 2.6 below.

The following lemma shows that Condition (ii) of Theorem 2.1 holds when

$\mathbf{X}$ is distributed as a mixture of normals.

Lemma 2.2. *Let the distribution of the covariate vectors (excluding the first coordinate in the case that it is a constant 1) be normal, or a finite mixture of normals, or an infinite mixture of normals with covariance matrices in a set Ξ, and $\inf_{\Sigma\in\Xi}\lambda_{min}(\Sigma)>0$, where λ_{min} denotes the smallest eigenvalue. Then, for $n>p+5$, Condition* (ii) *of Theorem 2.1 is satisfied.*

More generally, the rth moments of the entries of $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ are bounded under the conditions of Lemma 2.2 provided that $n>p+2r-1$ (see von Rosen [19], Theorem 4.1). Note that the condition on λ_{min} guarantees that $\mathbf{X}$ is bounded away from exact multicollinearity.

2.4. Discrete covariates

When $\mathbf{X}$ contains discrete covariates, the probability that the matrix $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ does not exist is positive, and expressions like $\widehat{\boldsymbol{\beta}}_n$ of (2.2) and hence $R(n,p)$ of (2.4) may not exist. When the components of $\mathbf{X}$ are bounded, we provide the following limiting approach. Set

$$H_n := \big\{\mathbb{X}_n \,:\, \lambda_{min}(\mathbb{X}_n'\mathbb{X}_n/n)\geq\lambda_{min}(\mathbb{Q})/2\big\}, \tag{2.14}$$

where λ_{min} is the smallest eigenvalue, and $\widetilde{R}(n,\mathcal{P}) := E\big[\big(Y-\mathbf{X}'\widehat{\boldsymbol{\beta}}_n\big)^2 \mid H_n\big]$. We have

Theorem 2.3. *Suppose that Y has all moments, the components of $\mathbf{X}$ are bounded, and $\mathbb{Q}$ is invertible; then for some $a\in(0,1)$,*

$$\big|\widetilde{R}(n,\mathcal{P})-AR(n,\mathcal{P})\big| = O(1/n^{3/2}) \text{ and } P(H_n)>1-a^{n\lambda_{min}(\mathbb{Q})}.$$

Moreover, all quantities appearing in Theorem 2.1 are well defined on H_N, and can be defined in an arbitrary way outside of H_N, and the results (2.10)–(2.13) *hold.*

Thus, apart from the complement H_n^c, which has exponentially small probability, the approximation rate of $AR(n,\mathcal{P})$ to the prediction error is the same as in (2.9) and the rest of Theorem 2.1 still holds. The result follows from Theorem 2.1 and Lemma 2.3 given in the Appendix.

2.5. Approximations and consistency

The focus of this section is on choosing a subset of covariates for prediction of future responses on the basis of a single dataset of size N. The linear model parameters are estimated from a sample of size n, with the understanding that different n's may (and should) lead to different choices of subsets; more specifically, a larger n naturally gives rise to a larger set of covariates. Asymptotic

results in n are not of major interest in this context; however, they may contribute some understanding when n is not small. Such results are discussed in this section.

In Proposition 2.5 we show that under the conditions of Theorem 2.1, choosing a subset of covariates in the set $\arg\min_{\mathcal{P}} C^{(\mathcal{P})}(n,N)$ guarantees that for increasing n and N we choose the best linear model with probability converging to 1, that is, the model minimizing $R(n,\mathcal{P}) = E\big(Y - \mathbf{X}^{(\mathcal{P})'}\widehat{\boldsymbol{\beta}}_n^{(\mathcal{P})}\big)^2$, with notation defined after (2.4). In Proposition 2.6 we show that for fixed n, using $C^{(\mathcal{P})}(n,N)$, we choose an adequate model in the sense defined in Section 2.2, with probability converging to 1 as $N \to \infty$.

Below $\arg\min_{\mathcal{P}}$ is taken over all subsets of covariates. For a given n, define the following sets:

$$\mathcal{P}^*(n) := \arg\min_{\mathcal{P}} R(n,\mathcal{P}) = \arg\min_{\mathcal{P}} E\big(Y - \mathbf{X}^{(\mathcal{P})'}\widehat{\boldsymbol{\beta}}_n^{(\mathcal{P})}\big)^2,$$

$$\pi^*(n) := \arg\min_{\mathcal{P}} AR(n,\mathcal{P}) = \arg\min_{\mathcal{P}} \Big\{E\big(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2 + \frac{1}{n}tr(\mathbb{V}^{(\mathcal{P})})\Big\},$$

$\mathcal{P}^* := \arg\min_{\mathcal{P}\in\mathcal{M}} |\mathcal{P}|$, where $\mathcal{M} := \arg\min_{\mathcal{P}} E\big(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2$ and $|\mathcal{P}|$ denotes the number of covariates in the model $\mathcal{P}$,

$$\widehat{\pi^*}(n,N) := \arg\min_{\mathcal{P}} C^{(\mathcal{P})}(n,N).$$

The following proposition shows that the first two sets defined above by arg min converge to the third, which is a singleton. Note that $\mathcal{P}^*$ is the best linear model in the sense of being the most parsimonious model minimizing the expected square of the projection error $Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}$. We deal with the convergence of $\widehat{\pi^*}(n,N)$ in Proposition 2.5.

Proposition 2.4. *Suppose that the conditions of Theorem 2.1 hold. Then*
(i) *Any two sequences in $\pi^*(n)$ and $\mathcal{P}^*(n)$ are equally good, that is, any sequence of models in $\pi^*(n)$ is adequate in the sense of Section 2.2.*
(ii) *The set $\mathcal{P}^*$ is a singleton, and the sets $\pi^*(n)$ and $\mathcal{P}^*(n)$ converge to the singleton $\mathcal{P}^*$ as $n \to \infty$.*

The proof shows that essentially $\mathcal{M}$ is a singleton; that is, besides $\mathcal{P}^*$, $\mathcal{M}$ may only contain models having the same covariates and regression coefficients as those of $\mathcal{P}^*$, and further covariates whose coefficients vanish. Note that since the number of models is finite, it follows that $\mathcal{P}^*(n) = \pi^*(n) = \mathcal{P}^*$ for large enough n; that is, the same model $\mathcal{P}^*$ minimizes both $R(n,\mathcal{P})$ and $AR(n,\mathcal{P})$. The model $\mathcal{P}^*$ is the minimal best linear predictive model that one would ideally use if the projection coefficients $\boldsymbol{\beta}^{(\mathcal{P})}$ were known.

The next proposition shows that minimizing the statistic $C^{(\mathcal{P})}(n,N)$ leads to correct selection asymptotically, that is, to selecting the model that minimizes the prediction error $R(n,\mathcal{P})$ with probability converging to 1.

Proposition 2.5. *Under the conditions of Theorem 2.1, with both $n, N \to \infty$, and $n/N \to 0$, we have $P\big(\widehat{\pi^*}(n,N) = \mathcal{P}^*(n)\big) \to 1$.*

The proof is given in the Appendix, where we also show that the condition $n/N \to 0$ is necessary. The case $n = N$ (with nonrandom covariates) corresponds to the standard Mallows C_p, which is inconsistent; more specifically, it is well known that for $n = N$, the choice $\widehat{\pi^*}(n, N)$ may lead to models $\mathcal{Q}$ that strictly contain $\mathcal{P}^*$; see, e.g., Nishii [18]. The equality $\widehat{\pi^*}(n, N) = \mathcal{P}^*(n)$, which holds for large enough n and N with high probability, implies that $\widehat{\pi^*}(n, N)$ is a singleton (by Proposition 2.4 (ii)), and that selecting a model according to the statistic $\widehat{\pi^*}(n, N)$ yields a model that minimizes the prediction error. Furthermore, the choice of a model by $\widehat{\pi^*}(n, N)$ leads asymptotically to the choice of $\mathcal{P}^*$, the smallest model in terms of the number of covariates in $\mathcal{M}$, that is, the most parsimonious model $\mathcal{P}$ that minimizes $E\big(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2$. This property is often referred to as consistency; see, e.g., Shao [24].

In the case of fixed n, Equation (2.13) readily implies that $C^{(\mathcal{P})}(n, N) - AR(n, \mathcal{P}) = O_p(1/\sqrt{N})$. Therefore, as N goes to infinity, the left-hand side converges to zero (at a rate of $1/\sqrt{N}$), implying

Proposition 2.6. *Under the condition of Theorem 2.1, we have for any fixed n, $P\Big(\widehat{\pi^*}(n, N) \subseteq \pi^*(n)\Big) \overset{N\to\infty}{\longrightarrow} 1$.*

In words, Proposition 2.6 says that a model that minimizes $C^{(\mathcal{P})}$ will minimize $AR(n, \mathcal{P})$ with high probability for fixed n and a suitably large N. Proposition 2.4 (i) asserts that minimizing $AR(n, \mathcal{P})$ by $\pi^*(n)$ is close to minimizing $R(n, \mathcal{P})$ by $\mathcal{P}^*(n)$, which is our goal.

3. Several datasets

Our main focus is on the case where several regression datasets are observed. We first discuss the case where we observe datasets from all the regressions of interest, and then, in Section 3.3, we consider a hierarchical situation where the data consist of a random sample of regression datasets from a structured collection of regression models.

3.1. Model selection observing all regressions

We consider a population of distributions $\mathcal{G} = \{G_j : j = 1, \ldots, \mathcal{K}\}$ with $\mathcal{J} = \mathcal{K} < \infty$, that is, the training set comprises of all regression datasets in the population. Thus, we observe data $D_j = \{(\mathbf{X}_{ij}, Y_{ij}) \sim G_j,\ i = 1, \ldots, N_j\}$, $j = 1, \ldots, \mathcal{J}$, and $\mathbf{X}_{ij} \in \mathbb{R}^d$.

For a given n, the goal is to select a common set of covariates $\mathcal{P}$ to be used for prediction of the response Y from $\mathbf{X} = \mathbf{X}^{(\mathcal{P})}$ (the subvector with coordinates in $\mathcal{P}$) for each individual distribution G_j from the population, or equivalently for a random G_J, see below (3.1), where the coefficients $\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$, which are allowed to vary with j, are estimated with a sample of size n. The relevant prediction error for this task is (3.1) below. When predicting for individual j, it may be natural to set $n = N_j$. However, other values of n may be of interest in studying

the contribution of covariates as a function of the sample size. Later (in Section 3.3), we use the $\mathcal{J}$ datasets as a training set for choosing a model to predict for any out-of-sample G_J on the basis of n future observations, where n is not determined in advance since J is not in the training set. In this case we use the chosen subset of covariates, and estimate its parameters on the basis of a dataset of size n from G_J. The value of n may vary, being the size of the dataset G_J.

Let $\mathbf{X} := \mathbf{X}^{(\mathcal{P})} \in \mathbb{R}^p$, where for now $\mathcal{P}$ and its size p are suppressed in the notation. For each j and generic observation $(\mathbf{X}, Y)$ from the distribution G_j, we define

$$\boldsymbol{\beta}_j := \arg\min_{\boldsymbol{\beta}} E_{G_j}(Y - \mathbf{X}'\boldsymbol{\beta})^2 = \mathbb{Q}_j^{-1} E_{G_j}(\mathbf{X}Y);$$

see (2.1), where $\mathbb{Q}_j := E_{G_j}(\mathbf{X}\mathbf{X}')$. Assuming finite fourth moments, we have for a sample size $n \to \infty$, for each j, as in (2.3),

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{j,n} - \boldsymbol{\beta}_j) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbb{Q}_j^{-1}\mathbb{W}_j\mathbb{Q}_j^{-1}) \text{ where } \widehat{\boldsymbol{\beta}}_{j,n} := (\mathbb{X}'_{j,n}\mathbb{X}_{j,n})^{-1}\mathbb{X}'_{j,n}\mathbf{Y}_{j,n},$$

$\mathbb{X}_{j,N}$ and $\mathbf{Y}_{j,N}$ are the jth versions of $\mathbb{X}_N$, and $\mathbf{Y}_N$, and $\mathbb{W}_j := E_{G_j}(\mathbf{X}\mathbf{X}'e^2)$, a $p \times p$ matrix, assumed to be positive definite. We further use the notation $\mathbb{V}_j$ for the jth version of $\mathbb{V}$, that is, when expectations are taken with respect to G_j, and similar notation when $N = N_j$ observations are used for the estimators $\widehat{\mathbb{Q}}_{j,N}$, $\widehat{\mathbb{V}}_{j,N}$, and $\widehat{\mathbb{W}}_{j,N}$ instead of $\widehat{\mathbb{Q}}_N$, $\widehat{\mathbb{V}}_N$, and $\widehat{\mathbb{W}}_N$.

We consider prediction for a random individual regression dataset of size n from the population $\mathcal{G}$, based on a model, that is, a subset of covariates $\mathcal{P}$. As above we suppress $\mathcal{P}$ and write $\mathbf{X}$ and $\boldsymbol{\beta}$ rather than $\mathbf{X}^{(\mathcal{P})}$ and $\boldsymbol{\beta}^{\mathcal{P}}$, etc. The relevant prediction error (see (1.1) and around for a discussion) is

$$\mathbf{R}(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} R_j(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}'\widehat{\boldsymbol{\beta}}_{j,n})^2, \tag{3.1}$$

where $(\mathbf{X}, Y) \sim G_j$ independently of $\widehat{\boldsymbol{\beta}}_{j,n}$, and the expectation on the right-hand side of (3.1) is also applied to $\widehat{\boldsymbol{\beta}}_{j,n}$. The risk $\mathbf{R}(n, \mathcal{P})$ can be interpreted as an expectation over G_J for a uniform choice of a single $J \in \mathcal{G}$ or equivalently, as the risk per task average for the multi-task of predicting for all $G_j \in \mathcal{G}$ if all datasets sizes (or task size) were n. In $\mathbf{R}(n, \mathcal{P})$ above and similar expressions below, we suppress the number of datasets $\mathcal{J}$. In the case that any of the distributions G_j involves discrete covariates, we replace $E_{G_j}(Y - \mathbf{X}'\widehat{\boldsymbol{\beta}}_{j,n})^2$ by a conditional expectation as in Section 2.4, where the conditioning is on a set whose complement is exponentially small. In the definition given in Equation (1.1), (3.1), and others below we use boldface letters when $\mathcal{J} > 1$. Next define

$$\mathbf{AR}(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} AR_j(n, \mathcal{P}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \Big\{ E_{G_j}(Y - \mathbf{X}'\boldsymbol{\beta}_j)^2 + \frac{tr(\mathbb{V}_j)}{n} \Big\}. \tag{3.2}$$

Using (2.9) we have

$$\mathbf{R}(n, \mathcal{P}) = \mathbf{AR}(n, \mathcal{P}) + O(1/n^{3/2}). \tag{3.3}$$

Set

$$C_j^{(\mathcal{P})}(n, N_j) := \frac{1}{N_j}||\mathbf{Y}_{j,N_j} - \mathbb{X}_{j,N_j}\widehat{\boldsymbol{\beta}}_{j,N_j}||^2 + tr(\widehat{\mathbb{V}}_{j,N_j})(1/n + 1/N_j), \text{ and}$$

$$\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} C_j^{(\mathcal{P})}(n, N_j), \tag{3.4}$$

where $\mathbf{N} = (N_1, \ldots, N_{\mathcal{J}})$. We define the jackknife bias-corrected $\mathbf{C}^{(\mathcal{P})}$ by

$$\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N}) := \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \mathbb{C}_j^{(\mathcal{P})}(n, N_j), \tag{3.5}$$

where $\mathbb{C}_j^{(\mathcal{P})}(n, N_j)$ is the bias-corrected $C_j^{(\mathcal{P})}(n, N_j)$; see Efron [9], Equation (2.8), for a precise definition of the jackknife correction we use.

Theorem 3.1 below parallels Theorem 2.1 concerning the error of $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ as an estimator of $\mathbf{AR}(n, \mathcal{P})$.

Theorem 3.1. *Suppose that the conditions of Theorem 2.1 are satisfied when $(\mathbf{X}, Y) \sim G_j$ for each $j = 1, \ldots, \mathcal{J}$. Then,*

$$\begin{aligned}
&\mathbf{AR}(n, \mathcal{P}) - \mathbf{C}^{(\mathcal{P})}(n, \mathbf{N}) \\
&= \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \mathcal{E}_{j,N_j} + \frac{1}{n\mathcal{J}} \sum_{j=1}^{\mathcal{J}} \Big\{ tr(\mathbb{V}_j) - tr(\widehat{\mathbb{V}}_{j,N_j}) \Big\} + \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} o_p\Big(\frac{1}{N_j}\Big),
\end{aligned}$$

where $\mathcal{E}_{j,N_j}$ is the jth version of $\mathcal{E}_N$ defined in (2.11), and the o_p terms do not depend on n.

Moreover, assume that $\lim_{\mathbf{N}\to\infty} N_1/N_j := a_j$ exists for all j, where $0 < a_j < \infty$; then

$$\sqrt{N_1}\big\{\mathbf{AR}(n, \mathcal{P}) - \mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})\big\} \xrightarrow{\mathcal{D}} N(0, \tau_{\mathcal{J}}^2),$$

as $\mathbf{N} \to \infty$, where $\tau_{\mathcal{J}}^2 = \frac{1}{\mathcal{J}^2} \sum_{j=1}^{\mathcal{J}} a_j \tau_j^2$ and τ_j^2 is the asymptotic variance under G_j as in Theorem 2.1, Equation (2.13).

Notice that if τ_j^2 and a_j are bounded (in j), then the asymptotic variance of $\sqrt{N_1}\left\{\mathbf{AR}(n, \mathcal{P}) - \mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})\right\}$ decreases like $1/\mathcal{J}$, which means that the error is decreasing in $\mathcal{J}$. Theorem 3.1 and (3.3) imply properties of $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ as an estimator of $\mathbf{R}(n, \mathcal{P})$ as discussed next.

3.2. Consistency

Analogously to the definitions in Section 2.5, where now the optimal sets of the multi-task problem are denoted using boldface, define

$$\boldsymbol{\mathcal{P}}^*(n) := \arg\min_{\mathcal{P}} \mathbf{R}(n,\mathcal{P}) = \arg\min_{p} \sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}^{(\mathcal{P})'}\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})})^2,$$

$$\boldsymbol{\pi}^*(n) := \arg\min_{\mathcal{P}} \mathbf{AR}(n,\mathcal{P}) = \arg\min_{\mathcal{P}} \sum_{j=1}^{\mathcal{J}} \Big\{ E_{G_j}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}_j^{(\mathcal{P})})^2 + \frac{tr(\mathbb{V}_j^{(\mathcal{P})})}{n} \Big\}, \tag{3.6}$$

$$\boldsymbol{\mathcal{P}}^* := \arg\min_{\mathcal{P}\in\boldsymbol{\mathcal{M}}} |\mathcal{P}| \text{ where } \boldsymbol{\mathcal{M}} := \arg\min_{\mathcal{P}} \sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}_j^{(\mathcal{P})})^2,$$

$$\widehat{\boldsymbol{\pi}^*}(n,\mathbf{N}) := \arg\min_{\mathcal{P}} \mathbf{C}^{(\mathcal{P})}(n,\mathbf{N}).$$

The next result is similar to Proposition 2.4, with essentially by the same proof. The notions *equally good* and *adequate* are the same as that of Section 2.2.

Proposition 3.2. *Suppose that the conditions of the first part of Theorem 3.1 hold. Then*
(i) *Any two sequences in $\boldsymbol{\pi}^*(n)$ and $\boldsymbol{\mathcal{P}}^*(n)$ are equally good, that is, any sequence of models in $\boldsymbol{\pi}^*(n)$ is adequate.*
(ii) *The set $\boldsymbol{\mathcal{P}}^*$ is a singleton and the sets $\boldsymbol{\mathcal{P}}^*(n)$ and $\boldsymbol{\pi}^*(n)$ converge to the singleton $\boldsymbol{\mathcal{P}}^*$ as $n \to \infty$.*

The following proposition generalizes Propositions 2.6 and 2.5 to $\mathcal{J} > 1$. Here we consider a uniform bound (3.7). Technically, the constant C provides a measure of the notion of "sufficiently homogeneous" of Section 1.2 when referring to the set of distributions $\mathcal{G}$; informally we mean that the regression datasets have enough in common to justify common subsets for prediction.

Proposition 3.3. *1. Assume that the conditions of the first part of Theorem 3.1 hold. Then for fixed n we have*

$$\lim_{\mathbf{N}\to\infty} P\left(\widehat{\boldsymbol{\pi}^*}(n,\mathbf{N}) \subseteq \boldsymbol{\pi}^*(n)\right) = 1.$$

2. Let n/N_j be bounded for all $j = 1,\ldots,\mathcal{J}$, and let C be a constant satisfying for all j and $\mathcal{P}$

$$n/N_j, \lambda_{max}(\mathbb{W}_j^{(\mathcal{P})}), 1/\lambda_{min}(\mathbb{W}_j^{(\mathcal{P})}), \lambda_{max}(\mathbb{Q}_j^{(\mathcal{P})}) \le C. \tag{3.7}$$

Then

$$\lim \inf_{n/N_j\le C,\, n\to\infty,\, \mathbf{N}\to\infty} P\left(\widehat{\boldsymbol{\pi}^*}(n,\mathbf{N}) = \boldsymbol{\mathcal{P}}^*(n)\right) \ge 1 - K_C/\mathcal{J},$$

where K_C depends only on C.

The existence of C follows from the assumption on n/N_j since only a finite number of bounded terms appear in (3.7) besides n/N_j.

Part 1 of the above proposition extends Propositions 2.6. Part 2 extends Proposition 2.5; however, a stronger condition was needed before, namely that $n/N \to 0$, to obtain consistency. Here, we obtain approximate consistency for large $\mathcal{J}$, assuming only that n/N_j is bounded, along with the other terms in (3.7). This is useful since in our application it is natural to consider the possibility that $n = N_j$.

The rate $K_C/\mathcal{J}$ in the theorem was achieved by using Chebyshev's inequality. Since under our assumptions all moments are bounded, a similar argument using a bound on $2m$ moments leads in the same way to the rate $K_{C,m}/\mathcal{J}^m$, where $K_{C,m}$ depends also on m, and with further effort, a large deviation rate (in $\mathcal{J}$) can be achieved.

3.3. A population of distributions

We now consider the situation where we have a sample of $\mathcal{J}$ regression datasets from a given, finite or infinite, population of such datasets, and we are interested in predictions for a random (possibly out-of-sample) further regression or several regressions from the same population. In terms of the application considered in this paper, this situation corresponds to the case that we have a training sample of $\mathcal{J}$ doctors out of many more, and our goal is to select a subset of covariates to be used to predict service durations for a random doctor from the population who may not be in the training sample.

Formally, let $(\Theta, \mathscr{T}, \mathscr{P})$ be a probability space and let $\{G_\theta : \theta \in \Theta\}$ be a family of distributions; see, e.g., Çinlar [7], Chapter VI for a formulation of random measures. Let $\{\theta_1, \ldots, \theta_{\mathcal{J}}\}$ be a sample where $\theta_j \sim \mathscr{P}$, and as in Section 3.1 we observe a training set consisting of datasets $D_j = \{(\mathbf{X}_{ij}, Y_{ij}) \sim G_j,\ i = 1, \ldots, N_j\}$, $j = 1, \ldots, \mathcal{J}$, where G_j stands for G_{θ_j}. Given $\theta \in \Theta$, we consider $D = \{(\mathbf{X}_i, Y_i) \sim G_\theta,\ i = 1, \ldots, n\}$. For any function f for which the conditional expectation $E_{G_\theta} f(D)$ of $f(D)$ given G_θ is well defined, we assume that so is $E_{\mathscr{P}} E_{G_\theta} f(D) = \int E_{G_\theta} f(D) \mathscr{P}(d\theta)$, where the outer expectation is over $\theta \sim \mathscr{P}$.

We now fix a set of covariates $\mathcal{P}$, which is suppressed in most of the notation as before. For any G_θ we define $\widehat{\boldsymbol{\beta}}_{\theta,n}$ to be the least squares estimator for the given dataset D. If G_θ is sampled randomly from $\mathscr{P}$ then the population prediction error is defined as

$$\mathbf{R}_{pop}(n, \mathcal{P}) := \int R_\theta(n, \mathcal{P}) \mathscr{P}(d\theta) := \int E_{G_\theta}(Y - \mathbf{X}' \widehat{\boldsymbol{\beta}}_{\theta,n})^2 \mathscr{P}(d\theta), \tag{3.8}$$

where the expectation E_{G_θ} is over $\widehat{\boldsymbol{\beta}}_{\theta,n}$ and $(\mathbf{X}, Y) \sim G_\theta$ that are independent of $\widehat{\boldsymbol{\beta}}_{\theta,n}$. Let $\boldsymbol{\beta}_\theta := \arg\min_{\boldsymbol{\beta}} E_{G_\theta}(Y - \mathbf{X}'\boldsymbol{\beta})^2$, $\mathbb{Q}_\theta := E_{G_\theta}(\mathbf{X}\mathbf{X}')$, $\mathbb{W}_\theta := E_{G_\theta}(\mathbf{X}\mathbf{X}' e^2)$, and $\mathbb{V}_\theta := \mathbb{W}_\theta \mathbb{Q}_\theta^{-1}$. As before, $\mathbf{R}_{pop}(n, \mathcal{P})$ is approximated by

$$\mathbf{AR}_{pop}(n, \mathcal{P}) := \int AR_\theta(n, \mathcal{P}) \mathscr{P}(d\theta) := \int \Big\{ E_{G_\theta}(Y - \mathbf{X}'\boldsymbol{\beta}_\theta)^2 + \frac{tr(\mathbb{V}_\theta)}{n} \Big\} \mathscr{P}(d\theta), \tag{3.9}$$

where the latter integrand defines $AR_\theta(n,\mathcal{P})$ as in (2.7). The quantity $AR_\theta(n,\mathcal{P})$, whose estimation was already discussed, is now a random variable, since it depends on G_θ with $\theta \sim \mathscr{P}$; its expectation, given by (3.9), is the basis of our estimation of $\mathbf{R}_{pop}(n,\mathcal{P})$ of (3.8). Lemma 3.4 below generalizes (2.9).

Lemma 3.4. *Suppose that the conditions of Theorem 2.1 hold uniformly in $\theta \in \Theta$; that is, for each k, the kth moment with respect to G_θ of each entry of $\mathbf{X}$ and Y is bounded uniformly in θ, and the entries of $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ have third moments with respect to G_θ that are bounded uniformly in n and θ. Then*

$$\mathbf{R}_{pop}(n,\mathcal{P}) = \mathbf{AR}_{pop}(n,\mathcal{P}) + O(1/n^{3/2}).$$

The lemma clearly holds if Θ is finite, and in general it follows readily by the uniform boundedness of moments in θ and the proof of (2.9) given in the Appendix. Recall Lemma 2.2, where we showed that the moment conditions of Theorem 2.1 hold when $\mathbf{X}$ is a mixture of normals and $\inf_{\Sigma\in\Xi}\lambda_{min}(\Sigma) > 0$. For the bound on moments as assumed in Lemma 3.4 to hold uniformly, it suffices that $\inf_\Sigma \lambda_{min}(\Sigma) > 0$, where now the infimum is over all covariance matrices of all the mixing normal distributions involved in all the distributions G_θ for all $\theta \in \Theta$. This technical assumption means that the covariates that are taken into account for the model selection are "bounded away" from multicollinearity. For discrete variables we redefine the prediction error by conditioning as in Section 2.4. Theorem 2.3 extends easily when we assume that all covariates are uniformly bounded in θ, and that $\lambda_{min}(\mathbb{Q}_\theta) > c$ for some $c > 0$, for all θ.

Lemma 3.4 suggests that a consistent estimator of $\mathbf{AR}_{pop}(n,\mathcal{P})$ will lead to selection of an adequate model in the sense of Section 2.2, that is, a model that is as good as the model that minimizes $\mathbf{R}_{pop}(n,\mathcal{P})$.

Recall the definition of $\mathbf{AR}(n,\mathcal{P})$ in (3.2); now this quantity is considered random as it is a function of the sampled distributions $G_1,\dots,G_{\mathcal{J}}$. In order to generalize the consistency results of Theorem 3.1 to this case, we need to bound $\mathbf{AR}_{pop}(n,\mathcal{P}) - \mathbf{AR}(n,\mathcal{P})$ as in the lemma below.

Lemma 3.5. *Under the conditions of Lemma 3.4,*

$$\mathbf{AR}_{pop}(n,\mathcal{P}) - \mathbf{AR}(n,\mathcal{P}) = O_p(1/\sqrt{\mathcal{J}}) \tag{3.10}$$

uniformly in n. Moreover, for any fixed n, $\sqrt{\mathcal{J}}\big(\mathbf{AR}_{pop}(n,\mathcal{P}) - \mathbf{AR}(n,\mathcal{P})\big)$ is asymptotically normal.

We now consider the population versions of the quantities defined in Section 3.2.

$$\boldsymbol{\mathcal{P}}^*_{pop}(n) := \arg\min_{\mathcal{P}} \mathbf{R}_{pop}(n,\mathcal{P}) = \arg\min_{\mathcal{P}} \int E_{G_\theta}(Y - \mathbf{X}^{(\mathcal{P})'}\widehat{\boldsymbol{\beta}}^{(\mathcal{P})}_{\theta,n})^2 \mathscr{P}(d\theta),$$

$$\boldsymbol{\pi}^*_{pop}(n) := \arg\min_{\mathcal{P}} \mathbf{AR}_{pop}(n,\mathcal{P})$$

$$= \arg\min_{\mathcal{P}} \int \Big\{ E_{G_\theta}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}_\theta)^2 + \frac{tr(\mathbb{V}^{(\mathcal{P})}_\theta)}{n} \Big\} \mathscr{P}(d\theta),$$

$$\boldsymbol{\mathcal{P}}^*_{pop} := \arg\min_{\mathcal{P}\in\boldsymbol{\mathcal{M}}_{pop}} |\mathcal{P}| \text{ where } \boldsymbol{\mathcal{M}}_{pop} := \arg\min_{\mathcal{P}} \int E_{G_\theta}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}_\theta)^2 \mathscr{P}(d\theta).$$

$\widehat{\boldsymbol{\pi}}^*(n,\mathbf{N})$ and $\boldsymbol{\pi}^*(n)$ are defined as in (3.6), however the fact that now the G_j's are random adds randomness to $\widehat{\boldsymbol{\pi}}^*(n,\mathbf{N})$, and makes $\boldsymbol{\pi}^*(n)$ a random variable.

Proposition 3.6 parallels Proposition 3.3; it shows consistency properties of $\widehat{\boldsymbol{\pi}}^*(n,\mathbf{N})$, as defined in (3.6) using (3.4). Below, the probability P is obtained by first conditioning on $\theta_1,\ldots,\theta_{\mathcal{J}}$, and then unconditioning by taking expectation over $\theta_1,\ldots,\theta_{\mathcal{J}}$ with respect to the product measure $\mathscr{P}^{\mathcal{J}}$.

Proposition 3.6. *Assume that the conditions of Lemma 3.4 hold, and in addition that* $n/N_j, \lambda_{max}(\mathbb{W}_\theta^{(\mathcal{P})}),\ 1/\lambda_{min}(\mathbb{W}_\theta^{(\mathcal{P})}), \lambda_{max}(\mathbb{Q}_\theta^{(\mathcal{P})}) \le C$ *for all* θ *and* $\mathcal{P}$.

1. *When* n *is fixed,*

$$\lim_{\mathbf{N}\to\infty}\inf P\Big(\widehat{\boldsymbol{\pi}}^*(n,\mathbf{N}) \subseteq \boldsymbol{\pi}^*_{pop}(n)\Big) \ge 1-\frac{K_C}{\mathcal{J}}.$$

2. *Letting* $n,\mathbf{N}\to\infty$,

$$\lim_{n/N_j\le C,\, n\to\infty,\, \mathbf{N}\to\infty}\inf P\left(\widehat{\boldsymbol{\pi}}^*(n,\mathbf{N}) = \boldsymbol{\mathcal{P}}^*_{pop}(n)\right) \ge 1-\frac{K_C}{\mathcal{J}},$$

where K_C *depends only on* C.

The proof of Proposition 3.6 also shows that $\boldsymbol{\mathcal{P}}^*_{pop}$ is a singleton and $\boldsymbol{\mathcal{P}}^*_{pop}(n)$ and $\boldsymbol{\pi}^*_{pop}(n)$ converge to it when $n\to\infty$.

4. GENO

4.1. Definition of GENO

Given a model (i.e., a set of covariates) $\mathcal{P}$ with coefficients estimated by a sample of n observations, we can say that it is equivalent to another model $\mathcal{Q}$ with m observations if their expected prediction errors satisfy $R(m,\mathcal{Q}) = R(n,\mathcal{P})$. Using the approximation $AR(n,\mathcal{P})$ to $R(n,\mathcal{P})$ given in (2.7), (3.2), and (3.9) for each of the scenarios we consider, we define GENO by

$$\mathrm{GENO}(n;\mathcal{P},\mathcal{Q}) := \big\{m : AR(m,\mathcal{Q}) = AR(n,\mathcal{P})\big\}. \tag{4.1}$$

If $AR(m,\mathcal{Q}) > (<) AR(n,\mathcal{P})$ for all m, we set $\mathrm{GENO}(n,\mathcal{P},\mathcal{Q}) = \infty(0)$, indicating that model $\mathcal{P}$ with n observations is better than model $\mathcal{Q}$ with any number of observations (model $\mathcal{Q}$ with any number of observations is better than $\mathcal{P}$ with n). A direct calculation shows that for $\mathcal{J}=1$ we have

$$\mathrm{GENO}(n;\mathcal{P},\mathcal{Q}) = tr(\mathbb{V}^{(\mathcal{Q})})\big\{AR(n,\mathcal{P}) - AR(n,\mathcal{Q}) + \frac{1}{n}tr\mathbb{V}^{(\mathcal{Q})}\big\}^{-1}.$$

For $\mathcal{J}>1$ with $\mathbf{AR}$ defined in (3.2) we have

$$\mathrm{GENO}(n;\mathcal{P},\mathcal{Q}) = \Big[\frac{1}{\mathcal{J}}\sum_j tr(\mathbb{V}_j^{(\mathcal{Q})})\Big]\Big\{\mathbf{AR}(n,\mathcal{P}) - \mathbf{AR}(n,\mathcal{Q}) + \frac{1}{\mathcal{J}n}\sum_j tr(\mathbb{V}_j^{(\mathcal{Q})})\Big\}^{-1}.$$

For the case of (3.9), j is replaced by θ, and the averages by integrals $\mathscr{P}(d\theta)$.

$\mathrm{GENO}(n;\mathcal{P},\mathcal{Q}) = m$ means that model $\mathcal{P}$ with n observations is equivalent in terms of expected prediction error to model $\mathcal{Q}$ with m observations. Note that the larger $\mathrm{GENO}(n;\mathcal{P},\mathcal{Q})$ is, the better model $\mathcal{P}$ (with n observations) is relative to model $\mathcal{Q}$. For each model $\mathcal{P}$ and sample size n, we define

$$\mathrm{GENO}(n,\mathcal{P}) = \min_{\mathcal{R}} \mathrm{GENO}(n;\mathcal{P},\mathcal{R}), \tag{4.2}$$

where the minimum is over all subsets of covariates $\mathcal{R}$. It follows that the inequality $\mathrm{GENO}(n,\mathcal{P}) \le n$ holds always, where equality means that $\mathcal{P}$ is the best model for n observations, as no other model can achieve the same prediction error with fewer observations. On the other hand, $\mathrm{GENO}(n,\mathcal{P}) = m < n$ means that there is a model that achieves, with $m < n$ observations, the same prediction error as $\mathcal{P}$ with n observations. Thus, small values of $\mathrm{GENO}(n,\mathcal{P})$ suggest considering another model. By the monotonicity of $AR(n,\mathcal{P})$ in n, if the inequality $\mathrm{GENO}(n;\mathcal{P},\mathcal{R}) \ge \mathrm{GENO}(n;\mathcal{Q},\mathcal{R})$ holds for some model $\mathcal{R}$, then it holds all $\mathcal{R}$. This readily implies

$$\begin{aligned} AR(n,\mathcal{P}) \le AR(n,\mathcal{Q}) &\Leftrightarrow \mathrm{GENO}(n;\mathcal{P},\mathcal{R}) \ge \mathrm{GENO}(n;\mathcal{Q},\mathcal{R}) \quad \text{for all } \mathcal{R} \\ &\Leftrightarrow \mathrm{GENO}(n,\mathcal{P}) \ge \mathrm{GENO}(n,\mathcal{Q}). \end{aligned} \tag{4.3}$$

4.2. Estimation of GENO

In the case $\mathcal{J} = 1$, (2.13) shows the consistency of $C^{(\mathcal{P})}(n,N)$ as an estimator of $AR(n,\mathcal{P})$ for fixed n as $N \to \infty$. In view of (4.1) we define a consistent estimator of $\mathrm{GENO}(n;p,q)$ by

$$\widehat{\mathrm{GENO}}(n;p,q) := \big\{m : C^{(\mathcal{Q})}(m,N) = C^{(\mathcal{P})}(n,N)\big\}.$$

To avoid cumbersome notation we suppress N in $\widehat{\mathrm{GENO}}$. Using (2.8) we obtain, as before,

$$\widehat{\mathrm{GENO}}(n;\mathcal{P},\mathcal{Q}) = tr(\widehat{\mathbb{V}}_N^{(\mathcal{Q})})\Big\{C^{(\mathcal{P})}(n,N) - C^{(\mathcal{Q})}(n,N) + \frac{tr(\widehat{\mathbb{V}}_N^{(\mathcal{Q})})}{n}\Big\}^{-1}, \tag{4.4}$$

setting it to be ∞ if the expression in curly brackets is negative or zero.

In the case of $\mathcal{J} > 1$ datasets of Section 3.1, or in the population case of Section 3.3, the above expression (4.4) remains unchanged except that now $C^{(\mathcal{P})}(n,N)$ is replaced by $\mathbf{C}^{(\mathcal{P})}(n,\mathbf{N})$ defined in (3.4), and $\widehat{\mathbb{V}}_N^{(\mathcal{Q})}$ is replaced by $\frac{1}{\mathcal{J}}\sum_j tr(\widehat{\mathbb{V}}_{j,N_j}^{(\mathcal{Q})})$. We can also define the estimator of (4.4) in terms of the jackknife bias-corrected $\mathbb{C}^{(\mathcal{P})}(n,\mathbf{N})$ of (3.5). This is done in estimating GENO in Section 6.3. The results below hold in the same way for all these cases. Similarly to (4.2), we define the statistic

$$\widehat{\mathrm{GENO}}(n,\mathcal{P}) := \min_{\mathcal{R}} \widehat{\mathrm{GENO}}(n;\mathcal{P},\mathcal{R}),$$

which is an estimate the minimal number of observations required by the best competing model to achieve the same prediction error as model $\mathcal{P}$ with sample size n.

As in (4.3), we have

$$C^{(\mathcal{P})}(n,N) < C^{(\mathcal{Q})}(n,N) \Leftrightarrow \widehat{\mathrm{GENO}}(n;\mathcal{P},\mathcal{R}) \geq \widehat{\mathrm{GENO}}(n;\mathcal{Q},\mathcal{R})\ \forall \mathcal{R}$$
$$\Leftrightarrow \widehat{\mathrm{GENO}}(n,\mathcal{P}) \geq \widehat{\mathrm{GENO}}(n,\mathcal{Q}).$$

The next proposition follows from (2.13) by applying the δ-method to the inverse function in (4.4). In particular, it shows the consistency of $\widehat{\mathrm{GENO}}(n;\mathcal{P},\mathcal{Q})$ for fixed n as $N \to \infty$.

Proposition 4.1. *Under the conditions of Theorem 3.1 (which include the case $\mathcal{J}=1$), we have for any fixed n*

$$\sqrt{N_1}\big(\widehat{\mathrm{GENO}}(n;\mathcal{P},\mathcal{Q}) - \mathrm{GENO}(n;\mathcal{P},\mathcal{Q})\big) \xrightarrow{\mathcal{D}} N(0,\eta^2) \text{ as } \mathbf{N} \to \infty,$$

for some $\eta^2 > 0$.

The variance η^2 is not given explicitly since it is too complicated to be useful, and it can be be computed by the bootstrap. See Theorem 3.1 and the ensuing comment, which show that (under certain conditions) the variance decreases at a rate of $1/\mathcal{J}$.

A similar problem is to estimate for a given model $\mathcal{P}$ and a certain prescribed prediction error E the sample size n that satisfies $AR(n,\mathcal{P}) = E$. When $\mathcal{J} = 1$, using (2.7) this quantity is given by $\dfrac{tr(\mathbb{V}^{(\mathcal{P})})}{E - E\big(Y - \mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2}$ and can be estimated by

$$\frac{tr(\widehat{\mathbb{V}}_N)}{E - \frac{1}{N}\{||\mathbf{Y}_N - \mathbb{X}_N^{(\mathcal{P})}\widehat{\boldsymbol{\beta}}_N^{(\mathcal{P})}||^2 + tr(\widehat{\mathbb{V}}_N^{(\mathcal{P})})\}} \tag{4.5}$$

(Since $\frac{1}{N}\{||\mathbf{Y}_N - \mathbb{X}_N^{(\mathcal{P})}\widehat{\boldsymbol{\beta}}_N^{(\mathcal{P})}||^2 + tr(\widehat{\mathbb{V}}_N^{(\mathcal{P})})\}$ is an unbiased estimator of $E\big(Y - \mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2$); the extensions to the cases $\mathcal{J} > 1$ and to the population setup are straightforward.

5. Simulations

In this section we evaluate by simulations the prediction error $R(n,\mathcal{P})$, its approximation $AR(n,\mathcal{P})$, and its estimation using $C^{(\mathcal{P})}$. We start with a single dataset ($\mathcal{J}=1$) and then we consider the case of several datasets. This simple example demonstrates the well-known difficulty involved in model selection for a single given dataset with methods such as Mallows C_p, AIC, BIC, as well as our version $C^{(\mathcal{P})}$. In Section 5.2 we compare the case of model selection for one dataset to that of choosing a common model for successful prediction on the average when we have data from several datasets, that is, a multi-task. Section 5.3 compares the prediction error when model selection is done according to $\mathbf{C}^{(\mathcal{P})}$ to the prediction error under other methods.

5.1. A single dataset

Suppose that the distribution of $(\mathbf{X}, Y)$ for $\mathbf{X} \in \mathbb{R}^5$ is given by

$$Y = b_0 + b_1 X_1 + \ldots + b_5 X_5 + a(X_1^2 - 1) + \sigma\varepsilon, \tag{5.1}$$

with $X_1, \ldots, X_5, \varepsilon \sim^{iid} N(0,1)$. Setting all models to include the intercept, there are 2^5 possible submodels; for simplicity, we focus for now on two models consisting of the subsets of covariates $\mathcal{P}_1 = \{1, X_1\}$, $\mathcal{P}_2 = \{1, X_1, \ldots, X_5\}$; more explicitly, we have model $\mathcal{P}_1$: $Y = \beta_0 + \beta_1 X_1 + e$ and model $\mathcal{P}_2$: $Y = \beta_0 + \beta_1 X_1 + \ldots \beta_5 X_5 + e$. These two models are wrong (as linear conditional expectation function models, see Hansen [12] Section 2.15) since the residual e includes the nonlinear term $X_1^2 - 1$. By the orthogonality of the variables in (5.1), the projection parameters β_k are equal to b_k for these models; see (2.1). This is used in computing the first part of $AR(n, \mathcal{P}_\ell)$ for $\ell = 1, 2$, and since in this case $\mathbb{Q} = I$, it is also easy to compute $tr(\mathbb{V})$ for each model. We obtain

$$AR(n, \mathcal{P}_1) = \sum_{k=2}^{5} b_k^2 + 2a^2 + \sigma^2 + \frac{2(\sum_{k=2}^{5} b_k^2 + \sigma^2) + 12a^2}{n}, \tag{5.2}$$

$$AR(n, \mathcal{P}_2) = 2a^2 + \sigma^2 + \frac{6\sigma^2 + 20a^2}{n}; \tag{5.3}$$

notice that the above functions do not depend on b_0, b_1. For a concrete example, we set in (5.1)

$$(b_0, b_1) = (1, 3),\ (b_2, \ldots, b_5) = (1, \ldots, 1),\ a = 1,\ \sigma = 7. \tag{5.4}$$

Figure 1 plots $R(n, \mathcal{P}_\ell)$ (see (2.4)-(2.5)) (solid lines) and $AR(n, p_\ell)$ (see (5.2)) (dashed line), $\ell = 1, 2$, as functions of n for the above parameters. We evaluated $R(n, \mathcal{P}_\ell)$, where $\ell = 1, 2$, by a simulation based on 10^3 repetitions and using the decomposition (see (7.1) and recall that $\mathbb{Q} = I$)

$$R(n, \mathcal{P}_\ell) = E\big(Y - (\mathbf{X}^{(\mathcal{P}_\ell)})'\boldsymbol{\beta}^{(\mathcal{P}_\ell)}\big)^2 + E\|\hat{\boldsymbol{\beta}}^{(\mathcal{P}_\ell)} - \boldsymbol{\beta}^{(\mathcal{P}_\ell)}\|^2;$$

the first expectation can be computed explicitly and the second is evaluated using simulations. For small n, $R(n, \mathcal{P}_2)$ differs from $AR(n, \mathcal{P}_2)$, and the approximation improves as n increases. For n smaller than about 50, model $\mathcal{P}_1$ has a smaller prediction error; for large n model $\mathcal{P}_2$ is better. This holds approximately for both R and AR. This makes sense as models with fewer parameters have a smaller prediction error for small n. The rest of the models are not optimal for any n (this observation is not shown in Figure 1).

Consider GENO as defined in (4.1). Careful inspection of Figure 1 shows, for example, that $\text{GENO}(49; \mathcal{P}_1, \mathcal{P}_2) = 49$, which means that in order to achieve the same prediction error as model $\mathcal{P}_1$ with $n = 49$ observations (the value of n where the dashed black line and red the line intersect), model $\mathcal{P}_2$ requires the same number of observations. Also, $\text{GENO}(60; p_2, p_1) = 95$, and therefore, to achieve the same prediction error as model $\mathcal{P}_2$ with $n = 60$, model $\mathcal{P}_1$ would

require 95 observations (the value of n where the dashed black line has the same level as the dashed red line at 60). Since the decrease of $AR(n, \mathcal{P}_1)$ (the black line) in n is slow, a small increase in n, will result in a much larger value of $\mathrm{GENO}(n; \mathcal{P}_2, \mathcal{P}_1)$; for example, $\mathrm{GENO}(65; \mathcal{P}_2, \mathcal{P}_1) = 142$. As mentioned before, GENO allows the statistician to compare the cost of additional observations to the cost of measuring additional variables, which may be expensive, or harmful, such as in the case of an invasive medical procedure or imaging that involves radiation.

By (4.5), the numbers of observations for models $\mathcal{P}_1$ and $\mathcal{P}_2$ to obtain a prediction error of 59 are about 29 and 39, respectively; i.e., model $\mathcal{P}_1$ can achieve this prediction error with a sample size that is smaller by 10 observations. On the other hand, for a prediction error of 56, model $\mathcal{P}_1$ requires 118 observations, while model $\mathcal{P}_2$ needs only 63 observations.

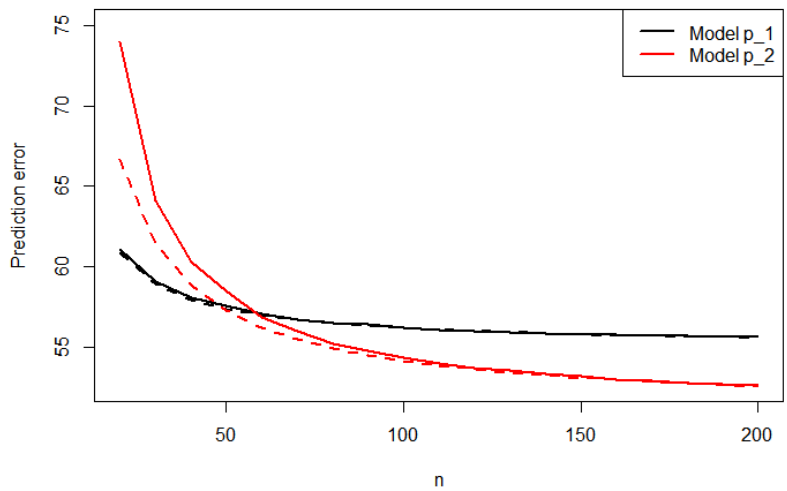
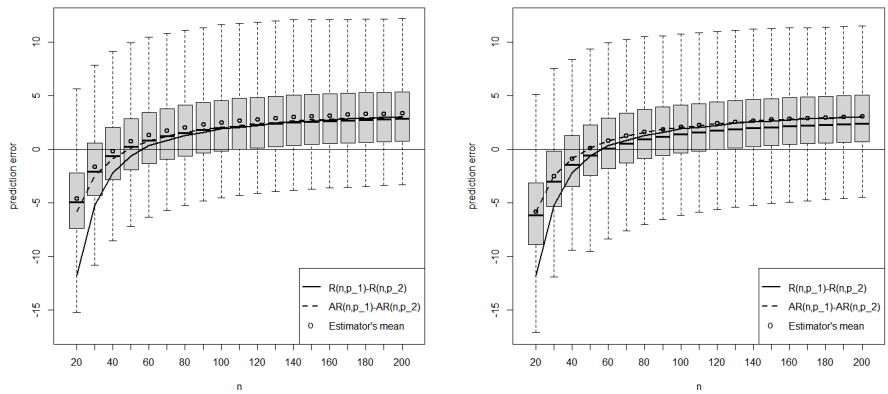


Fig 1: Simulation estimates of $R(n, \mathcal{P}_1)$ and $R(n, \mathcal{P}_2)$ (solid line) as well as the approximations AR (dashed line) given in (5.2).

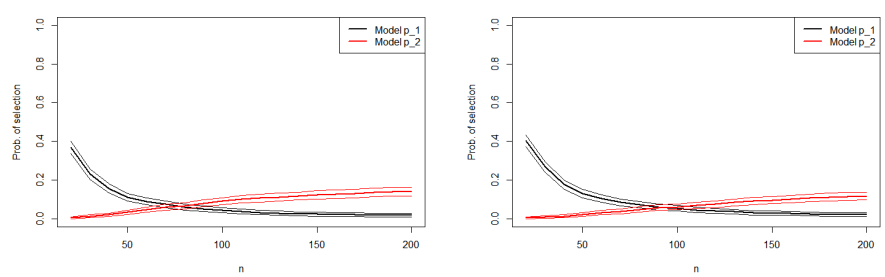
We now discuss estimation of the prediction error using $C^{(\mathcal{P})}(n, N)$ based on a single dataset of size $N = 100$. Figure 2 plots $R(n, \mathcal{P}_1) - R(n, \mathcal{P}_2)$ (solid line), $AR(n, \mathcal{P}_1) - AR(n, \mathcal{P}_2)$ (dashed line), and boxplots of the estimators $C^{(\mathcal{P}_1)}(n, N) - C^{(\mathcal{P}_2)}(n, N)$ on the left-hand side, and $\mathbb{C}^{(\mathcal{P}_1)}(n, N) - \mathbb{C}^{(p_2)}(n, N)$, the jackknife bias-corrected version, on the right-hand side, based on 10^3 simulations for each $n = 20, 40, \ldots, 200$. Their means are given by circles. We see that the jackknife corrects the bias of $C^{(\mathcal{P}_1)}(n, N) - C^{(\mathcal{P}_2)}(n, N)$ as an estimator of $AR(n, \mathcal{P}_1) - AR(n, \mathcal{P}_2)$; see the discussion following (2.8). Recall that the bias itself and the correction decrease in n. The mean of the difference $C^{(\mathcal{P}_1)}(n, N) - C^{(\mathcal{P}_2)}(n, N)$

and $\mathbb{C}^{(\mathcal{P}_1)}(n,N) - \mathbb{C}^{(\mathcal{P}_2)}(n,N)$ equals 0 at about $n = 40$ and $n = 50$, respectively; thus the jackknife leads to correct selection on average since it is optimal to select model $\mathcal{P}_1$ for about $n \leq 50$.



(a) Boxplots of $C^{(\mathcal{P}_1)}(n,N) - C^{(\mathcal{P}_2)}(n,N)$ (b) Boxplots of $\mathbb{C}^{(\mathcal{P}_1)}(n,N) - \mathbb{C}^{(\mathcal{P}_2)}(n,N)$

Fig 2: (a) Boxplots of the simulation results of $C^{(\mathcal{P}_1)}(n,N) - C^{(\mathcal{P}_2)}(n,N)$, where $\circ$ (circle) denotes the mean, and $R(n,\mathcal{P}_1) - R(n,\mathcal{P}_2)$ (respectively, $AR(n,\mathcal{P}_1) - AR(n,\mathcal{P}_2)$) is a solid (respectively, dashed) line. (b) Same as (a) for jackknifed version $\mathbb{C}^{(\mathcal{P}_1)}(n,N) - \mathbb{C}^{(\mathcal{P}_2)}(n,N)$.



(a) Probability of selection by $C^{(\mathcal{P})}(n,N)$ (b) Probability of selection by $\mathbb{C}^{(\mathcal{P})}(n,N)$

Fig 3: Probability of selecting model p_1; the thick line is the simulation mean and the thin lines are plus and minus two standard errors.

Figure 3 depicts simulation estimates of the probability of selecting models $\mathcal{P}_1$ and $\mathcal{P}_2$ as a function of n, using $C^{(\mathcal{P})}(n,\mathbf{N})$ and the jackknifed $\mathbb{C}^{(\mathcal{P})}(n,\mathbf{N})$, where all 2^5 possible sub-models $\mathcal{P}$ are considered; for clarity we present the curves of $\mathcal{P}_1$ and $\mathcal{P}_2$ only. For each n and for each simulated dataset, $C^{(\mathcal{P})}(n,\mathbf{N})$ and $\mathbb{C}^{(\mathcal{P})}(n,\mathbf{N})$ are calculated for all $\mathcal{P}$. The empirical averages over 100 simulations

of selecting models $\mathcal{P}_1$ and $\mathcal{P}_2$ out of the 2^5 sub-models for each n are plotted in Figure 3. The bias correction increases the probability of selecting model $\mathcal{P}_1$ for small n. This improves the selection for small or moderate values of n. For the problem of selecting a common model for $\mathcal{J}$ datasets, the bias correction becomes more significant, as demonstrated next.

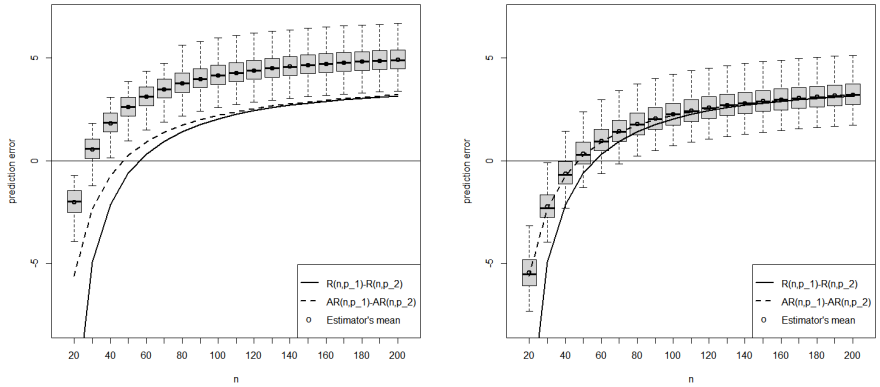
5.2. Multiple datasets

We now consider the case of $\mathcal{J} > 1$ datasets. Suppose that G_θ is given by model (5.1) with $b_{\theta,0} = 1$, $b_{\theta,k} = W_k(b_k + Z_k)$ for $k = 1, \ldots, 5$, where b_k is given in (5.4), $(Z_1, \ldots, Z_5) \sim N(0, 0.2^2)$, W_k is ± 1 with equal probability, and all the above random variables are independent (thus determining the distribution $\mathscr{P}$ of Section 3.3), and then fixed throughout this section. The expected $b^2_{\theta,k}$ is approximately equal to b^2_k in (5.4), but about half of the $b_{\theta,k}$'s are positive and half are negative. The number of regression datasets is $\mathcal{J} = 100$, and $N_j = 20, 100$, and 200 for $1 \le j \le 33$, $34 \le j \le 66$, $67 \le j \le 100$, respectively.

In the case of observing all regressions (see Section 3.1, Equation (1.1)), we wish to estimate $\mathbf{R}(n, \mathcal{P})$, whereas in the case of observing a sample of regressions from the distribution $\mathscr{P}$ (see Section 3.3, Equation (3.8)), the relevant quantity is $\mathbf{R}_{pop}(n, \mathcal{P})$. Computing the latter quantity is difficult, and instead we use the approximation $\mathbf{R}(n, \mathcal{P})$, which is justified by the law of large numbers and the central limit theorem (see Lemma 3.5). Thus we now focus on estimating $\mathbf{R}(n, \mathcal{P})$ and selecting according to its estimate. The plot of $\mathbf{R}(n, \mathcal{P})$ for $\mathcal{J}$=100 and $\mathcal{P} = \mathcal{P}_1, \mathcal{P}_2$ is similar to Figure 1 and therefore is not presented here.

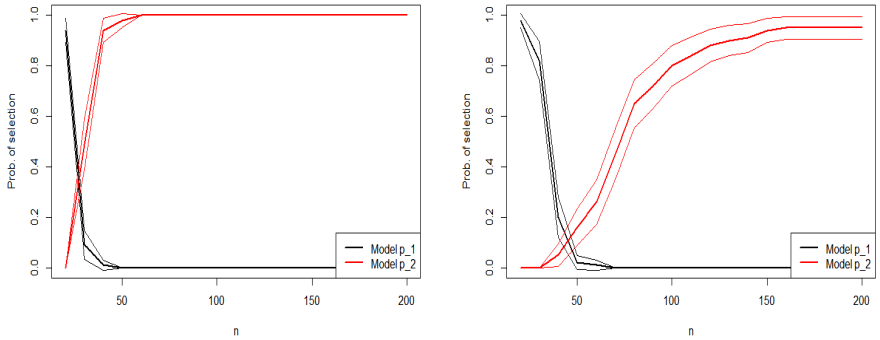
Figure 4 parallels Figure 2, where now in the case of $\mathcal{J}$ datasets, $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ and $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$ replace $C^{(\mathcal{P})}(n, N)$, and $\mathbb{C}^{(\mathcal{P})}(n, N)$, respectively; see (3.4) and (3.5); the number of simulations to evaluate $\mathbf{R}_{pop}(n, \mathcal{P})$ and to produce Figures 4 and 5 is 100. We see that the jackknife bias correction works well. Here the variances of the estimates are much smaller, indicating that several datasets can lead to better estimates and model selection procedures, as predicted by theory. The y-axis scale varies between Figures 4 and 2, in a way that undermines their difference.

Figure 5 plots the selection probabilities as a function of n (out of all 2^5 sub-models). Unlike the case $\mathcal{J} = 1$ (see Figure 3), model $\mathcal{P}_1$ (respectively, model $\mathcal{P}_2$) is selected with high probability for small n (respectively, large n). Recall that it is optimal to select model $\mathcal{P}_1$ (respectively, $\mathcal{P}_2$) when $n \le 50$, (respectively, $n \ge 50$). Selecting according to $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ leads to favoring $\mathcal{P}_2$ (or other models) when n is greater than approximately 25 (instead of 50) and $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$ corrects this bias. Thus, the probability of correct model selection is much higher when using the $\mathcal{J} = 100$ datasets (see Figure 1). Clearly, the probability of making a correct selection depends on the number of datasets $\mathcal{J}$, the similarity among the $\mathcal{J}$ models, the noise level in the models, and the sample size n.



(a) Boxplots of $\mathbf{C}^{(\mathcal{P}_1)}(n, \mathbf{N}) - \mathbf{C}^{(\mathcal{P}_2)}(\mathbf{n}, \mathbf{N})$ (b) Boxplots of $\mathbb{C}^{(\mathcal{P}_1)}(n, \mathbf{N}) - \mathbb{C}^{(\mathcal{P}_2)}(\mathbf{n}, \mathbf{N})$

Fig 4: Same plots as in Figure 2 when there are $\mathcal{J} = 100$ samples.



(a) Probability of selection by $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ (b) Probability of selection by $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$

Fig 5: Same plots as in Figure 3 when there are $\mathcal{J} = 100$ samples.

5.3. Comparisons to other approaches

We considered the possibility of concatenating the whole training sample and performing a single regression with an intercept for each j. In this simulation, since about half of the $b_{\theta,k}$'s are positive and half are negative, the resulting regression model leads to a higher prediction error than the one of $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$. The latter has estimated prediction error of 56.1 (SE=0.03) (see Table 1 below), while for the ordinary least squares applied to the concatenated dataset the corresponding number is 63.6 (SE=0.1), computed by averaging the prediction

error over 1000 independent datasets with the same distribution. For ridge and lasso estimators applied to the concatenated dataset (calculated using the glmnet package, where the tuning parameter was computed using cross-validation), the prediction error was slightly higher: 64.1 (SE=0.1) and 63.9 (SE=0.1) for ridge and lasso respectively.

Another approach is to consider a separate model selection algorithm for each of the $\mathcal{J}$ datasets. We considered three selection criteria: $\mathbb{C}^{(\mathcal{P})}(n, N_j)$ with $n = N_j$ as in (2.8) (applied to each dataset separately), Mallows' C_p and BIC. The means of the resulting prediction errors are given in the Table 1 below as well as that of $\mathbf{C}^{(\mathcal{P})}(n, N)$ (where the same model is selected for all j's with the same sample size N_j). The datasets are divided into three categories according to their sample sizes and the mean is reported for each category separately. Recall that $N_j = 20, 100$, and 200 for $1 \leq j \leq 33$, $34 \leq j \leq 66$, and $67 \leq j \leq 100$, respectively. The prediction error $R_j(N_j, \mathcal{P}_\ell)$ was evaluated as in Figure 1. Table 1 shows that $\mathbf{C}^{(\mathcal{P})}(n, N)$ leads to smaller prediction errors and the improvements is higher for smaller sample-sizes, where borrowing power from other datasets is more important.

TABLE 1
The means of the prediction errors $R_j(N_j, \mathcal{P}_\ell)$, where $\mathcal{P}_\ell$ is selected by different methods. The standard errors are about 0.03.

	Model selection method			
N_j	$\mathbf{C}^{(\mathcal{P})}(n, N)$	$\mathbb{C}^{(\mathcal{P})}(n, N_j)$	Mallows' C_p	BIC
20	61.4	67.4	66.4	66.1
100	54.4	55.2	55.2	56.1
200	52.7	53.4	53.4	54.5
Average	56.1	58.7	58.3	58.9

6. Prediction of durations of medical examinations

In this section we analyze a dataset of outpatients' hospital visits. Different models are considered in order to predict the actual appointments' durations as opposed to the planned durations.

6.1. Description of the data

The dataset analyzed is taken from the SEE Lab at the Technion. It consists of information on 140,924 hospital visits that took place in a certain US hospital for about two years between 2013 and 2015. For each visit, both the planned time and the actual time are reported. The goal was to provide a more accurate prediction of the actual duration than the planned one. In this dataset there is information on 44,516 patients and 258 doctors, out of whom 34 doctors had fewer than 50 visits. We shall focus on the rest, which corresponds to 99.5% of all visits. The regression coefficients will differ between doctors, and the goal

is to select one common subset of covariates (for each n) for all doctors for prediction of visit durations.

The distribution of the planned duration is given in Table 2 and Figure 6 plots the estimated density (a normal kernel estimate using the R command "density") of the actual durations for the time slots of 15, 30, and 60 minutes. Actual durations are obtained by a real-time location system (RTLS). The means are 16.7, 21.3, and 41.2, respectively.

TABLE 2
The distribution of the planned duration.

minutes	15	30	45	60	other
percentage	29.8%	52.6%	1.8%	15.5%	0.3%

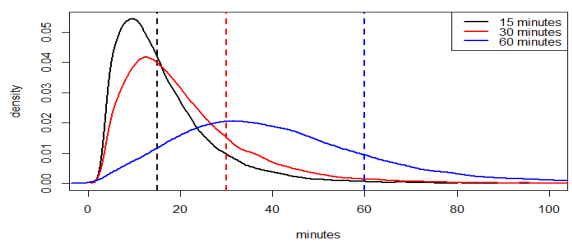


Fig 6: Estimated density of the actual duration for the time slots of 15, 30, and 60 minutes. The vertical dashed lines are at 15, 30, and 60 minutes.

6.2. A regression model

The original dataset contains a large number of covariates, of which many did not seem to have any predictive power relative to visit durations. For simplicity of presentation, we focus on a small number of covariates that seem most relevant. We aim to predict actual duration, using the following covariates:

- duration_planned = the planned duration of the visit in minutes.
- duration_planned_2 = the planned duration in minutes of the visit, squared.
- last = the planned minus the actual duration of the previous visit of the same patient (taken to be 0 for the first visit of the patient).
- hour_end = whether the exam is planned to end on the hour. It turns out that these kinds of visits tend to be somewhat longer.
- type = there are two types of examinations: consultation/examination only, or the above plus treatment. In either case, only the first part counts as duration.

Standard statistical inference of the linear regression model of the whole dataset (ignoring the doctors' index) reveals that all of the above covariates besides "type" are statistically significant; however, the standard error of the residuals is 15.33, and $R^2 = 0.227$, suggesting that the prediction error is quite large.

6.3. $\mathbb{C}^{(\mathcal{P})}$ *and model selection*

In our notation, each doctor is indexed by j, and N_j is the number of visits to doctor j in the dataset; N_j varies between 50 and 2135. We demonstrate our approach by focusing on four candidate models that have the smallest (or nearly smallest) $\mathbb{C}^{(\mathcal{P})}$ from all submodels of the five covariates (all models included the intercept term) for relevant sample sizes n. These models are $\mathcal{P}_1$ – the model with the covariates: duration_planned, duration_planned_2; $\mathcal{P}_2$ – the model with the same covariates as in $\mathcal{P}_1$ and additionally, the variable "last"; $\mathcal{P}_3$ – the model with the same covariates of $\mathcal{P}_2$ and additionally, the variable "type"; and $\mathcal{P}_4$ – the full model. For certain submodels estimation is possible only for a subset of the doctors since $\mathbb{X}_{j,N_j}^{(\mathcal{P})'}\mathbb{X}_{j,N_j}^{(\mathcal{P})}$ is not always invertible. Therefore $\mathcal{J}$ varies between the models. For the models $\mathcal{P}_1$ and $\mathcal{P}_2$, invertibility held for 96 doctors and for the models $\mathcal{P}_3$ and $\mathcal{P}_4$, the corresponding number is 95, and so for these models $\mathcal{J} = 96$ or $\mathcal{J} = 95$. In this case, $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$ is based only on this subset.

Figure 7 plots $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$ for $\mathcal{P} = \mathcal{P}_\ell$ where $\ell = 1, \ldots, 4$ and n is between 50 and 500. For n smaller than approximately 80, model $\mathcal{P}_1$ is the best among the candidate models; for n between 80 and 450, $\mathcal{P}_2$ has a smaller $\mathbb{C}^{(\mathcal{P})}$, and for larger n, $\mathcal{P}_3$ is the best, but $\mathcal{P}_2$ is very close. In terms of GENO, we have, for example, that for $n = 50$, $\widehat{\mathrm{GENO}}(n, \mathcal{P}_1, \mathcal{Q})$ for $\mathcal{Q} = \mathcal{P}_2, \mathcal{P}_3$, and $\mathcal{P}_4$ equals 54, 63, 73, respectively. The latter number means that model $\mathcal{P}_4$ (the full model) would require 73 observations to achieve the same prediction error as model $\mathcal{P}_1$ with $n = 50$ observations. Also, $\widehat{\mathrm{GENO}}(200, \mathcal{P}_2, \mathcal{P}_1) = 370$; if one considers using only the planned duration ($\mathcal{P}_1$) or using model $\mathcal{P}_2$, that is, adding the variable "last" with the information on the last visit, which may not be available for some patients, then the estimated prediction error by the model $\mathcal{P}_2$ with $n = 200$ observations can be achieved without knowing "last" by $\mathcal{P}_1$, with $n = 370$. It is then left to the user to decide whether to invest in measuring "last" or in using a larger sample, if such a sample is available.

Table 3 reports $\mathbb{C}^{(\mathcal{P}_\ell)}(n, \mathbf{N})$ for different sample sizes n. Standard deviations estimated by the bootstrap, and cross-validation estimates of $\mathbf{R}(n, \mathcal{P})$, are also provided. The latter estimates are computed only for j's where $N_j > n$. For each such j, the data were split at random into a training set with n observations, and a testing set of size $N_j - n$. The estimates $\hat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$ are based on the training set and the prediction error $R_j(n, \mathcal{P})$ is evaluated using the testing set. This procedure was repeated 1,000 times and the average prediction error is reported. The cross-validation estimates are mostly within one standard error of the $\mathbb{C}^{(\mathcal{P})}$ values, and the two approaches lead to selection of the same models.

Table 4 reports the values of $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N}) - \mathbb{C}^{(\mathcal{Q})}(n, \mathbf{N})$ together with a bootstrap estimate of the standard deviation for different values of n and various

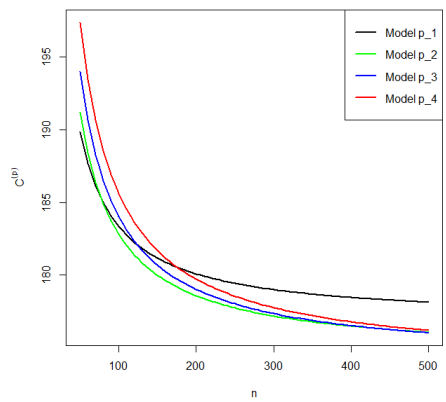


Fig 7: A plot of $\mathbb{C}^{(p)}(n, \mathbf{N})$ for $p = p_1, p_2, p_3, p_4$ and $n = 50, 55, \ldots, 500$.

pairs of models. Also the differences of the corresponding cross-validation estimates are given. The standard deviations of Table 4 are much smaller than those of Table 3. This is consistent with our theoretical results that comparison of two similar models leads to a small estimation error (see the discussion after Theorem 2.1). The table shows which pairs $\mathcal{P}$, $\mathcal{Q}$ differ significantly, and for which values of n.

TABLE 3
$\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N})$ *for* $\mathcal{P} = \mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4$ *and for* $n = 50, 150, 500$. *Bootstrap standard deviations (SD) and cross-validation (CV) estimates are also provided.*

	Model $\mathcal{P}_1$		Model $\mathcal{P}_2$		Model $\mathcal{P}_3$		Model $\mathcal{P}_4$	
n	$\mathbb{C}^{(\mathcal{P}_1)}$ (SD)	CV	$\mathbb{C}^{(\mathcal{P}_2)}$ (SD)	CV	$\mathbb{C}^{(\mathcal{P}_3)}$ (SD)	CV	$\mathbb{C}^{(\mathcal{P}_4)}$ (SD)	CV
50	189.9 (3.4)	183.9	191.2 (3.5)	185.9	194.0 (3.5)	190.1	197.4 (3.8)	194.5
150	181.1 (3.3)	181.1	180.0 (3.3)	180.4	180.7 (3.4)	181.3	181.7 (3.4)	183.0
500	178.1 (3.3)	183.0	176.1 (3.1)	181.3	176.0 (3.4)	181.2	176.2 (3.3)	181.3

TABLE 4
The values of $\mathbb{C}^{(\mathcal{P})}(n, \mathbf{N}) - \mathbb{C}^{(\mathcal{Q})}(n, \mathbf{N})$ *for* $n = 50, 150, 500$. *Bootstrap standard deviations (SD) and cross-validation (CV) estimates are also provided. Boldface numbers indicate differences that are significantly (more than two SD's) non-zero.*

		$n = 50$			$n = 150$			$n = 500$		
$\mathcal{P}$	$\mathcal{Q}$	$\mathbb{C}^{(\mathcal{P})} - \mathbb{C}^{(\mathcal{Q})}$ (SD)		CV	$\mathbb{C}^{(\mathcal{P})} - \mathbb{C}^{(\mathcal{Q})}$ (SD)		CV	$\mathbb{C}^{(\mathcal{P})} - \mathbb{C}^{(\mathcal{Q})}$ (SD)		CV
$\mathcal{P}_1$	$\mathcal{P}_2$	**-1.3**	(0.3)	-2.0	**1.2**	(0.3)	0.7	**2.1**	(0.3)	1.7
$\mathcal{P}_1$	$\mathcal{P}_3$	**-4.1**	(0.4)	-6.2	0.5	(0.4)	-0.2	**2.1**	(0.4)	1.8
$\mathcal{P}_1$	$\mathcal{P}_4$	**-7.5**	(0.5)	-10.6	-0.5	(0.5)	-1.9	**1.9**	(0.5)	1.7
$\mathcal{P}_2$	$\mathcal{P}_3$	**-2.7**	(0.2)	-4.2	**-0.7**	(0.2)	-0.9	0.0	(0.2)	0.1
$\mathcal{P}_2$	$\mathcal{P}_4$	**-6.2**	(0.4)	-8.6	**-1.7**	(0.3)	-2.6	-0.1	(0.3)	0.0
$\mathcal{P}_3$	$\mathcal{P}_4$	**-3.4**	(0.3)	-4.4	**-1.0**	(0.2)	-1.7	-0.2	(0.2)	-0.1

6.4. Comparisons to other approaches

As in Section 5.3 we compare our method to other approaches. One possibility is to concatenate the whole training sample and add a categorical variable for the doctors. The (10-fold) cross-validation estimate of the prediction error of the OLS is 204.5. The corresponding numbers for the ridge and lasso estimates (applied to the concatenated data) are similar: 205.0 and 204.8. The estimates of prediction errors of our method are smaller: they vary between 190 and 176 for $50 \leq n \leq 500$ (See Figure 7 and Table 3).

A different approach is to preform a separate model selection for each of the $\mathcal{J}$ datasets. As in Section 5.3, three selection criteria are considered, $\mathbb{C}^{(\mathcal{P})}(N_j, N_j)$ (the bias-corrected $C^{(\mathcal{P})}(n, N)$ with $n = N_j$), Mallows' C_p and BIC. Figure 8 plots the cross-validation estimates of the prediction errors of the selected models by the three criteria as a function of the sample size $n = N_j$. A normal-kernel smoothing is drawn to illustrate the average prediction error as a function of the sample size. Also plotted is the prediction error of the common model selection $\mathcal{P}_2$ (which is close to optimal for sample sizes between 50 and 500) as estimated by $\mathbf{C}^{(\mathcal{P}_2)}$. Table 3 shows that the latter estimate is rather close to its cross validation estimate. Figure 8 shows that: a. the differences between the three selection criteria are small; b. a common model selection by $\mathbf{C}^{(\mathcal{P})}$ is better on average than a separate model selection; c. the latter statement is especially true for small sample sizes where borrowing strength is more important.

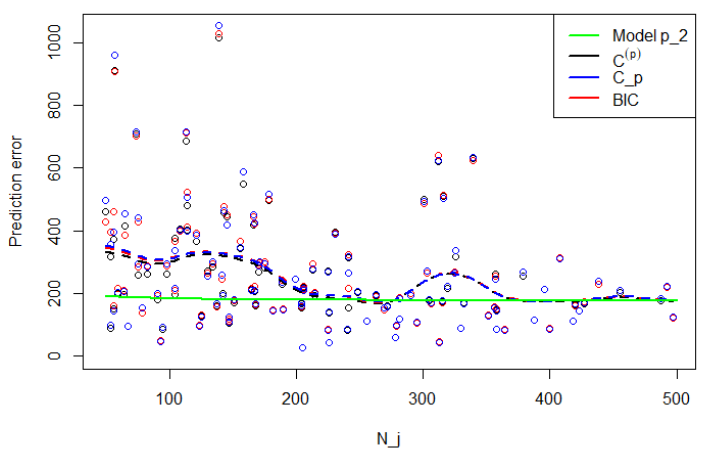


Fig 8: Prediction errors (estimated by cross-validation) of the model selection methods $\mathbb{C}^{(p)}(N_j, N_j)$, Mallows' C_p and BIC applied to each dataset separately compared to the common model $\mathcal{P}_2$. The dashed lines are Gaussian-kernel smoothing. The green line is $\mathbf{C}^{(\mathcal{P}_2)}(n, \mathbf{N})$ as a function of $n \in [50, 500]$.

Acknowledgment: We are grateful to Avishai Mandelbaum for providing access to the SEE Lab dataset we analyzed, and to Ella Nadjharov for creating the files we needed. We also wish to thank the associate editor and the two referees for their very useful comments.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Andrade, D., Okajima, Y. (2020). Adaptive Covariate Acquisition for Minimizing Total Cost of Classification. *Machine Learning (2021)* **110**, 1067–1104

[3] Anderson, D., Burnham, K. (2002). *Model selection and multi-model inference*. New York: Springer..

[4] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[5] Brown, L. D. (2016). Mallows C_p for out-of-sample prediction. http://www-stat.wharton.upenn.edu/~lbrown/Papers/2016f%20Mallows.pdf

[6] Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhao, L., Zhang, K. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, **34**, 523–544.

[7] Çinlar, E. (2011) *Probability and Stochastics*. New York: Springer.

[8] Claeskens, G., Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, **98**, 900-916.

[9] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[10] Erev, I., Roth, A. E., Slonim, R. L., Barron, G. (2007). Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory*, **33**, 29–51.

[11] Groves, T., Rothenberg, T. (1969). A note on the expected value of an inverse matrix. Biometrika, **56**, 690–691.

[12] Hansen, B. E. (2021). *Econometrics*. https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf

[13] Horn, R. A., Johnson, C. R. (2013). *Matrix Analysis* 2nd Ed. Cambridge University Press.

[14] Lindsay, B., Liu, J. (2009). Model assessment tools for a model false world. *Statistical Science*, **24**, 303–318.

[15] Mallows, C. L. (1973). Some comments on CP. *Technometrics*, **15**, 661–675.

[16] Obozinski, G., Taskar, B., Jordan, M. I. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, **20**/2, 231–252.

[17] Nevo, D., Ritov, Y. (2017) Identifying a minimal class of models for high-dimensional data, *Journal of Machine Learning Research*, **18**, 1–29.

[18] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, **12**, 758–765.

[19] von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, **15**, 97–109.

[20] Rosset, S., Tibshirani, R. (2020). From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, **115**, 138–151.
[21] Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: A GLM Approach*. London: Sage.
[22] Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, **69**, 682–689
[23] Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
[24] Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221–264.
[25] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, **12**, 389–434.
[26] Vonesh, E., Chinchilli, V. M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. CRC Press.
[27] Wagenmakers, E. J., Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic bulletin & review*, **11**, 192–196.
[28] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.
[29] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950.
[30] Zacks, S. (1985). Pitman efficiency. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson (eds). New York: Wiley & Sons.
[31] Zhang, Y., Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

7. Appendix A: Proofs

Recall that we use $\mathcal{P}$ to denote a subset of the covariates, to which we sometimes refer as a model, and denote its size with the corresponding letter p. We suppress it in most of our notation below and instead of $\mathbf{X}^{(\mathcal{P})}$ we write $\mathbf{X}$ and assume it is in $\mathbb{R}^p$.

Proof of Theorem 2.1. We first prove (2.9). For $\widehat{\boldsymbol{\beta}}_n$ computed from a sample $D = \{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$, and a pair of new observations from the same distribution $(\mathbf{X}, Y)$, independent of D, we have

$$\begin{aligned} E(Y - \mathbf{X}'\widehat{\boldsymbol{\beta}}_n)^2 &= E(Y - \mathbf{X}'\boldsymbol{\beta})^2 + E[\mathbf{X}'(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})]^2 - 2E[(Y - \mathbf{X}'\boldsymbol{\beta})\mathbf{X}'(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})] \\ &= E(Y - \mathbf{X}'\boldsymbol{\beta})^2 + E[\mathbf{X}'(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})]^2, \quad (7.1) \end{aligned}$$

where the last term in the first line of (7.1) vanishes since $E[(Y - \mathbf{X}'\boldsymbol{\beta})\mathbf{X}] = E[e\mathbf{X}] = \mathbf{0}$ and Y and $\mathbf{X}$ are independent of $\widehat{\boldsymbol{\beta}}_n$. This argument holds also if we condition on H_n (see (2.14), and Theorem 2.3). By (7.1) we have that

$$n[R(n, \mathcal{P}) - AR(n, \mathcal{P})] = E[\mathbf{X}'\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})]^2 - tr(\mathbb{V}). \quad (7.2)$$

Using independence of $\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}_n$ again we have

$$E[\mathbf{X}'\sqrt{n}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})]^2 = tr\{E[\mathbf{X}\mathbf{X}']E[n(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})']\} = tr\{\mathbb{Q}E[n(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})']\}.$$

For the proof of Theorem 2.3 the expectations should be conditioned on the set H_n, whose probability is large, and the conditioning does not affect the rates we obtain.

By Equation (7.3) of Hansen [12],

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \widehat{\mathbb{Q}}_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i e_i = \widehat{\mathbb{Q}}_n^{-1}\mathbf{U}_n.$$

Since $E\{tr(\mathbf{U}_n\mathbf{U}_n'\mathbb{Q}^{-1})\} = tr(\mathbb{V})$ we rewrite the right-hand side of (7.2) as

$$\begin{aligned} E\{tr(\mathbb{Q}\widehat{\mathbb{Q}}_n^{-1}\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1}) - tr(\mathbb{Q}\mathbb{Q}^{-1}\mathbf{U}_n\mathbf{U}_n'\mathbb{Q}^{-1})\} \\ = tr(\mathbb{Q}E\{\widehat{\mathbb{Q}}_n^{-1}\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1}\mathbf{U}_n\mathbf{U}_n'\mathbb{Q}^{-1}\}). \end{aligned} \tag{7.3}$$

In order to prove (2.9) we now show that the latter expectation is of order $O(1/n^{1/2})$. To this end, notice that

$$\begin{aligned} &\widehat{\mathbb{Q}}_n^{-1}\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1}\mathbf{U}_n\mathbf{U}_n'\mathbb{Q}^{-1} \\ &= (\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1})\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1} + \mathbb{Q}^{-1}\mathbf{U}_n\mathbf{U}_n'(\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1}). \end{aligned}$$

We now deal with the first term on the right-hand side above, the other term being similar, and simpler. We have

$$\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1} = \mathbb{Q}^{-1}(\mathbb{Q} - \widehat{\mathbb{Q}}_n)\widehat{\mathbb{Q}}_n^{-1} \tag{7.4}$$

and therefore (recall (7.3)) we consider the expectation of

$$tr\Big(\mathbb{Q}(\widehat{\mathbb{Q}}_n^{-1} - \mathbb{Q}^{-1})\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1}\Big) = tr\Big((\mathbb{Q} - \widehat{\mathbb{Q}}_n)\widehat{\mathbb{Q}}_n^{-1}\mathbf{U}_n\mathbf{U}_n'\widehat{\mathbb{Q}}_n^{-1}\Big).$$

This matrix is a product of random matrices of the form $ABCD$ where $A = \mathbb{Q} - \widehat{\mathbb{Q}}_n$, $B = D = \widehat{\mathbb{Q}}_n^{-1}$, and $C = \mathbf{U}_n\mathbf{U}_n'$. The trace is a sum of products of entries from all the matrices appearing in the product. Different choices of powers can be made, but we use Hölder's inequality in the form $E|abcd| \leq (Ea^{12})^{1/12}(E|b|^3)^{1/3}(Ec^4)^{1/4}(E|d|^3)^{1/3}$ for simplicity. Here a is an entry from A, b an entry from from B, etc., and the triangle inequality can then be used to bound the sum comprising the trace.

For each element j, k of $\mathbb{Q} - \widehat{\mathbb{Q}}_n$ we have

$$E(\mathbb{Q} - \widehat{\mathbb{Q}}_n)_{j,k}^{12} = E\Big(\frac{1}{n}\sum_{i=1}^n [E(X_jX_k) - X_{ij}X_{ik}]\Big)^{12}.$$

The summands $E(X_jX_k) - X_{ij}X_{ik}$ have zero expectation; expanding $(\mathbb{Q} - \widehat{\mathbb{Q}}_n)_{j,k}^{12}$ we see that the number of nonvanishing terms when the expectation is taken

is of order n^6, and all these terms are bounded by our assumptions. Therefore, $[E(\mathbb{Q}-\widehat{\mathbb{Q}}_n)^{12}_{j,k}]^{1/12}$ is of order $1/\sqrt{n}$. (Actually, it is easy to see that 24 bounded moments suffice for this argument, and also for bounding the remaining terms, and 24 can be somewhat reduced by a better but more cumbersome choice of the powers in Hölder's inequality.) A similar computation for the matrix C shows that the required moments of its entries are bounded. The rest of the terms are bounded by our assumptions. Now (2.9) follows.

The proof required the bounded third power of $B = \widehat{\mathbb{Q}}_n^{-1}$, which means that with a mixture of normals we need $n > p + 5$. See the Proof of Lemma 2.2.

We now show (2.10). The definitions of $AR(n, \mathcal{P})$ and $C^{(\mathcal{P})}(n, N)$ imply that

$$\begin{aligned} &AR(n,\mathcal{P}) - C^{(\mathcal{P})}(n,N) \\ &= E(Y - \mathbf{X}'\boldsymbol{\beta})^2 - \frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 + \frac{1}{n}\left\{tr(\mathbb{V}) - tr(\widehat{\mathbb{V}}_N)\right\} - \frac{tr(\widehat{\mathbb{V}}_N)}{N}. \end{aligned} \tag{7.5}$$

Starting with the second term on the right-hand side of (7.5), we have

$$||\mathbf{Y}_N - \mathbb{X}_N\boldsymbol{\beta}||^2 = ||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 + ||\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})||^2 - 2(\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N)'(\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})),$$

where the last term vanishes since $\mathbb{X}'_N(\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N) = 0$. Hence,

$$\begin{aligned} &E(Y - \mathbf{X}'\boldsymbol{\beta})^2 - \frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\widehat{\boldsymbol{\beta}}_N||^2 \\ &= E(Y - \mathbf{X}'\boldsymbol{\beta})^2 - \frac{1}{N}||\mathbf{Y}_N - \mathbb{X}_N\boldsymbol{\beta}||^2 + \frac{1}{N}||\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})||^2. \end{aligned} \tag{7.6}$$

Recall the notation $\mathbf{U}_N = \mathbb{X}'_N\mathbf{e}_N/\sqrt{N}$. Since by Equation (7.3) of Hansen [12] $N^{-1/2}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) = N^{-1/2}(\mathbb{X}'_N\mathbb{X}_N)^{-1}\mathbb{X}'_N\mathbf{e}_N = (\mathbb{X}'_N\mathbb{X}_N)^{-1}\mathbf{U}_N$, we have

$$\begin{aligned} \frac{1}{N}||\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})||^2 &= \frac{1}{N}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})'\mathbb{X}'_N\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) = \mathbf{U}'_N(\mathbb{X}'_N\mathbb{X}_N)^{-1}\mathbf{U}_N \\ &= \frac{1}{N}tr\{\mathbf{U}_N\mathbf{U}'_N(\mathbb{X}'_N\mathbb{X}_N/N)^{-1}\} = \frac{1}{N}tr\{\mathbf{U}_N\mathbf{U}'_N\mathbb{Q}^{-1}\} + o_p(1/N), \end{aligned}$$

where the last equality holds true since $(\mathbb{X}'_N\mathbb{X}_N/N)^{-1} - \mathbb{Q}^{-1} = o_p(1)$. Also,

$$\begin{aligned} \frac{1}{N}||\mathbb{X}_N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})||^2 &= \frac{1}{N}\Big[tr(\mathbf{U}_N\mathbf{U}'_N\mathbb{Q}^{-1}) - tr(\mathbb{W}\mathbb{Q}^{-1}) + tr(\mathbb{W}\mathbb{Q}^{-1})\Big] + o_p(1/N) \\ &= \frac{1}{N}\Big[tr(\mathbf{U}_N\mathbf{U}'_N\mathbb{Q}^{-1}) - tr(\mathbb{V}) + tr(\widehat{\mathbb{V}}_N)\Big] + o_p(1/N), \end{aligned} \tag{7.7}$$

where for the last equality it suffices that $\widehat{\mathbb{Q}}_N$ and $\widehat{\mathbb{W}}_N$ are consistent estimates of $\mathbb{Q}$ and $\mathbb{W}$, and therefore $tr(\widehat{\mathbb{V}}_N)$ is consistent for $tr(\mathbb{V}) = tr(\mathbb{W}\mathbb{Q}^{-1})$. Equations (7.5), (7.6), and (7.7) imply (2.10).

Next we show that $\sqrt{N}\{tr(\widehat{\mathbb{V}}_N) - tr(\mathbb{V})\}$ is asymptotically normal starting with the asymptotic normality of $\sqrt{N}(\widehat{\mathbb{W}}_N - \mathbb{W})$. We have,

$$\begin{aligned}
\widehat{\mathbb{W}}_N &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\widehat{e}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'(Y_i - \mathbf{X}_i'\widehat{\boldsymbol{\beta}}_N)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\big(Y_i - \mathbf{X}_i'\boldsymbol{\beta} - \{\mathbf{X}_i'\widehat{\boldsymbol{\beta}}_N - \mathbf{X}_i'\boldsymbol{\beta}\}\big)^2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\big(Y_i - \mathbf{X}_i'\boldsymbol{\beta}\big)^2 \\
&- \frac{2}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\big(Y_i - \mathbf{X}_i'\boldsymbol{\beta}\big)\left(\mathbf{X}_i'(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})\right) + \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'(\mathbf{X}_i'\{\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}\})^2 \\
&=: A - B + C, \qquad (7.8)
\end{aligned}$$

respectively. Starting with the first term we have

$$\sqrt{N}(A - \mathbb{W}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{n}\left(\mathbf{X}_i\mathbf{X}_i'e_i^2 - E[\mathbf{X}\mathbf{X}'e^2]\right),$$

which is asymptotically normal. Now B is obtained by multiplying the sum $\frac{2}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\big(Y_i - \mathbf{X}_i'\boldsymbol{\beta}\big)\otimes\mathbf{X}_i'$ (which converges to a matrix of constants by the law of large numbers) by $I_p\otimes(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})$, where I_p is the identity matrix of order p, and $\otimes$ is Kronecker's product. By Equation (7.3) of Hansen [12],

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}) = \widehat{\mathbb{Q}}_N^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{X}_i e_i.$$

By the law of large numbers and the fact that $N(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})_j(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})_k = O_p(1)$, we have that the term C in (7.8) is $O_p(\frac{1}{N})$. Summing up,

$$\begin{aligned}
&\sqrt{N}(\widehat{\mathbb{W}}_N - \mathbb{W}) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{n}\left(\mathbf{X}_i\mathbf{X}_i'e_i^2 - E[\mathbf{X}\mathbf{X}'e^2]\right) - \mathbb{B}_N[I_p\otimes(\widehat{\mathbb{Q}}_N^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{n}\mathbf{X}_i e_i)] + O_p(1/N),
\end{aligned}$$

where $\mathbb{B}_N$ is the matrix $\frac{2}{N}\sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i'\big(Y_i - \mathbf{X}_i'\boldsymbol{\beta}\big)\otimes\mathbf{X}_i'$. Condition (i) implies that the second moment of $\mathbf{X}_i\mathbf{X}_i'e_i^2$ is finite. Hence, by the central limit theorem,

$$\frac{1}{\sqrt{N}}\Big(\sum_{i=1}^{n}\left\{\mathbf{X}_i\mathbf{X}_i'e_i^2 - E[\mathbf{X}\mathbf{X}'e^2]\right\}, \sum_{i=1}^{n}\mathbf{X}_i e_i\Big) \qquad (7.9)$$

is jointly asymptotically normal, and since $\mathbb{B}_N$ and $\widehat{\mathbb{Q}}_N$ converge to a matrix of constants, a version of Slutsky's theorem implies that $\sqrt{N}(\widehat{\mathbb{W}}_N - \mathbb{W})$ is asymptotically normal.

Another application of Slutsky's theorem implies that $\sqrt{N}(\widehat{\mathbb{W}}_N\widehat{\mathbb{Q}}_N^{-1} - \mathbb{W}\mathbb{Q}^{-1})$ is asymptotically normal (since $\widehat{\mathbb{Q}}_N \to \mathbb{Q}$ in probability) and therefore so is

$$\sqrt{N}\left\{tr\left(\widehat{\mathbb{W}}_N\widehat{\mathbb{Q}}_N^{-1}\right) - tr(\mathbb{W}\mathbb{Q}^{-1})\right\} = \sqrt{N}\left\{tr(\widehat{\mathbb{V}}_N) - tr(\mathbb{V})\right\}.$$

It follows that $tr(\widehat{\mathbb{V}}_N) - tr(\mathbb{V}) = O_p(1/\sqrt{N})$ and it is asymptotically normal.

Similar to previous arguments, the random variables in (7.9) and the first part of $\mathcal{E}_{\mathcal{N}}$ are jointly asymptotically normal and a version of Slutsky's theorem that allows us to ignore the term $O_p(1/N)$ together with (2.10) and (2.11), implies the last statement of Theorem 2.1 about the asymptotic normality of $\sqrt{N}\big(C^{(\mathcal{P})}(n,N) - AR(n,\mathcal{P})\big)$. □

Proof of Lemma 2.2. First assume that the first coordinate of the covariate vectors is 1 (corresponding to an intercept coefficient). Let $\widetilde{\mathbb{X}}$ denote the $n \times (p-1)$ matrix defined as $\mathbb{X}_n$ but without the first column of 1's. We suppress n here and in the following notation. Let $\overline{\mathbf{X}} := \widetilde{\mathbb{X}}'\mathbf{1}/n$, where $\mathbf{1}$ is an n-column vector of 1's, that is, $\overline{\mathbf{X}}$ is the $(p-1)$-column of covariates means, and let $\mathbb{S} := \widetilde{\mathbb{X}}'\widetilde{\mathbb{X}}/n - \overline{\mathbf{X}}\,\overline{\mathbf{X}}'$. Then by Horn and Johnson [13], page 25, Equation (0.8.5.6)

$$(\mathbb{X}'\mathbb{X}/n)^{-1} = \begin{bmatrix} 1 + \overline{\mathbf{X}}'\mathbb{S}^{-1}\overline{\mathbf{X}} & -\overline{\mathbf{X}}'\mathbb{S}^{-1} \\ -\mathbb{S}^{-1}\overline{\mathbf{X}} & \mathbb{S}^{-1} \end{bmatrix}. \tag{7.10}$$

Let $\widetilde{\mathbf{X}}$ denote a $(p-1)$-vector of the covariates without the first 1, and assume first that $\widetilde{\mathbf{X}} \sim N(\boldsymbol{\mu}, \Sigma)$. Then $n\mathbb{S} \sim Wishart_{p-1}(\Sigma, n-1)$, and $\overline{\mathbf{X}}$ and $\mathbb{S}$ are independent. The third moments of $\mathbb{S}^{-1}$ are uniformly (in n) bounded by $C\max\{1, 1/\lambda_{min}(\Sigma)^3\}$ for some $C > 0$, provided $n - p - 5 > 0$ by Theorem 4.1 of von Rosen [19]. By (7.10) the third moments of $(\mathbb{X}'\mathbb{X}/n)^{-1}$ are also bounded. If $\widetilde{\mathbf{X}}$ is distributed according to a mixture of normals, the assumption in the lemma that all these normal distributions have $\lambda_{min}(\Sigma) > c > 0$ implies the uniform boundedness for the mixture.

If the first coordinate is not 1, we append 1 to the covariates and now the matrix of interest in the above notation (with p replacing $p-1$) is $\widetilde{\mathbb{X}}'\widetilde{\mathbb{X}}/n$, which is now $p \times p$, and we wish to show that its inverse has bounded third moments. The eigenvalues of the latter matrix are larger (not strictly) than those of $\mathbb{S} = \widetilde{\mathbb{X}}'\widetilde{\mathbb{X}}/n - \overline{\mathbf{X}}\,\overline{\mathbf{X}}'$. The latter relation is revered for the inverses. Now use the inequality that for a positive definite matrix A we have $|a_{ij}| \le tr(A)/2$ to conclude that the entries of $(\widetilde{\mathbb{X}}'\widetilde{\mathbb{X}}/n)^{-1}$ are bounded by $tr(\mathbb{S}^{-1})$. By the first part of the theorem (with p replacing $p-1$) we know that $\mathbb{S}^{-1}$ has finite third moments and therefore also its trace (by Minkowski inequality), and the result follows. □

Lemma 2.3. *For $\mathbb{X}_n \in H_n$ (see (2.14)) the matrix $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ exists and all its entries are bounded uniformly in n. Moreover, if the components of $\mathbf{X}$ are bounded, then for some $a < 1$ we have, $P(H_n) > 1 - a^{n\lambda_{min}(\mathbb{Q})}$, which converge to 1 at an exponential rate in n.*

Proof. When $\mathbb{X}_n \in H_n$, then $\lambda_{min}(\mathbb{X}_n'\mathbb{X}_n/n) \ge \frac{1}{2}\lambda_{min}(\mathbb{Q}) > 0$, and therefore $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ exists. Since the entries of a positive semi-definite matrix are bounded by the maximal eigenvalue, all entries of $(\mathbb{X}_n'\mathbb{X}_n/n)^{-1}$ are bounded in this case by $2/\lambda_{min}(\mathbb{Q})$. For the moreover part, notice that when the elements of $\mathbf{X}$ are bounded then so is $\lambda_{max}(\mathbb{X}_n'\mathbb{X}_n/n)$. By Tropp [25], Theorem 1.1, $P(H_n^c) \le a^{n\lambda_{min}(\mathbb{Q})}$ for some $a < 1$. □

Proof of Theorem 2.3. Conditionally on H_n the arguments in the proof of Theorem 2.1 continue to apply with obvious modifications. Lemma 2.3 provides an exponentially small bound on $P(H_n^c)$, adding a term $O_p(e^{-\gamma n})$ in Equation (2.9) and $O_p(e^{-\gamma N})$ to (2.10) and (2.12) (b) for some $\gamma > 0$. Clearly, these terms do not affect the results. □

Proof of Proposition 2.4. For Part (i), set $\mathcal{P}(n) \in \pi^*(n)$ and $\widetilde{\mathcal{P}}(n) \in \mathcal{P}^*(n)$. We have $R(n, \mathcal{P}(n)) - R(n, \widetilde{\mathcal{P}}(n)) \geq 0$, and also $R(n, \mathcal{P}(n)) - R(n, \widetilde{\mathcal{P}}(n)) = [R(n, \mathcal{P}(n)) - AR(n, \mathcal{P}(n))] + [AR(n, \mathcal{P}(n)) - AR(n, \widetilde{\mathcal{P}}(n))] + [AR(n, \widetilde{\mathcal{P}}(n)) - R(n, \widetilde{\mathcal{P}}(n))]$. The middle term is negative and by (2.9) the two other terms are $o(1/n)$ and Part (i) follows.

For Part (ii), we first show that $\mathcal{P}^*$ is a singleton. Suppose that there are two models $\mathcal{P}, \mathcal{Q}$ in $\mathcal{P}^*$. By the definition of $\mathcal{P}^*$, the components of the projection coefficient vectors $\boldsymbol{\beta}^{(\mathcal{P})}$ and $\boldsymbol{\beta}^{(\mathcal{Q})}$ must all be non-zero. Since the function $(Y-a)^2$ is strictly convex in a, we have

$$\left(Y - \frac{\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} + \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}}{2}\right)^2 \leq \frac{\left(Y - \mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})}\right)^2 + \left(Y - \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}\right)^2}{2}, \tag{7.11}$$

with equality if and only if $\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}$. Unless $\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{P})}$ a.s., the model $\mathcal{P} \cup \mathcal{Q}$ would contradict the assumption that $\mathcal{P}$ and $\mathcal{Q}$ are in $\mathcal{M}$ by taking expectations in (7.11). We assumed that $E(\mathbf{X}\mathbf{X}')$ is invertible and hence $\mathbf{X}'\boldsymbol{\beta}$ vanishes a.s. only for $\boldsymbol{\beta} = \mathbf{0}$. Adding zeros to $\boldsymbol{\beta}^{(\mathcal{P})}$ and $\boldsymbol{\beta}^{(\mathcal{Q})}$, thus completing them to vectors in $\mathbb{R}^d$, we see that the completed vectors are identical. Since the two models are in $\mathcal{M}$ and their projection coefficients are all non-zero, it follows that $\mathcal{P} = \mathcal{Q}$, and therefore $\mathcal{P}^*$ is a singleton. The above discussion also shows that any model $\mathcal{Q} \in \mathcal{M}$ must satisfy, as sets of covariates, $\mathcal{Q} \supseteq \mathcal{P}^*$.

In order to show that $\pi^*(n) = \mathcal{P}^*$ for large n, note first that $\mathcal{Q} \in \pi^*(n)$ for large enough n implies $\mathcal{Q} \in \mathcal{M}$; if not then there is some $\widetilde{\mathcal{Q}} \in \mathcal{M}$ such that $E\big(Y - \mathbf{X}^{(\widetilde{\mathcal{Q}})\prime}\boldsymbol{\beta}^{(\widetilde{\mathcal{Q}})}\big)^2 < E\big(Y - \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}\big)^2$. For large n this $\widetilde{q}$ contradicts $\mathcal{Q} \in \pi^*(n)$. It follows that $\mathcal{Q} \supseteq \mathcal{P}^*$ as sets of covariates, and it suffices to show that $tr(\mathbb{V}^{(\mathcal{P})})$ is minimized over $\mathcal{M}$ by $\mathcal{P}^*$. Indeed we show that if $\mathcal{P}, \mathcal{Q} \in \mathcal{M}$ and $\mathcal{Q} \supsetneqq \mathcal{P}$ as sets of covariates, then $tr(\mathbb{V}^{(\mathcal{P})}) < tr(\mathbb{V}^{(\mathcal{Q})})$. Consider a Gram–Schmidt process on the space of square integrable random variables, with the inner product of two random variables being the expectation of their product. Starting with the indexes in $\mathcal{P}$, there exist linear transformations $\widetilde{\mathbf{X}}^{(\mathcal{P})} := A\mathbf{X}^{(\mathcal{P})}$ and $\widetilde{\mathbf{X}}^{(\mathcal{Q})} := B\mathbf{X}^{(\mathcal{Q})}$, where A and B are invertible $p \times p$ and $q \times q$ matrices, such that $E\big(\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})\prime}\big)$ and $E\big(\widetilde{\mathbf{X}}^{(\mathcal{Q})}\widetilde{\mathbf{X}}^{(\mathcal{Q})\prime}\big)$ are both identity matrices (with different dimensions). Also, we can assume that the first p rows of B can be obtained from those of A by adding $q-p$ zeros to each of these rows. Therefore, for $k \in \mathcal{P}$ we have $\widetilde{X}_k^{(\mathcal{P})} = \widetilde{X}_k^{(\mathcal{Q})}$, where $\widetilde{X}_k^{(\mathcal{P})}$ is the kth coordinate of $\widetilde{\mathbf{X}}^{(\mathcal{P})}$. The relation $\widetilde{\mathbf{X}}^{(\mathcal{P})} = A\mathbf{X}^{(\mathcal{P})}$ and straightforward algebra, using properties of the trace function, imply that $tr(\mathbb{V}^{(\mathcal{P})}) = tr(\widetilde{\mathbb{V}}^{(\mathcal{P})})$, where

$$\widetilde{\mathbb{V}}^{(\mathcal{P})} := E\left(\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})\prime}\{e^{(\mathcal{P})}\}^2\right)\left\{E\left(\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})\prime}\right)\right\}^{-1} = E\left(\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})\prime}\{e^{(\mathcal{P})}\}^2\right),$$

and similarly for $tr(\mathbb{V}^{(\mathcal{Q})})$. We have $e^{(\mathcal{P})} = e^{(\mathcal{Q})}$ for any $\mathcal{P}, \mathcal{Q} \in \mathcal{M}$ (with probability 1) since otherwise, by the argument in (7.11), $\mathcal{P} \cup \mathcal{Q}$ would contradict the assumption that both $\mathcal{P}$ and $\mathcal{Q}$ are in $\mathcal{M}$ as above. We conclude that,

$$tr(\mathbb{V}^{(\mathcal{P})}) = tr\left\{E\left(\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})'}\{e^{(\mathcal{P})}\}^2\right)\right\} = \sum_{k\in\mathcal{P}} E\left(\widetilde{X}_k^{(\mathcal{P})}e^{(\mathcal{P})}\right)^2$$
$$< \sum_{k\in\mathcal{Q}} E\left(\widetilde{X}_k^{(\mathcal{Q})}e^{(\mathcal{Q})}\right)^2 = tr(\mathbb{V}^{(\mathcal{Q})}). \quad (7.12)$$

The strict inequality follows from the fact that $\mathbb{W}$ is positive definite, and thus $\widetilde{\mathbf{X}}^{(\mathcal{P})}\widetilde{\mathbf{X}}^{(\mathcal{P})'}\{e^{(\mathcal{P})}\}^2 = A\mathbf{X}^{(\mathcal{P})}\mathbf{X}^{(\mathcal{P})'}\{e^{(\mathcal{P})}\}^2A'$ are matrices with positive definite expectations, and therefore positive diagonal elements. Summing up, the above discussion shows that $\mathcal{P}^*$ is a singleton, and that $tr(\mathbb{V}^{(\mathcal{P}^*)})$ is minimal among the models in $\mathcal{M}$. This implies that $\pi^*(n) \to \mathcal{P}^*$ as $n \to \infty$. By (2.9), for every model p, $R(n, \mathcal{P}) - AR(n, \mathcal{P}) = o(1/n)$, and therefore $\mathcal{P}^*(n) = \pi^*(n)$ for large enough n. Hence, also $\mathcal{P}^*(n) \to \mathcal{P}^*$ as $n \to \infty$. □

Proof of Proposition 2.5. It suffices to prove that $P(\widehat{\pi^*}(n, N) = \mathcal{P}^*) \to 1$ when $n, N \to \infty$ and $n/N \to 0$, since by Proposition 2.4, $\mathcal{P}^*(n) = \mathcal{P}^*$ for large enough n. Equivalently, we claim that for every $\mathcal{P} \neq \mathcal{P}^*$ we have $P(\widehat{\pi^*}(n, N) = \mathcal{P}) \to 0$, and since there is a finite number of models, the result follows. The latter claim is proved separately for $\mathcal{P} \notin \mathcal{M}$ and then for $\mathcal{P} \in \mathcal{M}$, (conditions (a) and (b) below):

(a) For $\mathcal{P} \notin \mathcal{M}$ we shall show that

$$C^{(\mathcal{P})}(n, N) - C^{(\mathcal{P}^*)}(n, N) = A - \frac{tr(\mathbb{V}^{(\mathcal{P})}) - tr(\mathbb{V}^{(\mathcal{P}^*)})}{n} + O_p(1/\sqrt{N}), \quad (7.13)$$

for a positive constant A. Since $\widehat{\pi^*}(n, N)$ is the minimizer of $C^{(\mathcal{P})}(n, N)$, it follows that $P(\widehat{\pi^*}(n, N) = \mathcal{P}) \to 0$ as both n, N go to infinity. To prove (7.13) note that by the definition of $AR(n, \mathcal{P})$ and Equations (2.7), (2.10), and (2.12), we have

$$\begin{aligned} C^{(\mathcal{P})}(n, N) - C^{(\mathcal{P}^*)}(n, N) &= AR(n, \mathcal{P}) - AR(n, \mathcal{P}^*) + O_p(1/\sqrt{N}) \\ &= E\big(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(p)}\big)^2 - E\big(Y - \mathbf{X}^{(\mathcal{P}^*)'}\boldsymbol{\beta}^{(\mathcal{P}^*)}\big)^2 \\ &\quad + \frac{tr(\mathbb{V}^{(\mathcal{P})}) - tr(\mathbb{V}^{(\mathcal{P}^*)})}{n} + O_p(1/\sqrt{N}). \end{aligned}$$

Since $\mathcal{P} \notin \mathcal{M}$ and $\mathcal{P}^* \in \mathcal{M}$, the difference of the expectations, which we denote by A, is positive.

(b) For $\mathcal{P} \in \mathcal{M}$ and $\mathcal{P} \neq \mathcal{P}^*$ we shall show that

$$C^{(\mathcal{P})}(n, N) - C^{(\mathcal{P}^*)}(n, N) = B/n + O_p(1/N) + O_p\left(\frac{1}{n\sqrt{N}}\right), \quad (7.14)$$

where B is a positive constant implying that $P(\widehat{\pi^*}(n, N) = \mathcal{P}) \to 0$ when both n, N go to infinity and $n/N \to 0$.

We now prove (7.14). Consider $\mathcal{P} \in \mathcal{M}$ and $\mathcal{P} \neq \mathcal{P}^*$. Since both models are in $\mathcal{M}$, we have

$$E\big(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2 - E\big(Y - \mathbf{X}^{(\mathcal{P}^*)'}\boldsymbol{\beta}^{(\mathcal{P}^*)}\big)^2 = 0$$

and therefore

$$AR(n,\mathcal{P}) - AR(n,\mathcal{P}^*) = \frac{tr(\mathbb{V}^{(\mathcal{P})}) - tr(\mathbb{V}^{(\mathcal{P}^*)})}{n}.$$

In Proposition 2.4 we showed that $tr(\mathbb{V}^{(\mathcal{P}^*)}) < tr(\mathbb{V}^{(\mathcal{P})})$, and therefore $AR(n,\mathcal{P}) - AR(n,\mathcal{P}^*) = B/n$, where B is a positive constant. Since both $\mathcal{P}$ and $\mathcal{P}^*$ are in $\mathcal{M}$, it follows that $\mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\mathcal{P}^*)'}\boldsymbol{\beta}^{(\mathcal{P}^*)}$ a.s. (see the proof of Proposition 2.4). Therefore, the first part of $\mathcal{E}_N^{(\mathcal{P})}$ and $\mathcal{E}_N^{(\mathcal{P}^*)}$ is equal, and hence, $\mathcal{E}_N^{(\mathcal{P})} - \mathcal{E}_N^{(\mathcal{P}^*)} = O_p(1/N)$. Recalling that $AR(n,\mathcal{P}) - AR(n,\mathcal{P}^*) = B/n$, (2.10) implies that

$$\begin{aligned} &C^{(\mathcal{P})}(n,N) - C^{(\mathcal{P}^*)}(n,N) \\ &\qquad = B/n + O_p(1/N) - \frac{tr(\mathbb{V}^{(\mathcal{P})}) - tr(\widehat{\mathbb{V}}_N^{(\mathcal{P})}) - tr(\mathbb{V}^{(\mathcal{P}^*)}) + tr(\widehat{\mathbb{V}}_N^{(\mathcal{P}^*)})}{n}, \end{aligned}$$

which implies (7.14) by (2.12) (b). □

Proof of Theorem 3.1. The first part follows from (2.10) of Theorem 2.1. That the o_p terms do not depend on n can be seen by inspecting the proof of (2.10) of Theorem 2.1. The moreover part follows from the asymptotic normality of each j; see (2.13). □

Proof of Proposition 3.3. Part 1 follows from the first part of Theorem 3.1.

The proof of Part 2 differs from that of Proposition 2.5 only in taking averages over $\mathcal{J}$ in similar expressions. The only real difference is in case (b) of the proof of Proposition 2.5, with $\mathcal{P}$ and $\boldsymbol{\mathcal{P}}^*$ both in $\boldsymbol{\mathcal{M}}$. The proof is achieved by showing that there exists $\mathcal{B} > 0$ such that

$$\lim \sup_{n/N_j \le C, n, \mathbf{N} \to \infty} P\left(n\{\mathbf{C}^{(\mathcal{P})}(n,\mathbf{N}) - \mathbf{C}^{(\boldsymbol{\mathcal{P}}^*)}(n,\mathbf{N})\} < \mathcal{B}/2\right) < K/\mathcal{J}. \quad (7.15)$$

We have

$$n\{\mathbf{C}^{(\mathcal{P})}(n,\mathbf{N}) - \mathbf{C}^{(\boldsymbol{\mathcal{P}}^*)}(n,\mathbf{N})\} = \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} B_j + \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} C_j + \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} D_j + \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} E_j \quad (7.16)$$

where,

$$B_j := tr(\mathbb{V}_j^{(\mathcal{P})}) - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)}), \qquad C_j := n(\mathcal{E}_{j,N_j}^{(\mathcal{P})} - \mathcal{E}_{j,N_j}^{(\boldsymbol{\mathcal{P}}^*)}),$$
$$D_j := tr(\mathbb{V}_j^{(\mathcal{P})}) - tr(\widehat{\mathbb{V}}_{j,N_j}^{(\mathcal{P})}) - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)}) + tr(\widehat{\mathbb{V}}_{j,N_j}^{(\boldsymbol{\mathcal{P}}^*)}),$$

and $E_j = no_p(1/N_j)$, arising from the last term in (2.10). The proof of (7.15) is accomplished by showing that $\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} B_j \ge \mathcal{B}$, to be defined below, and that the other three sums are small.

We start with the first term in (7.16), $\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} B_j = \frac{1}{\mathcal{J}}\sum_j [tr(\mathbb{V}_j^{(\mathcal{P})}) - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)}]$. Since $\mathcal{P}$ is in $\boldsymbol{\mathcal{M}}$ we have $\mathcal{P} \supseteq \boldsymbol{\mathcal{P}}^*$ as sets of covariates. By (7.12), $tr(\mathbb{V}_j^{(\mathcal{P})}) - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)})$ is bounded below by $E_{G_j}\left(\widetilde{X}_k^{(\mathcal{P})} e^{(\mathcal{P})}\right)^2$ for $k \in \mathcal{P} \setminus \boldsymbol{\mathcal{P}}^*$ (as sets). We have $\widetilde{X}_k = \mathbf{b}_k'\mathbf{X}$ where $\mathbf{b}_k'$ is the kth row of the matrix B defined in the proof of Proposition 2.4. We have $1 = E(\mathbf{b}_k'\mathbf{X})^2 = \mathbf{b}_k'\mathbb{Q}_j^{(\mathcal{P})}\mathbf{b}_k$ and therefore $\|\mathbf{b}_k'\{\mathbb{Q}_j^{(\mathcal{P})}\}^{1/2}\| = 1$. It follows that $\|\mathbf{b}_k\|^2 = \mathbf{b}_k'\{\mathbb{Q}_j^{(\mathcal{P})}\}^{1/2}\{\mathbb{Q}_j^{(\mathcal{P})}\}^{-1}\{\mathbb{Q}_j^{(\mathcal{P})}\}^{1/2}\mathbf{b}_k \geq \lambda_{min}(\{\mathbb{Q}_j^{(\mathcal{P})}\}^{-1}) = 1/\lambda_{max}(\mathbb{Q}_j^{(\mathcal{P})})$ and therefore $E_{G_j}\left(\widetilde{X}_k^{(\mathcal{P})} e^{(\mathcal{P})}\right)^2 = E_{G_j}(\mathbf{b}_k'\mathbf{X}^{(\mathcal{P})}e)^2 = \mathbf{b}'\mathbb{W}_j^{(\mathcal{P})}\widetilde{\mathbf{b}}_k \geq \lambda_{min}(\mathbb{W}_j^{(\mathcal{P})})/\lambda_{max}(\mathbb{Q}_j^{(\mathcal{P})}) > 1/C^2 > 0$. We obtained that $\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} B_j \geq 1/C^2 =: \mathcal{B}$.

We now deal with $\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} C_j$. By (2.10) and (2.11) and the fact that $\mathcal{P}, \mathcal{P}^* \in \boldsymbol{\mathcal{M}}$, this term equals,

$$\frac{n}{\mathcal{J}}\Big[\sum_j \frac{1}{N_j}\Big\{tr\left[\mathbf{U}_{j,N_j}^{(\boldsymbol{\mathcal{P}}^*)}\mathbf{U}_{j,N_j}^{(\boldsymbol{\mathcal{P}}^*)'}(\mathbb{Q}_j^{(\boldsymbol{\mathcal{P}}^*)})^{-1}\right] - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)}) \\ - tr\left[\mathbf{U}_{j,N_j}^{(p)}\mathbf{U}_{j,N_j}^{(\mathcal{P})'}(\mathbb{Q}_j^{(\mathcal{P})})^{-1}\right] + tr(\mathbb{V}_j^{(\mathcal{P})})\Big\}\Big]. \quad (7.17)$$

Since $\mathbf{U}_{j,N_j}^{(\mathcal{P})}$ converges in distribution to $\mathbf{Z}_j^{(\mathcal{P})} \sim N(0, \mathbb{W}_j^{(\mathcal{P})})$ and C is an upper bound on n/N_j we have that the limit of the probability that the expression in (7.17) exceeds ε is bounded by $P(T_{\mathcal{J}} > \varepsilon)$ where

$$T_{\mathcal{J}} := \frac{C}{\mathcal{J}}\Big|\sum_j \Big\{tr\left[\mathbf{Z}_j^{(\boldsymbol{\mathcal{P}}^*)}\mathbf{Z}_j^{(\boldsymbol{\mathcal{P}}^*)'}(\mathbb{Q}_j^{(\boldsymbol{\mathcal{P}}^*)})^{-1}\right] - tr(\mathbb{V}_j^{(\boldsymbol{\mathcal{P}}^*)}) \\ - tr\left[\mathbf{Z}_j^{(\mathcal{P})}\mathbf{Z}_j^{(\mathcal{P})'}(\mathbb{Q}_j^{(\mathcal{P})})^{-1}\right] + tr(\mathbb{V}_j^{(\mathcal{P})})\Big\}\Big|.$$

Note that the expression within the absolute value sign has mean zero. Writing $T_{\mathcal{J}} = \frac{C}{\mathcal{J}}|\sum_{j=1}^{\mathcal{J}} A_j|$, Markov's inequality implies that in order to obtain $P(T_{\mathcal{J}} > \varepsilon) \leq K/\mathcal{J}$ it is enough to bound $Var(A_j)$ uniformly in j, which holds when $(\mathbb{Q}_j^{(\mathcal{P})})^{-1}, \mathbb{W}_j^{(\mathcal{P})}$ are bounded (element-wise) for all models $\mathcal{P}$ (of which there is a finite number) and uniformly for all j. This follows from our eigenvalue assumptions (see (3.7)) and the fact that the entries of a positive-definite matrix are bounded by its maximal eigenvalue. Finally, it suffices to show that $D_j \to 0$ and $E_j \to 0$ as $n, N_j \to \infty$ with n/N_j bounded. The first follows from (2.12), and the second is obvious. □

Proof of Lemma 3.5. First notice that when the moments appearing in (i) of Theorem 2.1 are bounded uniformly in $\theta \in \Theta$, then $E_{G_\theta}(Y - \mathbf{X}'\boldsymbol{\beta}_\theta)^2$ is bounded in θ. Also, the matrix $\mathbb{W}_\theta$ is bounded (element-wise) uniformly in θ. Finally, because $E\{(\mathbb{X}_n{}'\mathbb{X}_n/n)^{-1}\} - \mathbb{Q}_\theta^{-1}$ is positive semi-definite (see Groves and Rothenberg [11]), then uniform boundedness of the moment condition (ii) of Theorem 2.1 implies that $\mathbb{Q}_\theta^{-1}$ is uniformly bounded and therefore so is $tr(\mathbb{V}_\theta) = tr(\mathbb{W}_\theta\mathbb{Q}_\theta^{-1})$.

We have

$$\begin{aligned}\mathbf{AR}_{pop}(n,\mathcal{P}) - \mathbf{AR}(n,\mathcal{P}) \\ = \Big\{ \int E_{G_\theta}(Y - \mathbf{X}'\boldsymbol{\beta}_\theta)^2 \mathscr{P}(d\theta) - \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}'\boldsymbol{\beta}_j)^2 \Big\} \\ + \frac{1}{n}\Big\{ \int tr(\mathbb{V}_\theta)\mathscr{P}(d\theta) - \frac{1}{\mathcal{J}} \sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}_j) \Big\}. \quad (7.18)\end{aligned}$$

The above two sums contain random variables that are bounded, and hence so are their variances. The central limit theorem applied twice, implies (3.10) and the claimed asymptotic normality. It is easy to see directly from (7.18) that the O_p term in (3.10) is uniform in n. □

For the proof of Proposition 3.6 we need the following lemma:

Lemma 7.1. *Suppose that the conditions of Lemma 3.4 hold and also that $\lambda_{min}(\mathbb{W}_\theta)$ is bounded away from zero uniformly in θ; then*

1. *The set $\boldsymbol{\mathcal{P}}^*_{pop}$ is a singleton and as $n \to \infty$ both $\boldsymbol{\pi}^*_{pop}(n) \to \boldsymbol{\mathcal{P}}^*_{pop}$ and $\boldsymbol{\mathcal{P}}^*_{pop}(n) \to \boldsymbol{\mathcal{P}}^*_{pop}$, and therefore also $\boldsymbol{\pi}^*_{pop}(n) = \boldsymbol{\mathcal{P}}^*_{pop}(n)$ for large n.*
2. *There exists a constant K_C depending only on C, such that for $\boldsymbol{\pi}^*(n)$ defined in (3.6),*

$$P\Big(\boldsymbol{\pi}^*(n) \subseteq \boldsymbol{\pi}^*_{pop}(n)\Big) \geq 1 - \frac{K_C}{\mathcal{J}} \quad \forall n.$$

Proof of Lemma 7.1. Part 1. The proof is similar to that of Proposition 2.4. We sketch the proof. Let $\mathcal{P}$ and $\mathcal{Q}$ be in $\boldsymbol{\mathcal{P}}^*_{pop}$. By convexity as in (7.11),

$$\frac{\big(Y - \mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})}\big)^2 + \big(Y - \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}\big)^2}{2} - \left(Y - \frac{\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} + \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}}{2}\right)^2 \geq 0, \quad (7.19)$$

with equality iff $\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{P})}$. This implies that $\boldsymbol{\mathcal{P}}^*_{pop}$ is a singleton as in the proof of Proposition 2.4, Part (ii). Since $\mathcal{P}$ and $\mathcal{Q}$ are in $\boldsymbol{\mathcal{M}}_{pop}$, the expectation of the left-hand side of (7.19) is zero. It follows that $\int P_{G_\theta}\Big(\mathbf{X}^{(\mathcal{P})\prime}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\mathcal{Q})\prime}\boldsymbol{\beta}^{(\mathcal{Q})}\Big)\mathscr{P}(d\theta) = 1$, and therefore for every model $\mathcal{P}$ in $\boldsymbol{\mathcal{M}}_{pop}$ we have that $\boldsymbol{\mathcal{P}}^*_{pop} \subseteq \mathcal{P}$. By the assumptions on moments being uniformly bounded, it follows that $\lambda_{max}(\mathbb{Q}_\theta)$ is bounded above and $\lambda_{min}(\mathbb{W}_\theta)$ is positive and bounded away from zero, both uniformly in θ. Now (7.12) and the discussion in the paragraph above (7.17) imply that if $\boldsymbol{\mathcal{P}}^*_{pop} \subseteq \mathcal{P}$ as sets of covariates, $\mathcal{P} \in \boldsymbol{\mathcal{M}}_{pop}$, and $\boldsymbol{\mathcal{P}}^*_{pop} \neq \mathcal{P}$ then $\int tr\Big(\mathbb{V}_\theta^{(\boldsymbol{\mathcal{P}}^*_{pop})}\Big)\mathscr{P}(d\theta) < \int tr\left(\mathbb{V}_\theta^{(\mathcal{P})}\right)\mathscr{P}(d\theta)$. Therefore, $\boldsymbol{\mathcal{P}}^*_{pop}$ has a minimal trace among $\boldsymbol{\mathcal{M}}_{pop}$. It follows that $\boldsymbol{\pi}^*_{pop}(n) \to \boldsymbol{\mathcal{P}}^*_{pop}$ as $n \to \infty$. Furthermore, Lemma 3.4 implies that $\boldsymbol{\pi}^*_{pop}(n)$ and $\boldsymbol{\mathcal{P}}^*_{pop}(n)$ coincide for large n. The result now follows from the convergence of $\boldsymbol{\pi}^*_{pop}(n)$ to $\boldsymbol{\mathcal{P}}^*_{pop}$.

Part 2. By Part 1, there exists n_1 such that for every $n \geq n_1$ $\boldsymbol{\pi}^*_{pop}(n) = \boldsymbol{\mathcal{P}}^*_{pop}$, and both are singletons.

For $n > n_1$ the set $\boldsymbol{\mathcal{P}}^*_{pop}$ is a singleton, and we now show that for n sufficiently large

$$P\left(\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\right) \leq \frac{K_C}{\mathcal{J}}. \tag{7.20}$$

We have that

$$P\left(\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\right) \leq \sum_{\mathcal{P} \neq \boldsymbol{\mathcal{P}}^*_{pop}} P\left(\{\mathcal{P} \in \boldsymbol{\pi}^*(n)\} \cap \{\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\}\right).$$

The event $\{\mathcal{P} \in \boldsymbol{\pi}^*(n)\}$ implies that $\mathbf{AR}(n, \mathcal{P}) < \mathbf{AR}(n, \mathcal{Q})$ for every $\mathcal{Q} \notin \boldsymbol{\pi}^*(n)$. In particular,

$$P\left(\{\mathcal{P} \in \boldsymbol{\pi}^*(n)\} \cap \{\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\}\right) \leq P\left(\mathbf{AR}(n, \mathcal{P}) < \mathbf{AR}(n, \boldsymbol{\mathcal{P}}^*_{pop})\right). \tag{7.21}$$

We consider now two cases for $\mathcal{P}$: $\mathcal{P} \in \boldsymbol{\mathcal{M}}_{pop}$ and $\mathcal{P} \notin \boldsymbol{\mathcal{M}}_{pop}$. Starting with the former case, since both $\mathcal{P}$ and $\boldsymbol{\mathcal{P}}^*_{pop}$ are in $\boldsymbol{\mathcal{M}}_{pop}$, by the argument ensuing (7.19), $\int P_{G_\theta}\left(\mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\boldsymbol{\mathcal{P}}^*_{pop})'}\boldsymbol{\beta}^{(\boldsymbol{\mathcal{P}}^*_{pop})}\right)\mathscr{P}(d\theta) = 1$, and therefore for almost every θ, $P_{G_\theta}(\mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})} = \mathbf{X}^{(\boldsymbol{\mathcal{P}}^*_{pop})'}\boldsymbol{\beta}^{(\boldsymbol{\mathcal{P}}^*_{pop})}) = 1$; hence,

$$\sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}^{(\mathcal{P})}_j)^2 = \sum_{j=1}^{\mathcal{J}} E_{G_j}(Y - \mathbf{X}^{(\boldsymbol{\mathcal{P}}^*_{pop})'}\boldsymbol{\beta}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_j)^2,$$

with probability 1. By the definition of $\mathbf{AR}(n, \mathcal{P})$,

$$\mathbf{AR}(n, \boldsymbol{\mathcal{P}}^*_{pop}) - \mathbf{AR}(n, \mathcal{P}) = \frac{\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_j) - \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\mathcal{P})}_j)}{n}$$

Therefore, going back to (7.21), we have

$$P\Big(\mathbf{AR}(n, \mathcal{P}) < \mathbf{AR}(n, \boldsymbol{\mathcal{P}}^*_{pop})\Big) = P\Big(\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_j) - \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\mathcal{P})}_j) > 0\Big).$$

By Part 1,

$$\int tr\left(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_\theta\right)\mathscr{P}(d\theta) - \int tr\left(\mathbb{V}^{(\mathcal{P})}_\theta\right)\mathscr{P}(d\theta) \leq -\varepsilon,$$

where ε is the difference between $\int tr\big(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_\theta\big)\mathscr{P}(d\theta)$ and the second best. Therefore, $E\big(\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_j) - \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\mathcal{P})}_j)\big) \leq -\varepsilon$; also, $Var(tr(V_\theta))$ is bounded (by a constant that depends on C). Chebyshev's inequality implies that

$$P\Big(\frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\boldsymbol{\mathcal{P}}^*_{pop})}_j) - \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}} tr(\mathbb{V}^{(\mathcal{P})}_j) > 0\Big) \leq K_C/\mathcal{J},$$

and therefore, $P\Big(\{\mathcal{P} \in \boldsymbol{\pi}^*(n)\} \cap \{\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\}\Big) \leq K_C/\mathcal{J}$.

Next consider the case $\mathcal{P} \notin \boldsymbol{\mathcal{M}}_{pop}$. By definition, there exists $\varepsilon > 0$ such that for any $\mathcal{P} \notin \boldsymbol{\mathcal{M}}_{pop}$

$$\int E_{G_\theta}(Y - \mathbf{X}^{(\mathcal{P})'}\boldsymbol{\beta}_\theta^{(\mathcal{P})})^2 \mathscr{P}(d\theta) - \int E_{G_\theta}(Y - \mathbf{X}^{(\boldsymbol{\mathcal{P}}^*_{pop})'}\boldsymbol{\beta}_\theta^{(\boldsymbol{\mathcal{P}}^*_{pop})})^2 \mathscr{P}(d\theta) > \varepsilon.$$

It is easy to see that for n_2 large enough this implies

$$E\big(\mathbf{AR}(n, \boldsymbol{\mathcal{P}}^*_{pop}) - \mathbf{AR}(n, \mathcal{P})\big) < -\varepsilon/2 \qquad \forall n \geq n_2.$$

By an argument as above $P\Big(\{\mathcal{P} \in \boldsymbol{\pi}^*(n)\} \cap \{\boldsymbol{\mathcal{P}}^*_{pop} \notin \boldsymbol{\pi}^*(n)\}\Big) \leq K_C/\mathcal{J}$. Since the number of models is finite, (7.20) follows.

Now, for fixed n that satisfies $n < n_0 := \max\{n_1, n_2\}$ again a similar argument shows that for any $\mathcal{P} \in \boldsymbol{\pi}^*(n)$,

$$P\Big(\mathcal{P} \notin \boldsymbol{\pi}^*_{pop}(n)\Big) \leq \frac{K_C(n)}{\mathcal{J}},$$

where $K_C(n)$ may depend on n (and on C). Since there are only finite such n's the result of Part 2 follows. □

Proof of Proposition 3.6. The first part of Proposition 3.6 follows from Part 1 of Proposition 3.3, which shows that $\widehat{\boldsymbol{\pi}^*}(n, \mathbf{N}) \subseteq \boldsymbol{\pi}^*(n)$ with probability converging to 1, and Part 2 of Lemma 7.1, which shows that $\boldsymbol{\pi}^*(n) \subseteq \boldsymbol{\pi}^*_{pop}(n)$ with high probability.

The second part of Proposition 3.6 follows from a combination of several statements: $\widehat{\pi^*}(n, \mathbf{N}) = \boldsymbol{\mathcal{P}}^*(n)$ with high probability (Proposition 3.3, Part 2); $\boldsymbol{\mathcal{P}}^*(n) = \boldsymbol{\pi}^*(n)$ for large n (Proposition 3.2); $\boldsymbol{\pi}^*(n) \subseteq \boldsymbol{\pi}^*_{pop}(n)$ with high probability (Lemma 7.1 Part 2); and for large n, $\boldsymbol{\pi}^*_{pop}(n)$ is a singleton, and $\boldsymbol{\pi}^*_{pop}(n) = \boldsymbol{\mathcal{P}}^*_{pop}(n)$ (Lemma 7.1, Part 1). □

8. Appendix B: A table of notation

Expression	Description
$\mathcal{J}$	Number of observed regression datasets
N_j	Number of observations in the the jth regression dataset
Y_{ij}	The response of the ith observation from the jth regression
$\mathbf{X}_{ij} \in \mathbb{R}^d$	The covariate vector of the ith observation from the jth regression
$(\mathbf{X}, Y)$	A generic observation (whose distribution is G_j)
$D_j = \{(\mathbf{X}_{ij}, Y_{ij})\}$	The jth regression dataset
G_j	The distribution of the jth regression, i.e., $\{(\mathbf{X}_{ij}, Y_{ij})\} \sim^{iid} G_j$
$\mathcal{G}$	A set of distributions to which G_j belongs (the cases $\lvert\mathcal{G}\rvert = 1$, $\lvert\mathcal{G}\rvert = \mathcal{J}$ and $\mathcal{J} < \lvert\mathcal{G}\rvert \leq \infty$ appear in Sections 2, 3.1, and 3.3, respectively)
$\mathcal{P}$	A subset of $\{1, \ldots, d\}$, used to denote a subset of covariates. Its size is denoted by p.
$R(n, \mathcal{P})$	The prediction error of the linear model with covariates in $\mathcal{P}$ with n observations for the case $\lvert\mathcal{G}\rvert = 1$; $\mathbf{R}(n, \mathcal{P})$ and $\mathbf{R}_{pop}(n, \mathcal{P})$ denote the cases of $\lvert\mathcal{G}\rvert = \mathcal{J}$ and $\mathcal{J} < \lvert\mathcal{G}\rvert$, respectively
$AR(n, \mathcal{P})$	Approximate prediction error; $\mathbf{AR}(n, \mathcal{P})$ and $\mathbf{AR}_{pop}(n, \mathcal{P})$ are approximations of $\mathbf{R}(n, \mathcal{P})$ and $\mathbf{R}_{pop}(n, \mathcal{P})$, respectively
In the notation below j and $^{(\mathcal{P})}$ are sometimes suppressed	
$\mathbb{X}_{j,N_j}^{(\mathcal{P})}$	The $N_j \times p$ design matrix of the jth regression
$\mathbf{Y}_{j,N_j}$	The vector of responses for the jth regression
$\boldsymbol{\beta}_j^{(\mathcal{P})}$	Projection coefficients under G_j for model $\mathcal{P}$
$e_j^{(\mathcal{P})}$	The residual; $e_j^{(\mathcal{P})} = Y - \mathbf{X}_j^{(\mathcal{P})'} \boldsymbol{\beta}_j^{(\mathcal{P})}$; $\mathbf{e}_{j,N_j}$ denotes the vector of the residuals of dimension N_j
$\widehat{\boldsymbol{\beta}}_{j,n}^{(\mathcal{P})}$	The least squares estimate of $\boldsymbol{\beta}_j^{(\mathcal{P})}$ based on n observations.
$\mathbb{Q}_j^{(\mathcal{P})}$	$E_{G_j}(\mathbf{X}^{(\mathcal{P})} \mathbf{X}^{(\mathcal{P})'})$
$\mathbb{W}_j^{(\mathcal{P})}$	$E_{G_j}(\mathbf{X}^{(\mathcal{P})} \mathbf{X}^{(\mathcal{P})'} e^2)$
$\mathbb{V}_j^{(\mathcal{P})}$	$\mathbb{W}_j^{(\mathcal{P})} \{\mathbb{Q}_j^{(\mathcal{P})}\}^{-1}$
$\widehat{\mathbb{Q}}_{j,N_j}^{(\mathcal{P})}$	The empirical estimate of $\mathbb{Q}_j^{(\mathcal{P})}$
$\widehat{\mathbb{W}}_{j,N_j}^{(\mathcal{P})}$	The empirical estimate of $\mathbb{W}_j^{(\mathcal{P})}$
$\widehat{\mathbb{V}}_{j,N_j}^{(\mathcal{P})}$	The empirical estimate of $\mathbb{V}_j^{(\mathcal{P})}$
$\mathbf{U}_{j,N_j}^{(\mathcal{P})}$	$\frac{1}{\sqrt{N_j}} \mathbb{X}_{j,N_j}^{(\mathcal{P})'} \mathbf{e}_{j,N_j}$ (it is not a statistic)
$C^{(\mathcal{P})}(n, N)$	An estimate of $AR(n, \mathcal{P})$; $\mathbf{C}^{(\mathcal{P})}(n, \mathbf{N})$ corresponds to the case $\mathcal{J} > 1$; $\mathbb{C}^{(\mathcal{P})}(n, N)$ and $\boldsymbol{\mathbb{C}}^{(\mathcal{P})}(n, N)$ denote a jackknife bias correction
$\mathcal{P}^*(n)$	$\arg\min_{\mathcal{P}} R(n, \mathcal{P})$ (the best model for n observations); $\boldsymbol{\mathcal{P}}^*(n)$ corresponds to the case $\mathcal{J} > 1$
$\pi^*(n)$	$\arg\min_{\mathcal{P}} AR(n, \mathcal{P})$; $\boldsymbol{\pi}^*(n)$ corresponds to the case $\mathcal{J} > 1$
$\mathcal{P}^*$	The limit of both $\mathcal{P}^*(n)$ and $\pi^*(n)$ as $n \to \infty$; $\boldsymbol{\mathcal{P}}^*$ corresponds to the case $\mathcal{J} > 1$
$\widehat{\pi^*}(n, \mathbf{N})$	$\arg\min_{\mathcal{P}} C^{(\mathcal{P})}(n, \mathbf{N})$;$\widehat{\boldsymbol{\pi}^*}(n, \mathbf{N})$ corresponds to the case $\mathcal{J} > 1$