Electronic Journal of Statistics
Vol. 18 (2024) 5449–5483
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2308

Model selection and inference for estimation of causal parameters

Dominik Rothenhäusler

Department of Statistics, Stanford University,
e-mail: rdominik@stanford.edu

Abstract: In causal inference there are often multiple reasonable estimators for a given target quantity. For example, one may reasonably use inverse probability weighting, an instrumental variables approach, or construct an estimate based on proxy outcomes if the actual outcome is difficult to measure. Ideally, the practitioner decides on an estimator before looking at the data. However, this might be challenging in practice since a priori it might not be clear to a practitioner how to choose the method. If the final model is chosen after peeking at the data, naive inferential procedures may fail. This raises the need for a model selection tool, with rigorous asymptotic guarantees. Since there is usually no loss function available in causal inference, standard model selection techniques do not apply.

We propose a model selection procedure that estimates the squared ℓ_2-deviation of a finite-dimensional estimator from its target. The procedure relies on knowing an asymptotically unbiased (potentially highly variable) estimate of the parameter of interest. The resulting estimator is discontinuous and does not have a Gaussian limit distribution. Thus, standard asymptotic expansions do not apply. We derive asymptotically valid confidence intervals for low-dimensional settings that take into account the model selection step.

The performance of the approach for estimation and inference for average treatment effects is evaluated on simulated data sets in low-dimensional settings, including experimental data, instrumental variables settings and observational data with selection on observables.

MSC2020 subject classifications: Primary 62F10, 62F35; secondary 62D20.
Keywords and phrases: Causal inference, model selection, data fusion, efficiency.

Received June 2022.

1. Introduction

Model selection is a fundamental task in statistical practice. Usually, the aim is to find a model that optimizes overall model fit. Overall model fit is usually measured with a risk functional that can be estimated unbiasedly.

Model selection is far less developed in settings where one wants to infer a finite-dimensional parameter on a parametric or semiparametric model, with rigorous statistical guarantees. For example, in causal inference researchers might might want to choose a treatment effect estimate (a potentially one-dimensional quantity) among a set of candidate estimators. Some of these estimates might be unbiased, but have high variance. Other estimates might have low variance

but might be less trustworthy in the sense that they might rely on assumptions that the researcher has not verified yet. For example, a practitioner might have a small set of experimental units and a large pool of observational data. Estimating the causal effect only based on the experimental data might lead to high variance. Combining observational and experimental data might result in improved precision, but can be problematic if the observational units are highly confounded. There is a growing literature that proposes estimators that combine observational and experimental data, see for example Kallus et al. (2018); Rosenman et al. (2020); Hussain et al. (2022). If the decision between such estimators is made ad-hoc, the resulting estimate might be unreliable or naive confidence intervals might not cover the parameter of interest. This raises the need for a reliable model selection tool.

One complicating issue is that causal parameters in general do not admit a risk, that means they can in general not be written as the minimizer of a risk functional that can be estimated unbiasedly. Thus, standard model selection procedures such as loss-based cross-validation, Mallows C_p, or the Bayesian information criterion do not apply. As a result, there has been interest in model-selection procedures that are tailored to causal parameter estimation problems (Van der Laan et al., 2011; Cui and Tchetgen-Tchetgen, 2024).

From a theoretical perspective, model selection can result in non-regular estimators, which invalidates naive inferential approaches based on Gaussian asymptotics. Even worse, model selection that "buys" better behaviour in some parts of the parameter space can be at the expense of erratic behaviour in other parts of the parameter space. Hodges' estimator (Le Cam, 1953) is a famous example where the rescaled excess risk goes to infinity for some parts of the parameter space. Thus, it is important to carefully study and quantify the resulting behaviour of the final estimate. To summarize, our goal is to develop a model selection criterion that satisfies the following three desiderata:

1. The model selection criterion should not depend on a loss functional.
2. Based on variance estimates of the individual estimators, it should be possible to quantify the uncertainty of the selected estimator in the sense that the uncertainty can be propagated through the model selection step.
3. The model selection criterion should be reliable; in the sense that uniformly across parts of the parameter space the excess risk should stay within reasonable bounds.

1.1. Related work

Model selection has a long history in statistics and machine learning. For optimizing loss-based estimators, the most commonly used methods include cross-validation, the Akaike information criterion, and the Bayesian information criterion (Akaike, 1974; Schwarz, 1978; Friedman et al., 2001; Arlot and Celisse, 2010). Since there is usually no loss function available in causal inference, these standard model selection techniques do not apply.

There has been a surge in interest in inference for statistical parameters after model selection. Selective inference (Fithian et al., 2014; Taylor and Tibshirani, 2015; Loftus and Taylor, 2015; Lee et al., 2016; Yang et al., 2016; Hyun et al., 2018) conditions on the selection event; and provide valid coverage for the resulting estimand. Berk et al. (2013) provide uniformly valid coverage by widening conventional confidence intervals. Different from this line of work, we consider the target functional as fixed. The main reasoning behind this is that some of the candidate estimators might have considerable bias. Thus, switching between different estimands might lead to problematic performance in practice.

In a sequence of papers, Leeb and Pötscher (Leeb and Pötscher, 2003, 2005, 2006) warn that (asymptotic) inference after model selection can hide some finite-sample issues and is usually not uniformly valid. Their results warrant a careful investiation of the proposed confidence intervals. Thus, we will study minimal realized coverage of the proposed confidence intervals in Section 4.

Lepski's method is a tool to select the bandwidth in nonparametric smoothing problems (Lepskii, 1991). Lepski's method is usually used to estimate an infinite-dimensional object. On the other hand, our goal is to estimate a finite-dimensional (often one-dimensional) parameter such as the average treatment effect. Even though the overall goals and asymptotic setups are different, Lepski's procedure can be applied in our setting. We will compare the proposed procedure to Lepski's method in Section 4.

The focused information criterion is a model selection criterion which, for a given focus parameter, estimates the mean-squared error of submodels (Claeskens and Hjort, 2003, 2008). It relies on knowing an asymptotically unbiased estimator of the parameter of interest. Its theoretical justification is given in a local misspecification framework.

More recently, in the context of causal inference Cui and Tchetgen-Tchetgen (2024) introduce a model selection tool for finite-dimensional functionals in a semiparametric model if a doubly robust estimation function is available. It is based on a pseudo-risk criterion that has a robustness property if one of the estimators is biased. However, the final estimate can still be biased, even asymptotically, if two of the nuisance models are biased. Our approach differs from Cui and Tchetgen-Tchetgen (2024) in that we assume that the data scientist trusts one estimator more than some of the other estimators. Our approach results in a model selection criterion that is robust even if several nuisance models are biased.

For the task of model selection when estimating heterogeneous treatment effects, several methods have been developed (Kapelner et al., 2014; Rolling and Yang, 2014; Athey and Imbens, 2016; Nie and Wager, 2021; Zhao et al., 2017; Powers et al., 2018). Most of the methodologies are specific to the considered model class. A comparison of this line of work for individual treatment effects can be found in Schuler et al. (2018).

Van der Laan and Robins (2003) propose a loss-based approach for parameter-specific model selection. In this work, the authors recommend minimizing an empirical estimate of the overall risk $R(\hat{\theta}^{(g)}, \hat{\eta})$, where $\hat{\theta}^{(g)}$, $g = 1, \ldots, G$ are candidate estimators and $\hat{\eta}$ is an efficient estimator of the nuisance parame-

ter, computed on the training data set. Our approach is more generic in the sense that we do not assume that parameter of interest minimizes a known loss function.

Closest to our work is the sample-splitting criterion developed by Brookhart and Van Der Laan (2006). Roughly speaking, the data is split into a training and a test data set. Then, estimators are computed on the training and the test data set, and the squared deviation of estimators is aggregated across several splits. The criterion developed by Brookhart and Van Der Laan (2006) can be seen as a form of Monte Carlo cross-validation. In the following, we discuss a variant of this approach that splits the data into k folds and thus mimics k-fold cross-validation procedures, which are popular in practice. The data $D = (D_1, \ldots, D_n)$ is randomly split into K disjoint roughly equally-sized folds $D^{0,1}, \ldots, D^{0,K}$. Define $D^{1,k} = D \setminus D^{0,k}$. Assuming that the data are i.i.d., $D^{1,k}$ and $D^{0,k}$ are independent for each k. Let $\hat{\theta}^{(0)}$ be an unbiased estimator of the parameter of interest $\theta^{(0)} \in \mathbb{R}^d$. If several unbiased estimators are available, aggregation procedures such as inverse variance weighting can be used in a preprocessing step to obtain $\hat{\theta}^{(0)}$. Let $\hat{\theta}^{(g)}$ be candidate estimators, $g = 0, \ldots, G$. Then, we can compute the risk criterion

$$\tilde{R}(g) = \frac{1}{K} \sum_{k=1}^{K} \|\hat{\theta}^{(g)}(D^{1,k}) - \hat{\theta}^{(0)}(D^{0,k})\|_2^2. \tag{1}$$

Using independence of $D^{1,k}$ and $D^{0,k}$,

$$\mathbb{E}[\tilde{R}(g)] = \mathbb{E}[\|\hat{\theta}^{(g)}(D^{1,1}) - \theta^{(0)}\|_2^2] + \sum_{j=1}^{d} \mathrm{Var}(\hat{\theta}_j^{(0)}(D^{0,1})).$$

As $\mathrm{Var}(\hat{\theta}^{(0)}(D^{0,1}))$ is constant in g, the criterion in equation (1) can be used to select an estimator $\hat{\theta}^{(g)}$ with low mean-squared error for estimating $\theta^{(0)}$ among $\hat{\theta}^{(g)}, g = 0, \ldots, G$. The criterion in equation (1) is attractive as it is simple and widely applicable. We will compare the proposed model selection criterion to the criterion in equation (1), both from a theoretical perspective and in simulations.

1.2. Our contribution

We derive a model selection criterion that estimates the squared ℓ_2-deviation of an estimator from its target. By construction the proposed selection criterion does not depend on a loss functional. Instead, it relies on knowing an asymptotically unbiased (potentially highly variable) estimate of the target of interest.

Our main goal is to select models for causal estimation problems. In such settings, modern ML-based estimation techniques allow for $n^{-1/2}$-consistent inference of common target parameters (Chernozhukov et al., 2018). Furthermore, practitioners usually desire confidence intervals for the resulting estimate. Thus, we study a regime where the individual estimators are $n^{-1/2}$-consistent for their target quantity.

Even if the candidate estimators are asymptotically linear, the model selection procedure is discontinuous and will not result in a regular estimator, even for $n \to \infty$. Thus, the final estimator does not have a Gaussian limit distribution. We derive asymptotically valid confidence intervals for the resulting estimator in low-dimensional settings that takes into account the model selection step. Furthermore, we compare the asymptotic behaviour of the procedure to a competing procedure based on sample splitting.

Compared to the baseline procedure, for fixed n, model selection can lead to increased risk in parts of the parameter space. We provide a finite-sample bound that reveals that the excess risk due to model selection becomes negligible as the dimension of the target parameter grows.

In low-dimensional simulation settings, we compare the proposed procedure to competing procedures, including variants of cross-validation, Lepski's method, and selective machine learning. The proposed procedure shows very promising performance across several settings.

The code can be found at github.com/rothenhaeusler/tms.

1.2.1. Outline

In Section 2, we introduce a method for parameter-specific model selection and discuss an example. Theory for the method is discussed in Section 3. We evaluate the performance of the proposed procedure on simulated data in Section 4.

2. Targeted model selection

This section consists of two parts. We briefly discuss the setting in Section 2.1. Then, we introduce the method in Section 2.2 and discuss basic properties.

2.1. Setting and notation

We observe data $D = (D_i,\ i = 1, \ldots, n)$, where the D_i are independently drawn from some unknown distribution $\mathbb{P}$. Suppose we have access to estimators $\hat{\theta}^{(g)}(D)$, $g = 0, \ldots, G$, of some unknown parameter $\theta^{(0)}$. In the following, to simplify notation, we will write $\hat{\theta}^{(g)}$ instead of $\hat{\theta}^{(g)}(D)$. We assume that the baseline estimator $\hat{\theta}^{(0)}$ is asymptotically unbiased for $\theta^{(0)}$, i.e. that $\mathbb{E}[\hat{\theta}^{(0)}] = \theta^{(0)} + o(n^{-1/2})$. In practice, the data scientist may know several estimators that are asymptotically unbiased for the parameter of interest. In this case, one can use aggregation procedures such as inverse variance weighting to construct an optimally weighted aggregated estimator $\hat{\theta}^{(0)}$.

In addition, the data scientist may have access to estimators $\hat{\theta}^{(g)}$ for which the data scientist is not sure whether they are approximately unbiased for the effect of interest. The goal is to select among the set of estimators, minimizing the mean-squared error with respect to the target of interest $\theta^{(0)}$. We assume that $\mathbb{E}[\hat{\theta}^{(g)}] = \theta^{(g)} + o(n^{-1/2})$ for some unknown $\theta^{(g)}$ and that $\sqrt{n}(\hat{\theta}^{(g)} - \theta^{(g)})$

converges to a non-degenerate random variable. Please note that this assumption does not preclude modern inferential estimation strategies, such as debiased inference in high-dimensional settings or estimation based on machine-learning tools, see for example (Van der Laan et al., 2011; Zhang and Zhang, 2014; Chernozhukov et al., 2018). We chose this setting mainly due to the fact that we want to apply the method in causal inference, where it is usually desired to have asymptotically valid confidence intervals for the resulting parameter estimates.

We write $\sigma_j^{(g)}$ for the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(g)})$. Similarly, we assume that $\sqrt{n}(\hat{\theta}^{(g)} - \hat{\theta}^{(0)} - (\theta^{(g)} - \theta^{(0)}))$ converges to a non-degenerate random variable for $g \neq 0$ and write $\tau_j^{(g)}$ for the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(g)} - \hat{\theta}_j^{(0)} + \theta_j^{(0)})$.

2.2. The method

We aim to find an estimator g that minimizes

$$R(g) = \mathbb{E}[\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2]. \tag{2}$$

Here and in the following, we suppress the dependence of $R(g)$ and $\hat{\theta}^{(g)}$ on n. As bias and variance of $\hat{\theta}^{(g)}$ are unknown, the function $R(g)$ is unknown and one has to minimize a proxy of the risk $R(g)$ instead. We propose to estimate $R(g)$ in equation (2) via

$$\hat{R}(g) = \|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 + \sum_{j=1}^{d} \frac{(\hat{\sigma}_j^{(g)})^2}{n} - \frac{(\hat{\tau}_j^{(g)})^2}{n}, \tag{3}$$

where $\hat{\sigma}_j^{(g)}$ is an estimator of the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(g)})$ and $\hat{\tau}_j^{(g)}$ is an estimator of the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(g)} - \hat{\theta}_j^{(0)} + \theta_j^{(0)})$. The intuition is that for each j, $(\hat{\theta}_j^{(g)} - \hat{\theta}_j^{(0)})^2 - \frac{(\hat{\tau}_j^{(g)})^2}{n}$ is an estimate of the squared bias, while $\frac{(\hat{\sigma}_j^{(g)})^2}{n}$ is an estimate of the variance of $\hat{\theta}_j^{(g)}$. If the estimators are asymptotically linear (i.e. in some semi-parametric or low-dimensional parametric settings), consistent estimators $\hat{\tau}^{(g)}$ and $\hat{\sigma}^{(g)}$ are usually available via plug-in estimators of the variance of the influence function (Van der Vaart, 2000; Tsiatis, 2007). An example will be discussed below. We propose to choose a final estimate $\hat{\theta}^{(\bar{g})}$ by solving

$$\bar{g} = \arg\min_g \hat{R}(g). \tag{4}$$

Let us consider a linear regression example. This example was mainly chosen for expository simplicity; the main motivating examples for this method are drawn from causal inference. The causal inference examples need more discussion and will be explained in detail in Section 4.

Example 1 (Model selection for parameter estimation)**.** *Usually, when conducting model selection in the context of prediction, the goal is to find a model that can be estimated well and is a good approximation of some complex model of interest. However, if the purpose is parameter estimation, fitting complex models can reduce variance while potentially introducing bias. Such settings appear in causal inference and will be further discussed in Section 4. Here, we consider the task where the goal is to fit a regression with just one covariate; but there are additional covariates at our disposal that can be used to reduce variance, while potentially introducing some bias for the parameter of interest. Let* $Y_i = X_i\theta^{(0)} + \epsilon_i$*, where* $D_i = (Y_i, X_i)$ *are i.i.d. and the* ϵ_i *are centered noise terms that are uncorrelated of the* X_i*. Furthermore, for simplicity we assume that* Y_i*,* X_i *and* ϵ_i *are centered. We are interested in the parameter* $\theta^{(0)} = \arg\min \mathbb{E}[(Y - X\theta)^2]$ *and consider the baseline estimator*

$$\hat{\theta}^{(0)} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - X_i\theta)^2.$$

Let us assume that we have access to observations $Z_1, \ldots, Z_n$ *from some additional covariate. One may consider the estimator*

$$\hat{\theta}^{(1)} = \arg\min_{\theta} \min_{\eta} \sum_{i=1}^{n} (Y_i - X_i\theta - Z_i\eta)^2,$$

Let (X, Y, Z, ϵ) *denote a generic* $(X_i, Y_i, Z_i, \epsilon_i)$*. If* Z *is correlated with* Y *and only weakly correlated with* X*, this estimator may reduce asymptotic variance compared to* $\hat{\theta}^{(0)}$ *since intuitively speaking, adjusting for* Z *reduces unexplained variation in the residuals. On the other hand if* Z *is strongly correlated with* X*,* $\hat{\theta}^{(1)}$ *may converge to a different parameter than* $\hat{\theta}^{(0)}$*. Under regularity conditions (Van der Vaart, 2000, Section 5),* $\hat{\theta}^{(0)}$ *is asymptotically linear and unbiased for* $\theta^{(0)} := \arg\min_{\theta} \mathbb{E}[(Y - X\theta)^2]$*, i.e.*

$$\sqrt{n}(\hat{\theta}^{(0)} - \theta^{(0)}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}[X^2]^{-1} X_i \epsilon_i + o_P(1).$$

Similarly, under regularity conditions,

$$\begin{aligned}&\sqrt{n}(\hat{\theta}^{(1)} - \theta^{(1)})\\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_1^\intercal \mathbb{E}[(X, Z)^\intercal (X, Z)]^{-1} (X_i, Z_i)^\intercal (Y_i - X_i\theta^{(1)} - Z_i\eta^{(1)}) + o_P(1),\end{aligned}$$

where $(\theta^{(1)}, \eta^{(1)}) = \arg\min_{(\theta,\eta)} \mathbb{E}[(Y - X\theta - Z\eta)^2]$ *and where* e_j *denotes the* j*-th unit vector. Thus,*

$$\begin{aligned}(\sigma^{(0)})^2 &= Var(\mathbb{E}[X^2]^{-1} X\epsilon),\\ (\sigma^{(1)})^2 &= Var(e_1^\intercal \mathbb{E}[(X, Z)^\intercal (X, Z)]^{-1} (X, Z)^\intercal (Y - X\theta^{(1)} - Z\eta^{(1)}),\end{aligned}$$

$$(\tau^{(0)})^2 = 0,$$
$$(\tau^{(1)})^2 = Var\left(e_1^\intercal \mathbb{E}[(X,Z)^\intercal (X,Z)]^{-1}(X,Z)^\intercal (Y - X\theta^{(1)} - Z\eta^{(1)}) - \mathbb{E}[X^2]^{-1} X\epsilon\right).$$

These quantities can be consistently estimated via plug-in estimators in standard settings. For example, $(\sigma^{(1)})^2$ and $(\sigma^{(0)})^2$ can be consistently estimated via the sandwich estimator under regularity assumptions (Huber, 1967).

We will study the risk proxy in equation (3) in Section 3. The method is evaluated on simulated data sets in Section 4.

2.2.1. Improving precision

Recall that the risk criterion can be decomposed into several parts, i.e. $\hat{R}(g) = \sum_{j=1}^d \hat{R}_{\text{bias},j}(g) + \hat{R}_{\text{var},j}(g)$, where

$$\hat{R}_{\text{bias},j}(g) = (\hat{\theta}_j^{(g)} - \hat{\theta}_j^{(0)})^2 - \frac{\hat{\tau}_j^{(g)}}{n},$$

and

$$\hat{R}_{\text{var},j}(g) = \frac{\hat{\sigma}_j^{(g)}}{n}.$$

As the naming indicates, the first term can be interpreted as an estimate of the squared bias $(\theta^{(g)} - \theta^{(0)})^2$, whereas the second term is an estimate of the variance of $\hat{\theta}^{(g)}$. Since we know that squared bias terms are non-negative this motivates defining the following modified risk criterion:

$$\hat{R}^{\text{mod}}(g) = \left(\sum_{j=1}^d (\hat{\theta}_j^{(g)} - \hat{\theta}_j^{(0)})^2 - \frac{(\hat{\tau}_j^{(g)})^2}{n}\right)_+ + \sum_{j=1}^d \frac{(\hat{\sigma}_j^{(g)})^2}{n}. \tag{5}$$

Then the final estimator $\hat{\theta}^{(\bar{g})}$ is chosen such that $\bar{g}$ minimizes equation (5). We take the positive part of the sum (instead of the sum of positive parts) as this allows random errors to cancel out for large d. This will be important for the theory developed in Section 3.5. If there are ties, we select $\bar{g}$ as the one that minimizes $\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2$ among the g that satisfy $\hat{R}^{\text{mod}}(g) = \min_{g'} \hat{R}^{\text{mod}}(g')$. The criterion $\hat{R}^{\text{mod}}(g)$ is not asymptotially unbiased for $R(g)$, but has some favorable statistical properties that we will discuss in the following section.

3. Theory

In this section we discuss the theoretical underpinnings of the method introduced in Section 2. First, we show that the criterion $\hat{R}(g)$ is asymptotically unbiased for estimating the mean-squared error $R(g) = \mathbb{E}[\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2]$. Then, we discuss the asymptotic risk of the resulting estimator. We derive asymptotically valid

confidence intervals for the parameter of interest that takes into account the model selection step. Finally, we present a finite-sample bound that shows that if the dimension of the target parameter is large, the excess risk due to model selection becomes negligible.

3.1. Assumptions

We make two major assumptions, in addition to the assumptions outlined in Section 2.1. The first major assumption is a slightly stronger version of asymptotic linearity. Asymptotic linearity is an assumption that is commonly made to justify asymptotic normality of an estimator (Van der Vaart, 2000; Tsiatis, 2007). As our goal is to estimate the mean-squared error of an estimator, we use a slightly stronger version that guarantees convergence of second moments. The second major assumption is that the variance estimates are consistent.

Assumption 1. *We make two major assumptions.*

1. *Let* $\hat{\theta}^{(g)}$, $g = 0, \dots, G$ *be estimators such that*

$$\hat{\theta}^{(g)} - \theta^{(g)} = \frac{1}{n} \sum_{i=1}^{n} \psi^{(g)}(D_i) + e_g(n),$$

   where $\psi^{(g)}(D_i)$ *are centered and have finite nonzero second moments, and* $\mathbb{E}[\|e_g(n)\|_2^2] = o(1/n)$. *The parameters* $\theta^{(g)}$ *might depend on* n *(for example, one might have* $\theta^{(g)} - \theta^{(0)} = \frac{c_g}{\sqrt{n}}$, *but we suppress this in the notation). To avoid trivial special cases, in addition we assume that the covariance matrix of* $(\psi^{(0)}, \dots, \psi^{(G)})$ *is positive definite.*
2. *The estimators of variance are consistent, that means*

$$(\hat{\tau}^{(g)})^2 = (\tau^{(g)})^2 + o_P(1),$$
$$(\hat{\sigma}^{(g)})^2 = (\sigma^{(g)})^2 + o_P(1).$$

Let us compare the first part of the assumption to asymptotic linearity. Asymptotic linearity assumes that $\|e_g(n)^2\|_2^2 = o_P(1/n)$ while we assume that $\mathbb{E}[\|e_g(n)^2\|_2^2] = o(1/n)$. Thus, our assumption is stronger than asymptotic linearity. Let us now turn to our theoretical results.

3.2. Asymptotic unbiasedness

Let us now turn to the asymptotic behaviour of the proposed procedure. First, if $\theta^{(g)} \neq \theta^{(0)}$ is fixed, then by Assumption 1 we immediately have $\hat{R}(g) \xrightarrow{p} \|\theta^{(g)} - \theta^{(0)}\|_2^2 = \lim_{n\to\infty} \mathbb{E}[\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2]$. It is more interesting to study how the model selection criterion behaves if $\theta^{(g)}$ is close to (but different) from $\theta^{(0)}$. Our first result shows that the proposed criterion is asymptotically unbiased for the mean-squared error, for $\theta^{(g)}$ in a neighborhood of $\theta^{(0)}$. More specifically, we allow $\theta^{(g)}$ to vary across n and keep $\theta^{(0)}$ fixed, but we notationally suppress the

dependence of $\theta^{(g)}$ on n. The proof of the following result can be found in the supplement.

Theorem 1 (Asymptotic unbiasedness of $\hat{R}(g)$)**.** *Let Assumption 1 hold. If* $\theta^{(g)} - \theta^{(0)} = \frac{c_g}{\sqrt{n}} + o(1/\sqrt{n})$*, for some fixed* c_g*, then*

$$n(\hat{R}(g) - \mathbb{E}[\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2])$$

converges weakly to a random variable with mean zero.

This result shows that the criterion is asymptotically unbiased, which is important for the interpretation of the risk criterion. However, it does not make any statement about the validity of the selected model. In the following section, we study the asymptotic risk of the selected model.

3.3. Asymptotic risk

Here, we focus on the case where the number of models G is small and fixed and $n \to \infty$. A finite-sample bound that applies to high-dimensional settings will be discussed in Section 3.5. First, we investigate the asymptotic behaviour of the proposed procedure in the case where the number of models is fixed and $n \to \infty$. The proof of the following result can be found in the supplement.

Corollary 1 (Asymptotic risk of selected model)**.** *Let Assumption 1 hold and let* $\theta^{(0)}, \ldots, \theta^{(G)}$ *be fixed. Consider a finite and fixed number of estimators* $g = 0, \ldots, G$*. Let*

$$\bar{g} = \arg\min \hat{R}^{mod}(g).$$

For $n \to \infty$*,*

$$\mathbb{P}[\theta^{(\bar{g})} = \theta^{(0)}] \to 1,$$

and

$$\mathbb{P}[R(\bar{g}) \leq R(0)] \to 1.$$

Furthermore, if there exists a g *that has better asymptotic risk than the baseline model, meaning that*

$$\theta^{(g)} = \theta^{(0)} \text{ and } \sum_{j=1}^{d} (\sigma_j^{(g)})^2 < \sum_{j=1}^{d} (\sigma_j^{(0)})^2,$$

then for $n \to \infty$ *we have* $R(\overline{g}) < R(0)$ *with probability exceeding* c*, where* $c > 0$ *is a constant that does not depend on* n*.*

In words, for $n \to \infty$, the proposed method selects models with lower or equal risk than the baseline estimator $\hat{\theta}^{(0)}$. This seems to be a relatively weak property and heavily desired from a model selection criterion. However, an analogous result does not hold for the cross-validation procedure (1). Cross-validation is inconsistent in general, see for example Yang (2007) and references therein. In the following, for completeness, we discuss an example for which the cross-validation procedure has higher risk than the baseline estimator, even as $n \to \infty$. The proof of the following result can be found in the supplement.

Proposition 1. *Consider the case $G = 1$, that means the case of two independent estimators $\hat{\theta}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\psi^{(1)}(D_i)$, where $\psi^{(1)}(D_i) \overset{i.i.d.}{\sim} \mathcal{N}(0,2)$ and $\hat{\theta}^{(0)} = \frac{1}{n}\sum_{i=1}^{n}\psi^{(0)}(D_i)$, where $\psi^{(0)}(D_i) \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let Asssumption 1 hold. For simplicity we assume that n is divisible by K. Then there exists $c > 0$ that does not depend on n such that with probability exceeding c the cross-validation procedure (1) selects a model $\tilde{g}$ with $R(\tilde{g}) > R(0)$.*

The issue with this example is that the alternative estimator, while unbiased, has higher variance than the baseline estimator. This will lead the cross-validation based procedure to select the alternative estimator with positive probability, even for $n \to \infty$. For the proposed procedure, we essentially avoid this issue due to the modification introduced in Section 2.2.1.

In Section 4 we will see in a numerical example that even for relatively large n, the selected model by cross-validation may have risk that is significantly larger than the risk of the baseline procedure.

3.4. Confidence intervals

Deriving confidence intervals that are valid in conjuction with a model selection step is a challenging topic and has attracted substantial interest in recent years, see for example Berk et al. (2013) and Taylor and Tibshirani (2015). Generally speaking, statistical inference after a model selection step can be unreliable if the uncertainty induced by the model selection step is ignored. In this section, we describe how to construct confidence intervals that take into account the uncertainty induced by model selection. Intuitively speaking, the challenge is that the final estimator is a discontinuous function of the data. To be more precise, the final estimator $\hat{\theta}^{(\bar{g})}$ is not regular and not asymptotically normal.

The goal in this section is to find I_1 and I_2 as a function of the data $D_1, \ldots, D_n$ such that

$$\mathbb{P}[\hat{\theta}_j^{(\bar{g})} - I_1 \leq \theta_j^{(0)} \leq \hat{\theta}_j^{(\bar{g})} + I_2] \to 1 - \alpha,$$

for some pre-determined $\alpha > 0$ and j and where

$$\bar{g} = \arg\min_g \hat{R}^{\text{mod}}(g).$$

The following theorem shows how to construct confidence intervals in low-dimensional settings; i.e. in settings where the number of models G is fixed and the sample size goes to infinity.

Theorem 2. *Define $\hat{\theta} = (\hat{\theta}^{(0)}, \hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(G)}) \in \mathbb{R}^{(G+1)d}$ and let $\theta = (\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(G)})$ be fixed. Assume that*

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \Sigma),$$

for some positive definite Σ. Let $\hat{\Sigma}(D) = \hat{\Sigma}_n(D_1, \ldots, D_n)$ be a consistent estimator of $\Sigma \in \mathbb{R}^{(G+1)d \times (G+1)d}$ and $\hat{\sigma}^{(g)} = \hat{\sigma}^{(g)}(D_1, \ldots, D_n)$ be a consistent

estimator of the asymptotic standard deviation of $\sqrt{n}(\hat{\theta}^{(g)} - \theta^{(g)})$ *and* $\hat{\tau}^{(g)} = \hat{\tau}^{(g)}(D_1, \ldots, D_n)$ *be a consistent estimator of the asymptotic standard deviation of* $\sqrt{n}(\hat{\theta}^{(g)} - \hat{\theta}^{(0)} - \theta^{(g)} + \theta^{(0)})$. *With some abuse of notation, conditionally on the data* $D = (D_1, \ldots, D_n)$ *draw* $(Z_0, \ldots, Z_G) \sim \mathcal{N}(\sqrt{n}\hat{\theta}(D), \hat{\Sigma}(D))$ *with* $Z_g \in \mathbb{R}^d$. *We define the event*

$$\begin{aligned} A_g^{est} = \{&\max(\|Z_g - Z_0\|_2^2 - \|\hat{\tau}^{(g)}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2 \\ &< \min_{g' \neq g} \max(\|Z_{g'} - Z_0\|_2^2 - \|\hat{\tau}^{(g')}\|_2^2, 0) + \|\hat{\sigma}^{(g')}\|_2^2\}. \end{aligned}$$

Now for some fixed j *define*

$$b_{D_1,\ldots,D_n}(\beta) = \sum_g \mathbb{P}[\{Z_{g,j} - \sqrt{n}\hat{\theta}_j^{(0)} \leq \beta\} \cap A_g^{est} | D_1, \ldots, D_n]. \tag{6}$$

Then:

1. *For* $n \to \infty$, *with probability converging to one, the inverse* $b^{-1}_{D_1,\ldots,D_n} : (0,1) \to \mathbb{R}$ *is well-defined.*
2. *For all* $\alpha > 0$,

$$\mathbb{P}\left[\hat{\theta}_j^{(\bar{g})} - \frac{b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)}{\sqrt{n}} \leq \theta_j^{(0)} \leq \hat{\theta}_j^{(\bar{g})} - \frac{b^{-1}_{D_1,\ldots,D_n}(\alpha/2)}{\sqrt{n}}\right] \to 1 - \alpha.$$

Note that by definition the conditional distribution of Z given $D_1, \ldots, D_n$ is known to the researcher. Also, the researcher often can construct estimators of the variances in parametric and semi-parametric settings via plug-in estimators of the influence function, see e.g. Van der Vaart (2000); Tsiatis (2007). In these cases, $b_{i,\alpha}(D_1, \ldots, D_n)$ can be computed by the researcher, for example by Monte-Carlo simulation. Thus, Theorem 2 allows us to construct asymptotically valid confidence intervals for the final estimator $\hat{\theta}^{(\bar{g})}$ in many parametric and semi-parametric settings.

This result is different from standard asymptotic arguments in the sense that $\hat{\theta}^{(\bar{g})} - \theta^{(0)}$ is not asymptotically Gaussian. However, as the result shows, it is possible to recover the exact asymptotic distribution of $\hat{\theta}^{(\bar{g})}$ in low-dimensional scenarios and use this information to conduct asymptotically valid statistical inference. We will evaluate the empirical performance of confidence intervals constructed via Theorem 2 in Section 4.

Inference after model selection, is usually not uniformly valid see Leeb and Pötscher (2005) for an overview. Our proposal falls into what Leeb and Pötscher call "Conservative Model Selection Framework", since model selection is asymptotically not consistent. Thus, one has to use these inferential tools with caution. In simulation settings (Section 4), we will investigate the coverage of confidence intervals uniformly across parts of the parameter space.

3.5. Finite-sample bound

In this section, we discuss a finite-sample bound that can be applied to high-dimensional settings. In particular, we will not need to assume that the para-

metric rates described in Section 2.1 hold, and instead give a risk bound that depends on tail bounds of the candidate estimators.

As discussed in Section 3.3, the proposed method selects a model that is asymptotically no worse than the baseline estimator. However, there is no free lunch. In transitional regimes, for fixed n, the estimator can perform worse than the baseline estimator $\hat{\theta}^{(0)}$. This is to be expected from statistical theory, see for example the discussion of the Hodges-Le Cam estimator on page 110 in Van der Vaart (2000). This makes it important to understand in which cases we can expect reliable performance of the proposed model selection procedure. In the case $d = 1$, a Bayesian bound (Gill and Levit, 1995) reveals that improving over the Cramér-Rao bound in some parts of the parameter space must lead to deteriorating performance in other parts of the parameter space. In the following, we will provide a finite-sample bound that shows that under strong regularity assumptions, for large d the excess risk becomes negligible, uniformly over a set of distributions.

Theorem 3. *Let* $\hat{\theta}^{(g)} = \theta^{(g)} + \epsilon^{(g)} = \theta^{(0)} + \delta^{(g)} + \epsilon^{(g)}$ *where* $\delta^{(g)} \in \mathbb{R}^d$ *is a constant vector and* $G > 1$. *We assume that the* $\epsilon_j^{(g)}$ *are centered, independent and sub-Gaussian random variables with variance proxy* $\eta_j^{(g)}/\sqrt{n}$,[1] *i.e. that* $\mathbb{E}[\exp\left(s\epsilon_j^{(g)}\right)] \leq \exp\left(\frac{(\eta_j^{(g)})^2 s^2}{2n}\right)$ *for all* $g = 1, \ldots, G$ *and* $j = 1, \ldots, d$ *and* $s \in \mathbb{R}$. *Define* $\tau_j^{(g)}$ *as the standard deviation of* $\sqrt{n}(\epsilon_j^{(g)} - \epsilon_j^{(0)})$ *and* $\sigma_j^{(g)}$ *as the standard deviation of* $\sqrt{n}\epsilon_j^{(g)}$. *We define* $b_\infty = \max_g \|\delta^{(g)}\|_2$ *and* $s_\infty = \max_{j,g} \eta_j^{(g)}$. *Define* $\iota_{n,d} := \max(\sup_{g,j} |(\hat{\sigma}_j^{(g)})^2 - (\sigma_j^{(g)})^2|, \sup_{g,j} |(\hat{\tau}_j^{(g)})^2 - (\tau_j^{(g)})^2|)$ *and* $M = \sup_g \|\hat{\theta}^{(g)} - \theta^{(0)}\|_2 + \|\delta^{(0)}\|_2$. *Furthermore, assume that* $\log G/d \leq c_1$ *for some constant* $c_1 > 0$. *Then, for every* $\kappa > 0$ *there exists a constant* C *that may depend on* c_1, *and* κ *(but not on* δ, s_∞ *or* b_∞*) such that with probability exceeding* $1 - \kappa$,

$$\frac{1}{d}\sum_{j=1}^{d}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)})^2 \leq \min_g \frac{1}{d}\sum_{j=1}^{d}(\hat{\theta}_j^{(g)} - \theta_j^{(0)})^2 + C\left(\frac{s_\infty^2}{n}\sqrt{\frac{\log G}{d}} + \frac{s_\infty b_\infty}{\sqrt{n}}\sqrt{\frac{\log G}{d}}\right) + \frac{6\iota_{n,d}}{n} + 2\frac{M}{d}\|\delta^{(0)}\|_2.$$

A few comments are in order. First, the candidate estimators $\hat{\theta}^{(g)}$ usually have smaller variance than the baseline estimator $\hat{\theta}^{(0)}$ – otherwise one would not consider model selection. Thus, usually one has $s_\infty = \max_j \eta_j^{(0)}$. Secondly, $\hat{\theta}^{(0)}$ should have small bias $\|\delta^{(0)}\|_2$ – otherwise it is not justified to use $\hat{\theta}^{(0)}$ as the baseline estimator. The excess risk due to model selection goes to zero as $\frac{d}{n \log G} \to \infty$ and if $\sup_{g,j} |(\hat{\sigma}_j^{(g)})^2 - \sigma_j^{(g)}| \to 0$ and $\sup_{g,j} |\hat{\tau}_j^{(g)} - \tau_j^{(g)}| \to 0$ and if

[1] The $\eta_j^{(g)}$ can be arbitrarily large. In modern causal inference, it is common to de-bias high-dimensional or semi-parametric estimators to arrive at a componentwise $\sqrt{n}$-rate. Thus, the re-scaling of the variance proxy with $\sqrt{n}$ is to facilitate interpretation of the bound.

the bias $\|\delta^{(0)}\|_2$ is negligible. Note that the latter is a strong requirement - for a high-dimensional estimation procedures tuned via standard cross-validation, the bias $\|\delta^{(0)}\|_2$ would not be negligible. In Section 4 we will see in an example that the excess risk declines for growing dimension of the estimand.

4. Applications

In this section, we discuss applications of the proposed method. As we will see, model selection can lead to drastic improvements in the mean-squared error. However, there is no free lunch. Compared to the baseline procedure, for fixed n, model selection can lead to increased risk in parts of the parameter space. Thus, in this section, we study the excess risk of the procedures across the parameter space. Leeb and Pötscher (2005) warn that pointwise valid confidence intervals after model selection are usually not uniformly valid. Thus, in addition to average coverage, we will also report minimal realized coverage.

In the following we will use potential outcomes to define causal effects (Rubin, 1974; Splawa-Neyman et al., 1990). We are interested in the causal effect of a treatment $T \in \{0, 1\}$ on an outcome Y. Let $Y(1)$ denote the potential outcome under treatment $T = 1$ and $Y(0)$ the potential outcome under treatment $T = 0$. We assume a superpopulation model, i.e. $Y(1)$ and $Y(0)$ are random variables. In the following, the goal is to estimate the average treatment effect within several subgroups,

$$\theta_s^{(0)} = \mathbb{E}[Y(1) - Y(0)|S = s]. \tag{7}$$

Many methods have been designed to estimate (7) and these methods operate under a variety of assumptions. We present several applications that are based on different sets of assumptions for identifying (7). In each of the cases, we compare the proposed method (5, termed "targeted selection") with the cross-validation procedure (1), with selective machine learning (Cui and Tchetgen-Tchetgen, 2024), with Lepski's method (Lepski et al., 1997), and with a baseline estimator. In addition, we compare with reverse K-fold cross-validation, which has shown promise for comparing close candidates (Shao, 1993; Yang, 2007; Zhang and Yang, 2015; Zhan and Yang, 2022). The code can be found at github.com/rothenhaeusler/tms.

4.1. Observational studies

In observational studies, it is common practice to estimate causal effects under the assumption of unconfoundedness and under the overlap assumption. Roughly speaking, the overlap assumption states that treatment assignment probabilities are bounded away from zero and one, conditional on covariates X. If these assumptions are met, it is possible to identify the average treatment effect via matching, inverse probability weighting, regression adjustment, or doubly robust methods (Hernan and Robins, 2020; Imbens and Wooldridge, 2009). However, if the overlap is limited, estimating the average treatment effect can be unreliable.

To deal with the issue of limited overlap, researchers sometimes switch to different estimands such as the average effect on the treated (ATT) or the overlap weighted effect (Crump et al., 2006). This raises issues of post selection inference. We will see that the proposed model selection tool results in trustworthy statistical inference, even in conjunction with the model selection step. In the following, we will focus on the overlap-weighted effect as it is the causal contrast that can be estimated with the lowest asymptotic variance in certain scenarios (Crump et al., 2006). The overlap-weighted effect is defined as

$$\theta^{(1)} = \frac{\mathbb{E}[p(T=1|X)(1-p(T=1|X))\tau(X)]}{\mathbb{E}[p(T=1|X)(1-p(T=1|X))]},$$

where $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$. Note that if the treatment effect is homogeneous $\tau(x) \equiv$ const., then the overlap-weighted effect and the average treatment effect coincide, that means $\theta^{(1)} = \theta^{(0)}$. Thus, shrinkage towards an efficient estimator of the overlap effect is potentially beneficial under treatment effect homogeneity.

We investigate shrinking between estimators of the average treatment effect and the overlap-weighted effect in a data-driven way. The proposed model selection tool will be used to trade off bias and variance.

4.1.1. The data set

We observe 1000 independent and identically distributed draws $(Y_i(T_i), T_i, X_i, S_i)$ of a distribution $\mathbb{P}$, where the X_i are covariates. The data generating process was chosen such that there is limited overlap, i.e. $\mathbb{P}[T = 1|X = 0] \approx 0$ and that the unconfoundedness assumptions, that means $(Y(0), Y(1)) \perp T|X$ (Rosenbaum and Rubin, 1983). As discussed above, the causal effect can be estimated via doubly robust methods such as augmented inverse probability weighting, among others (Hernan and Robins, 2020). The data are generated according to the following equations:

$$\begin{aligned} &S \text{ drawn from } \{1,2,3\} \text{ uniformly at random} \\ &\epsilon_Y \sim \mathcal{N}(0,1) \\ &X \sim \text{Ber}(.5) \\ &T \sim \begin{cases} \text{Ber}(.7) & \text{if } X = 1, \\ \text{Ber}(.05) & \text{if } X = 0, \end{cases} \\ Y(t) &= \frac{X}{2} + t + 3t\gamma^2 X + .1t \cdot 1_{S=1} + .2t \cdot 1_{S=2} - .1t \cdot 1_{S=3} + \epsilon_Y, \end{aligned} \tag{8}$$

where $\gamma \in [0,1]$. For $\gamma = 0$, the treatment effect is homogeneous across X. Thus, for $\gamma = 0$, the overlap-weighted effect coincides with the average treatment effect. In the Appendix, we present a variant of this simulation with a 10-dimensional covariate vector.

4.1.2. The estimators

In the following, we compute the estimators for each group $S = s$ separately on the data set $\{i : S_i = s\}$. For reasons of readability, notationally we suppress the dependence of the conditional probabilities and conditional expectations on s. We can estimate that average treatment effect via augmented inverse probability weighting (Robins et al., 1994),

$$\hat{\theta}_s^{(0)} = \hat{\mu}_1 - \hat{\mu}_0,$$

where

$$\hat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i 1_{T_i=a}}{\hat{p}(T_i = a|X_i)} - \frac{1_{T_i=a} - \hat{p}(T_i = a|X_i)}{\hat{p}(T_i = a|X_i)}\hat{Q}(X_i, a),$$

and where $\hat{Q}(x, t)$ is the empirical mean of Y given $X = x$ and $T = t$ and $\hat{p}(\cdot|\cdot)$ are empirical probabilities. Similarly as above, we can estimate the overlap effect by

$$\hat{\theta}_s^{(1)} = \frac{\hat{\eta}_1 - \hat{\eta}_0}{\frac{1}{n}\sum_i \hat{p}(T_i = 1|X_i)(1 - \hat{p}(T_i = 1|X_i))},$$

where

$$\begin{aligned}\hat{\eta}_a = \frac{1}{n}\sum_{i=1}^{n} & Y_i 1_{T_i=a}(1 - \hat{p}(T_i = a|X_i)) \\ & - (1_{T_i=a} - \hat{p}(T_i = a|X_i))(1 - \hat{p}(T_i = a|X_i))\hat{Q}(X_i, a).\end{aligned}$$

For $w \in \{1/10, \ldots, 9/10\}$ we define

$$\hat{\theta}(w) = (1 - w)\hat{\theta}^{(0)} + w\hat{\theta}^{(1)}. \tag{9}$$

For $\gamma \approx 0$, due to treatment effect homogeneity we expect $\mathbb{E}[(\hat{\theta}(1) - \theta^{(0)})^2] < \mathbb{E}[(\hat{\theta}^{(0)} - \theta^{(0)})^2]$. For $\gamma \approx 1$, we expect $\mathbb{E}[(\hat{\theta}(1) - \theta^{(0)})^2] > \mathbb{E}[(\hat{\theta}^{(0)} - \theta^{(0)})^2]$. In the first case, the optimal estimator is $\hat{\theta}(w)$ with $w \approx 1$. In the second case the optimal estimator is $\hat{\theta}(w)$ with $w \approx 0$.

4.1.3. Results

The mean-squared error of the estimator selected by targeted selection (the proposed procedure), 10-fold cross-validation, reverse 10-fold cross-validation, selective machine learning and Lepski's method is depicted in Figure 1. To study the influence of dimension d on the performance of the model selection procedure, we show how the method performs on the full data set (left-hand side) and how the method performs if it only has access to the subset of observations i for which $S_i = 1$ (right-hand side). Results are averaged across 100 simulation runs.

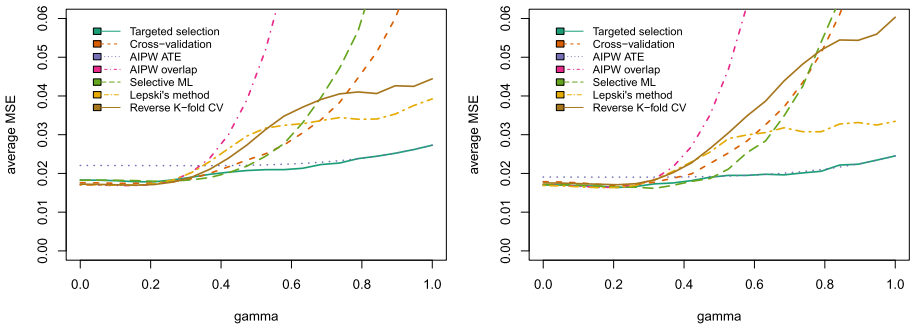

FIG 1. *Several model selection procedures are used to shrink between the AIPW ATE estimator and the AIPW overlap estimator. The proposed procedure is referred to as "targeted selection". On the left-hand side, we show the error* $\|\hat{\theta}_{\bullet}^{(\overline{g})} - \theta_{\bullet}^{(0)}\|_2^2/3$. *On the right-hand side the method is run on the subset of observations* i *for which* $S_i = 1$. *The data are drawn according to equation* (8).

We evaluate the realized coverage of confidence intervals with nominal coverage 95% as described in Section 3.4. Equation 6 is estimated using the non-parametric bootstrap. Across $\gamma \in [0, 1]$, the minimal realized coverage is 93% and the maximal realized coverage is 97%. Averaged across all $\gamma \in [0, 1]$, the overall coverage is 94.9%.

4.2. Instrumental variables and data fusion

The instrumental variables approach is a widely-used method to estimate causal effect of a treatment T on a target outcome Y in the presence of confounding (Wright, 1928; Bowden and Turkington, 1990; Angrist et al., 1996). Roughly speaking, the method relies on a predictor I (called the instrument) of the treatment T that is not associated with the error term of the outcome Y. We will not discuss the assumptions behind instrumental variables in detail, but refer the interested reader to Hernan and Robins (2020). We will focus on the case, where I, T and Y are one-dimensional. Under IV assumptions and linearity, the target quantity can be re-written as

$$\theta^{(0)} = \mathbb{E}[Y(1) - Y(0)] = \frac{\mathrm{Cov}(I, Y)}{\mathrm{Cov}(I, T)}.$$

Estimating this quantity can be challenging if the instrument is weak, i.e. if $\mathrm{Cov}(I, T) \approx 0$. In this case, the approach can benefit from shrinkage towards the ordinary least-squares solution (Nagar, 1959; Theil, 1961; Rothenhäusler et al., 2021; Jakobsen and Peters, 2022). Doing so may decrease the variance but generally introduces bias. We will focus on the case where we have some additional observational data, where we observe T and Y, but where the instrument I is unobserved.

4.2.1. The data set

We draw 500 i.i.d. observations according to the following equations:

$$\begin{aligned} S &\text{ drawn from } \{1,2,3\} \text{ uniformly at random} \\ I, H, \epsilon_T, \epsilon_Y &\sim \mathcal{N}(0,1) \\ T &= \frac{I}{2} + H + \epsilon_T \\ Y(t) &= t - \gamma^2 H + .1t \cdot 1_{S=1} + .2t \cdot 1_{S=2} - .1t \cdot 1_{S=3} + \epsilon_Y \end{aligned} \tag{10}$$

We vary $\gamma \in [0,2]$, which corresponds to the strength of confounding between T and Y. We observe $(T_i, Y_i(T_i), I_i)$ for $i = 1, \ldots, 500$. We also assume that we have access to a larger data set $i = 501, \ldots, 1000$ with incomplete observations. To be more precise, on this data set we only observe X and Y, but not the instrument I. Formally, for $i = 501, \ldots, 1000$ we observe $(T_i, Y_i(T_i))$ drawn according to equation (10).

4.2.2. The estimators

In the linear case, for each subset $S = s$, the instrumental variables estimator can be written as

$$(\hat{b}_{IV})_s = \frac{\hat{\mathrm{Cov}}(I, Y | S = s)}{\hat{\mathrm{Cov}}(I, T | S = s)},$$

where $\hat{\mathrm{Cov}}$ denotes the empirical covariance over the observations $i = 1, \ldots, 500$. To deal with the weak instrument, we will consider shrinking the instrumental variables estimator torwards ordinary least-squares,

$$(\hat{b}_{\mathrm{OLS}})_s = \operatorname*{argmin}_b \min_c \hat{\mathbb{E}}[(Y - Tb - c)^2 | S = s],$$

where $\hat{\mathbb{E}}$ denotes the empirical expectation over the observations $i = 1, \ldots, 1000$. Shrinking towards the ordinary least-squares solution will introduce some bias if $\gamma \neq 0$, but potentially decreases variance. As candidate estimators, for any $w \in \{0/10, 1/10, \ldots, 10/10\}$ we consider convex combinations of OLS and IV,

$$\hat{\theta}(w) = w\hat{b}_{\mathrm{OLS}} + (1 - w)\hat{b}_{\mathrm{IV}}.$$

Cross-validation was performed similarly as in the previous section.

4.2.3. Results

The results can be found in Figure 2. To summarize, in this setting the proposed procedure (targeted selection) performs better than the competing procedures. Compared to the baseline procedure (the green dotted line), targeted selection

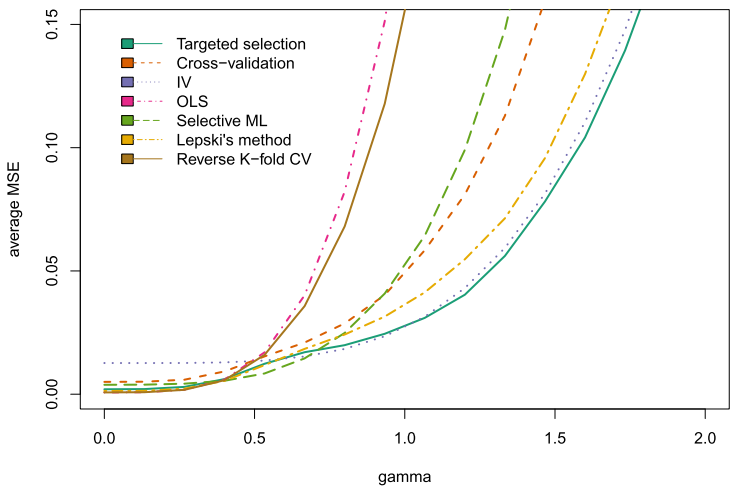

FIG 2. *The model selection procedures are used to stabilize the instrumental variables approach by shrinking the estimate towards ordinary least-squares. On the y-axis, we report the mean-squared error $R(\hat{w})$ where $\hat{w}$ is selected via cross-validation, reverse cross-validation, targeted selection (the proposed method), selective machine learning, or Lepski's method. The data are drawn according to equation* (10).

performs better under strong confounding and if the confounding is weak, but has slightly larger MSE in the transitional regime where $s \approx .5$. Similarly as discussed before, we evaluate the coverage of confidence intervals with nominal coverage .95. Since the confidence intervals are not uniformly valid, a drop in the realized coverage is expected in transitional regimes. Across all $\gamma \in [0, 2]$ the minimal realized coverage is 89%, while the overall realized coverage is 93%.

4.3. Experiment with proxy outcome

One of the most popular estimators for causal effects in experimental settings is difference-in-means. To improve variance, it is possible to adjust for pre-treatment covariates, see for example Lin (2013). This raises the question whether post-treatment covariates can be used to improve the precision of causal effect estimates. This is indeed the case under additional assumptions. For example, in some cases, the treatment effect can be written as the product

$$\theta^{(0)} = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \theta_{T\to P} \cdot \theta_{P\to Y}, \tag{11}$$

where $\theta_{T\to P} = \mathbb{E}[P|T=1] - \mathbb{E}[P|T=0]$ is the effect of the treatment on some surrogate or proxy outcome $P \in \{0, 1\}$; and $\theta_{P\to Y} = \mathbb{E}[Y|P=1] - \mathbb{E}[Y|P=0]$ is the effect of the proxy on the outcome. It is well-known that estimators that make use of such decompositions can outperform the standard difference-in-means estimator in terms of asymptotic variance (Tsiatis, 2007; Athey et al.,

2019; Guo and Perković, 2022). However, doing so can introduce bias if equation (11) does not hold. We will use the proposed model selection procedure to shrink between difference-in-means and an estimator that is unbiased if the treatment effect decomposition in equation (11) holds.

4.3.1. The data set

We consider a simple experimental setting with a post-treatment variable P. For simplicity, let us consider an experiment with binary treatment $T \in \{0,1\}$, a binary proxy outcome $P \in \{0,1\}$ and outcome Y. We draw 200 i.i.d. observations according to the following equations:

$$\begin{aligned} S &\text{ drawn from } \{1,2,3\} \text{ uniformly at random} \\ T &\sim \text{Ber}(.5) \\ \epsilon_P, \epsilon_Y &\sim \mathcal{N}(0,1) \\ P(t) &= 1_{\epsilon_p \le t} \\ Y(t) &= P(t) + \gamma^2 \left(.1t \cdot 1_{S=1} + .2t \cdot 1_{S=2} - .1t \cdot 1_{S=3}\right) + \epsilon_Y \end{aligned} \tag{12}$$

For $\gamma = 0$, the outcome $Y(T)$ is conditionally independent of the treatment, given the proxy $P(T)$. In this case, the average treatment effect can be written in product form, $\theta^{(0)} = \theta_{T\to P} \cdot \theta_{P\to Y}$, and this decomposition can be leveraged for estimation. For $\gamma \neq 0$, this decomposition does not hold.

4.3.2. The estimators

The standard estimator to estimate causal effects from experiments is difference-in-means,

$$\hat{\theta}^{(0)} = \frac{1}{\sum T_i} \sum_{i:T_i=1} Y_i - \frac{1}{\sum (1-T_i)} \sum_{i:T_i=0} Y_i. \tag{13}$$

If the proxy outcome is a valid surrogate, i.e. if

$$Y \perp T | P,$$

we can rewrite $\theta^{(0)}$ as

$$\begin{aligned} \theta^{(0)} &= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \\ &= \mathbb{E}[\mathbb{E}[Y|P=1]P + \mathbb{E}[Y|P=0](1-P)|T=1] \\ &\quad - \mathbb{E}[\mathbb{E}[Y|P=1]P + \mathbb{E}[Y|P=0](1-P)|T=0] \\ &= (\mathbb{E}[Y|P=1] - \mathbb{E}[Y|P=0])\,(\mathbb{E}[P|T=1] - \mathbb{E}[P|T=0]) \\ &= \theta_{T\to P} \cdot \theta_{P\to Y}. \end{aligned}$$

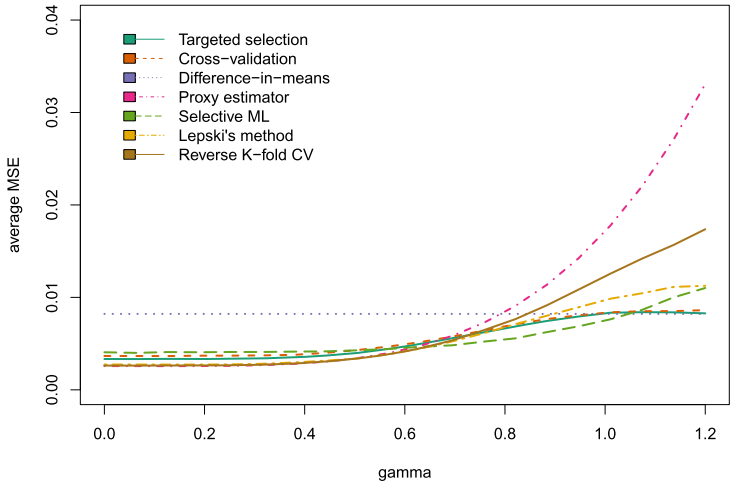

FIG 3. *Cross-validation, reverse cross-validation, selective ML, Lepski's method and targeted selection (our proposal) is used to stabilize the difference-in-means estimator by shrinking towards an estimator that makes use of a proxy outcome. The data are drawn according to equation* (12).

Thus, in this case, we can also consider the product estimator

$$\hat{\theta}^{(1)} = \left(\frac{1}{\sum T_i} \sum_{i:T_i=1} P_i - \frac{1}{\sum (1-T_i)} \sum_{i:T_i=0} P_i \right) \cdot \left(\frac{1}{\sum P_i} \sum_{i:P_i=1} Y_i - \frac{1}{\sum (1-P_i)} \sum_{i:P_i=0} Y_i \right) \tag{14}$$

On each subset $\{i : S_i = s\}$ we compute (13) and (14), yielding $\hat{\theta}_s^{(1)}$ and $\hat{\theta}_s^{(0)}$ for $s = 1, 2, 3$. We shrink between these two vectors, i.e. for $w \in \{1/10, \ldots, 9/10\}$ we define

$$\hat{\theta}(w) = (1-w)\hat{\theta}^{(0)} + w\hat{\theta}^{(1)}.$$

4.3.3. Results

The results are depicted in Figure 3. Similarly as above, targeted selection performs similar or better than cross-validation. Overall, selective ML performs similar compared to the proposed procedure. Lepski's method performs well for small γ, but is suboptimal for most of the parameter space. The minimal realized coverage of 95% confidence intervals is 93% and the overall realized coverage is 94.6%.

5. Conclusion

In causal inference, there are often multiple reasonable estimators for a given target quantity. Ideally, the practitioner decides on an asymptotically regular estimator before looking at the data and conducts inference as usual. However, this might be challenging in practice since a priori it might not be clear to a practitioner which method is the best. If the final model is chosen ad-hoc after peeking at the data, naive inferential procedures will fail. This necessitates the development of model selection tools, with rigorous inferential guarantees.

We have introduced a method that allows to conduct targeted parameter selection by estimating the bias and variance of candidate estimators. The theoretical justification of the method relies on a linear expansion of the estimator. The method can be used in both parametric and semi-parametric settings. Under regularity conditions, we showed that the proposed criterion provides an asymptotically unbiased estimate of the risk. In addition, we showed that for $n \to \infty$, the modified risk criterion selects models with lower or equal risk than the baseline estimator $\hat{\theta}^{(0)}$. Furthermore, we derived asymptotically valid confidence intervals in low-dimensional settings.

In low-dimensional simulation settings, we showed that the method selects reasonable models and performs similarly or better than cross-validation, selective ML and Lepski's method in simulations. The proposed method can decrease variance if the competing estimators are approximately unbiased. However, there is no free lunch. In transitional regimes, for fixed n, the estimator can perform worse than the baseline estimator $\hat{\theta}^{(0)}$. This is to be expected from statistical theory, see for example Van der Vaart (2000, page 110). However, as the finite-sample bound shows, the excess risk can go to zero in settings where the bias of the base procedure $\hat{\theta}^{(0)}$ is negligible. The simulations cover the cases of one-dimensional and three-dimensional target parameters. Such low-dimensional target parameters are common in causal inference, where the target parameter often represents the causal effect for a (sub-)population. For the three-dimensional target parameters the excess risk is smaller than for the one-dimensional parameter. A simulation with a 10-dimensional covariate vector can be found in Appendix A.

Since the confidence intervals are not uniformly valid, we expect the minimal coverage to be below .95 (Leeb and Pötscher, 2005). However, in simulations, the drop in coverage seems to be limited.

The theoretical justification of the proposed method relies on a linear approximation of the estimator in a neighborhood of the parameter values $\theta^{(g)}$. Thus, it would be important to understand the performance of the method in scenarios where parameter estimates of some of the estimators are far from the parameter values. In Section 4.2, we have seen some preliminary evidence that the proposed methodology may be used to combine knowledge across data sets. The proposed method is not tailored to this special case. Thus, we believe that it would be exciting to investigate whether the model selection can be further improved for data fusion tasks.

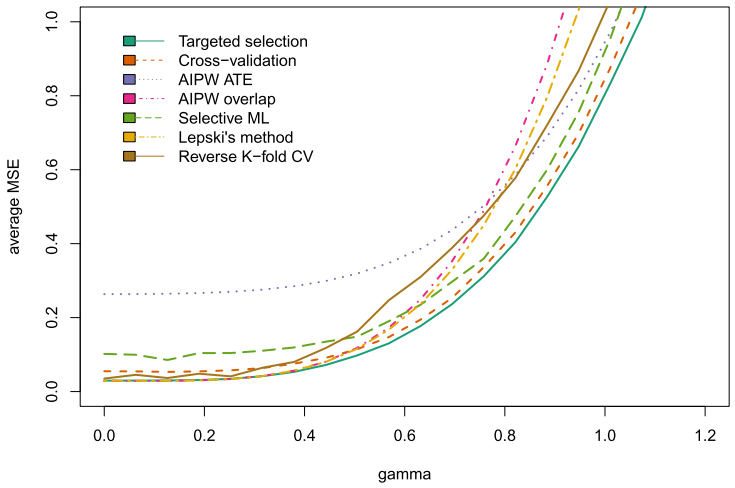

FIG 4. *Several model selection procedures are used to shrink between the AIPW ATE estimator and the AIPW overlap estimator. The proposed procedure is referred to as "targeted selection". The data are drawn according to equation* (15).

Appendix A: Additional numerical results

In this section, we present a variant of the simulation in Section 4.1, with a 10-dimensional covariate vector. To be specific, the data is generated according to

$$\begin{aligned} &S \text{ drawn from } \{1,2,3\} \text{ uniformly at random} \\ &\epsilon_Y \sim \mathcal{N}(0,1) \\ &X \sim \mathcal{N}(0, \mathrm{Id}_{10}) \\ &T \sim \mathrm{Ber}(\mathrm{expit}(X^\intercal \beta)) \\ Y(t) &= X^\intercal \beta + 3t\gamma^2(X_1+1) + .1t \cdot 1_{S=1} + .2t \cdot 1_{S=2} - .1t \cdot 1_{S=3} + \epsilon_Y. \end{aligned} \tag{15}$$

Here, β is a vector of coefficients randomly drawn from the unit sphere in $\mathbb{R}^{10}$. Analogously as in Section 4.1, we interpolate between augmented inverse probability weighting and the overlap effect. The results can be found in Figure 4.

Appendix B: Proofs

This appendix contains proofs for the theoretical results in the main paper.

B.1. Proof of Theorem 1

Proof. First, note that

$$n(\hat{R}(g) - \mathbb{E}[\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2]) = \sum_{j=1}^{d} n(\hat{\theta}_j^{(g)} - \hat{\theta}_j^{(0)})^2 - \hat{\tau}_j^{(g)} + \hat{\sigma}_j^{(g)} - \mathbb{E}[(\hat{\theta}_j^{(g)} - \theta_j^{(0)})^2].$$

Thus, it is sufficient to show that for each j

$$n(\hat{\theta}_j^{(g)} - \hat{\theta}_j^{(0)})^2 - \hat{\tau}_j^{(g)} + \hat{\sigma}_j^{(g)} - \mathbb{E}[(\hat{\theta}_j^{(g)} - \theta_j^{(0)})^2],$$

converges in distribution to a centered random variable. Thus, without loss of generality, we assume $d = 1$. As $\mathbb{E}[e_g(n)^2] = o(1/n)$ and as $\mathbb{E}[\hat{\theta}^{(g)}] - \theta^{(g)} = \frac{c_g}{\sqrt{n}} + o(1/\sqrt{n})$,

$$\hat{R}(g) = \frac{c_g^2}{n} + 2\frac{c_g}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n} \psi^{(g)}(D_i) - \psi^{(0)}(D_i) + \frac{1}{n}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi^{(g)}(D_i) - \psi^{(0)}(D_i)\right)^2 - \frac{(\tau^{(g)})^2}{n} + \frac{(\sigma^{(g)})^2}{n} + o_P(\frac{1}{n}),$$

and

$$\mathbb{E}[(\hat{\theta}^{(g)} - \theta^{(0)})^2] = \frac{c_g^2}{n} + \frac{1}{n}\mathrm{Var}(\psi^{(g)}(D)) + o(1/\sqrt{n}).$$

Thus,

$$n(\hat{R}(g) - \mathbb{E}[(\hat{\theta}^{(g)} - \theta^{(0)})^2]) = 2c_g\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi^{(g)}(D_i) - \psi^{(0)}(D_i) + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi^{(g)}(D_i) - \psi^{(0)}(D_i)\right)^2 - (\tau^{(g)})^2 + o_P(1) \tag{16}$$

Using the CLT, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi^{(g)}(D_i) - \psi^{(0)}(D_i)$ converges in distribution to a centered Gaussian random variable with variance $(\tau^{(g)})^2$. Using this fact in equation (16) concludes the proof. □

B.2. Proof of Corollary 1

Proof. By assumption, we have $\hat{\theta}^{(g)} - \hat{\theta}^{(0)} = \theta^{(g)} - \theta^{(0)} + O_P(1/\sqrt{n})$. If $\theta^{(g)} - \theta^{(0)} \not\equiv (0, \ldots, 0)$, then

$$\hat{R}^{\mathrm{mod}}(g) = \|\theta^{(g)} - \theta^{(0)}\|_2^2 + O_P(1/\sqrt{n}),$$

with $\|\theta^{(g)} - \theta^{(0)}\|_2^2 > 0$. On the other hand, if $\theta^{(g)} = \theta^{(0)}$,

$$\hat{R}^{\mathrm{mod}}(g) = O_P(1/n).$$

Thus,

$$\mathbb{P}[\theta^{(\bar{g})} = \theta^{(0)}] \to 1.$$

Now consider any g with $\theta^{(g)} = \theta^{(0)}$ and $\sum_j \text{Var}(\psi_j^{(g)}) > \sum_j \text{Var}(\psi_j^{(0)})$. Then, using Assumption 1,

$$\hat{R}^{\text{mod}}(g) \geq \sum_j \frac{1}{n} \text{Var}(\psi_j^{(g)}) + o_P(1/n),$$

and

$$\hat{R}^{\text{mod}}(0) = \sum_j \frac{1}{n} \text{Var}(\psi_j^{(0)})) + o_P(1/n).$$

Recall that by assumption $\sum_j \text{Var}(\psi_j^{(g)}) > \sum_j \text{Var}(\psi_j^{(0)})$. Thus, $\mathbb{P}[\hat{R}^{\text{mod}}(0) < \hat{R}^{\text{mod}}(g)] \to 1$ for $n \to \infty$. As this holds for all g with $\sum_j \text{Var}(\psi_j^{(g)}) > \sum_j \text{Var}(\psi_j^{(0)})$ for $n \to \infty$, this concludes the proof of the first statement. For the second statement, please note that due to the first part of the proof, without loss of generality we can assume that $\theta^{(g)} = \theta^{(0)}$ for all $g = 1, \ldots, G$. Thus, asymptotically $\sqrt{n}(\hat{\theta}^{(g)} - \hat{\theta}^{(0)})$ converge to non-degenerate centered Gaussians $Z^{(1)}, \ldots, Z^{(G)}$. Please note that $Z^{(0)} \equiv 0$. Now let there exist a $\overline{g}$ with $\sum_{j=1}^d (\sigma_j^{(\overline{g})})^2 < \sum_{j=1}^d (\sigma_j^{(0)})^2$. Thus, for $n \to \infty$, the probability that model $\overline{g}$ gets selected converges to the probability of the event

$$\begin{aligned} p = \mathbb{P}\Bigg[& \max\left(\|Z^{\overline{g}}\|_2^2 - \sum_j (\tau_j^{(\overline{g})})^2, 0 \right) + \sum_j (\sigma_j^{\overline{g}})^2 \\ & < \min_{g \neq \overline{g}} \max\left(\|Z^g\|_2^2 - \sum_j (\tau_j^{(g)})^2, 0 \right) + \sum_j (\sigma_j^g)^2 \Bigg] \end{aligned} \tag{17}$$

We now want to show that this probability is non-zero. Let's choose a fixed number z that is strictly larger than $\sum_j (\sigma_j^{\overline{g}})^2$ and strictly smaller than $\sum_j (\sigma_j^g)^2$. Then,

$$\begin{aligned} p \geq \mathbb{P}\Bigg[& \max\left(\|Z^{\overline{g}}\|_2^2 - \sum_j (\tau_j^{(\overline{g})})^2, 0 \right) + \sum_j (\sigma_j^{\overline{g}})^2 \\ & < z \text{ and } z < \min_{g \neq \overline{g}} \max\left(\|Z^g\|_2^2 - \sum_j (\tau_j^{(g)})^2, 0 \right) + \sum_j (\sigma_j^g)^2 \Bigg] \end{aligned} \tag{18}$$

Since the $Z^{(g)}$ are non-degenerate for $g \neq 0$, this probability is non-zero. □

B.3. Proof of Proposition 1

Proof. Note that we have $n = K|D^{0,1}|$. Then,

$$n\tilde{R}(1) = \frac{n}{K} \sum_{k=1}^K \left(\frac{1}{|D^{0,k}|} \sum_{i \in D^{0,k}} \psi^{(0)}(D_i) - \frac{1}{K-1} \sum_{k' \neq k} \frac{1}{|D^{0,k'}|} \sum_{i \in D^{0,k'}} \psi^{(1)}(D_i) \right)^2$$

$$= \sum_{k=1}^{K} \left(\frac{1}{\sqrt{|D^{0,k}|}} \sum_{i \in D^{0,k}} \psi^{(0)}(D_i) - \frac{1}{K-1} \sum_{k' \neq k} \frac{1}{\sqrt{|D^{0,k'}|}} \sum_{i \in D^{0,k'}} \psi^{(1)}(D_i) \right)^2$$

Thus, the distribution does not depend on n. In particular,

$$n\tilde{R}(1) \stackrel{d}{=} \sum_{k=1}^{K} \left(\psi^{(0)}(D_k) - \frac{1}{K-1} \sum_{k' \neq k} \psi^{(1)}(D_{k'}) \right)^2 \tag{19}$$

Similarly,

$$n\tilde{R}(0) \stackrel{d}{=} \sum_{k=1}^{K} \left(\psi^{(0)}(D_k) - \frac{1}{K-1} \sum_{k' \neq k} \psi^{(0)}(D_{k'}) \right)^2 \tag{20}$$

Furthermore, the joint distribution of the left-hand side of equations (19) and (20) is equal to the joint distribution of the right-hand side of equations (19) and (20). Thus, we have reduced the problem to the question whether

$$\sum_{k=1}^{K} \left(\psi^{(0)}(D_k) - \frac{1}{K-1} \sum_{k' \neq k} \psi^{(1)}(D_{k'}) \right)^2 < \sum_{k=1}^{K} \left(\psi^{(0)}(D_k) - \frac{1}{K-1} \sum_{k' \neq k} \psi^{(0)}(D_{k'}) \right)^2 \tag{21}$$

with positive probability. This is indeed easy to prove. First, almost surely we have $\sum_{k=1}^{K} (\psi^{(0)}(D_k) - \frac{1}{K-1} \sum_{k' \neq k} \psi^{(0)}(D_{k'}))^2 > 0$. Secondly, if $\psi^{(1)}(D_k) = \psi^{(0)}(D_{k'})$ for $k, k' = 1, \ldots, K$, the inequality holds. Thus, there exist a nonempty subset of configurations of $\psi^{(0)}(D_1), \ldots, \psi^{(0)}(D_K), \psi^{(1)}(D_1), \ldots, \psi^{(1)}(D_K) \in \mathbb{R}^{2K}$ for which the inequality holds. By continuity, the set of configurations for which this inequality holds is a nonempty open subset of $\mathbb{R}^{2K}$. Since the density of

$$\psi^{(0)}(D_1), \ldots, \psi^{(0)}(D_K), \psi^{(1)}(D_1), \ldots, \psi^{(1)}(D_K)$$

is positive on $\mathbb{R}^{2K}$, the probability of the event (21) must be nonzero. □

B.4. Proof of Theorem 2

Proof. We will first prove the first statement. Note that if $\hat{\Sigma}$ is positive definite, $\beta \mapsto b_{D_1,\ldots,D_n}(\beta)$ is continuous and strictly increasing and $\lim_{\beta \to \infty} b_{D_1,\ldots,D_n}(\beta) = 1$, $\lim_{\beta \to -\infty} b_{D_1,\ldots,D_n}(\beta) = 0$. As $\hat{\Sigma}$ converges in probability to a positive definite matrix, the inverse $b^{-1}_{D_1,\ldots,D_n} : (0,1) \to \mathbb{R}$ is well-defined, except on an event with vanishing probability as $n \to \infty$.

Let us now turn to the second statement. We use the decomposition

$$\mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \leq \beta] = \sum_{g} \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(0)}) \leq \beta; \bar{g} = g]. \tag{22}$$

Define the event

$$A_g = \{\bar{g} = g\}$$

$$= \{\max(n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \|\hat{\tau}^{(g)}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2$$
$$< \min_{g' \neq g} \max(n\|\hat{\theta}^{(g')} - \hat{\theta}^{(0)}\|_2^2 - \|\hat{\tau}^{(g')}\|_2^2, 0) + \|\hat{\sigma}^{(g')}\|_2^2\}.$$

It is crucial to understand how this event behaves in the limit. $\sqrt{n}(\hat{\theta} - \theta)$ converges to a centered Gaussian distribution with covariance matrix Σ. Thus, define

$$A_g^{lim} = \max(\|Z_g^0 - Z_0^0\|_2^2 - \|\tau^{(g)}\|_2^2, 0) + \|\sigma^{(g)}\|_2^2$$
$$< \min_{g' \neq g; \theta^{(g')} = \theta^{(0)}} \max(\|Z_{g'}^0 - Z_0^0\|_2^2 - \|\tau^{(g')}\|_2^2, 0) + \|\sigma^{(g)'}\|_2^2,$$

where $(Z_0^0, \ldots, Z_G^0) \sim \mathcal{N}(0, \Sigma)$, independent of the data $\{D_j\}_{j \in \mathbb{N}}$. Before we investigate the large-sample behaviour of equation (22), let us fix some notation.

$$f_g(\beta) := \mathbb{P}[\{\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(0)}) \leq \beta\} \cap A_g]$$
$$f_g^{lim}(\beta) := \begin{cases} \mathbb{P}[\{(Z_g^0)_j \leq \beta\} \cap A_g^{lim}] & \text{if } \theta^{(g)} = \theta^{(0)}, \\ 0 & \text{else.} \end{cases}$$
$$f_g^{comp}(\beta, D_1, \ldots, D_n) := \mathbb{P}[Z_{g,j} - \sqrt{n}\hat{\theta}_j^{(0)} \leq \beta; \max(\|Z_g - Z_0\|_2^2$$
$$- \|\hat{\tau}^{(g)}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2$$
$$< \min_{g' \neq g} \max(\|Z_g - Z_0\|_2^2 - \|\hat{\tau}^{(g)}\|_2^2, 0)$$
$$+ \|\hat{\sigma}^{(g)}\|_2^2 | D_1, \ldots, D_n],$$

where $(Z_0, \ldots, Z_G) \sim \mathcal{N}(\sqrt{n}\hat{\theta}(D), \hat{\Sigma}(D))$, conditionally on the data D. In the first step, let us consider g for which $\theta^{(g)} = \theta^{(0)}$. Recall that $\hat{\sigma}^{(g)} \to \sigma^{(g)}$, $\hat{\tau}^{(g)} \to \tau^{(g)}$, and $\hat{\Sigma} \to \Sigma$. Also, $\sqrt{n}(\hat{\theta} - \theta)$ converges weakly to a distribution that is equal to the distribution of Z^0. By weak convergence, for all β,

$$\lim_n f_g(\beta) = f_g^{lim}(\beta) = \lim_n f_g^{comp}(\beta, D_1, \ldots, D_n), \tag{23}$$

where the limit on the right-hand side is in probability. Now, let us focus on g for which $\theta^{(g)} \neq \theta^{(0)}$. Note that for all g with $\theta^{(g)} \neq \theta^{(0)}$, $n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 \to \infty$. Thus, for all g with $\theta^{(g)} \neq \theta^{(0)}$,

$$\mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(0)}) \leq \beta; \bar{g} = g]$$
$$\leq \mathbb{P}[\max(n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \|\hat{\tau}^{(g)}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2$$
$$< \min_{g' \neq g} \max(n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \|\hat{\tau}^{g}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2]$$
$$\leq \mathbb{P}[\max(n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \|\hat{\tau}^{(g)}\|_2^2, 0) + \|\hat{\sigma}^{(g)}\|_2^2 < \|\hat{\sigma}^{(0)}\|_2^2].$$

Since $\hat{\sigma}^{(0)} \to \sigma^{(0)}$ and $\hat{\sigma}^{(g)} \to \sigma^{(g)}$ and $\hat{\tau}^{(g)} \to \tau^{(g)}$ and $n\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 \to \infty$, for all g with $\theta^{(g)} \neq \theta^{(0)}$ we have $\mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(g)} - \theta_j^{(0)}) \leq \beta; \bar{g} = g] \to 0$. Thus for all g

with $\theta^{(g)} \neq \theta^{(0)}$, $f_g(\beta) \to 0$. Analogously it can be shown that $f_g^{comp}(\beta, D) \to 0$ for all g with $\theta^{(g)} \neq \theta^{(0)}$. Thus, for g with $\theta^{(g)} \neq \theta^{(0)}$,

$$\lim_n f_g(\beta) = f_g^{lim}(\beta) = \lim_n f_g^{comp}(\beta), \tag{24}$$

where the limit on the right-hand side is in probability. By equation (23) and equation (24), for all g,

$$\lim_n f_g(\beta) = f_g^{lim}(\beta) = \lim_n f_g^{comp}(\beta, D_1, \ldots, D_n),$$

Now note that as Σ is positive definite, $(Z_0^0, \ldots, Z_G^0)$ has positive density on $\mathbb{R}^{dG}$ and thus

$$\beta \mapsto f_0^{lim}(\beta)$$

is strictly increasing. By definition, for all $g = 1, \ldots, G$, $\beta \mapsto f_g^{lim}(\beta)$ is non-decreasing. Thus, the function

$$\beta \mapsto \sum_g f_g^{lim}(\beta)$$

is strictly increasing. Note that by construction $\{A_g^{lim} : \theta^{(g)} = \theta^{(0)}\}$ form a disjoint partition of the sample space. Thus, using the definition of f_g^{lim},

$$\lim_{\beta \to \infty} \sum_g f_g^{lim}(\beta) = 1 \text{ and } \lim_{\beta \to -\infty} \sum_g f_g^{lim}(\beta) = 0.$$

Similarly, by definition, $\beta \mapsto \sum_g f_g(\beta)$ and $\beta \mapsto \sum_g f_g^{comp}(\beta)$ are increasing with $0 \leq \sum_g f_g(\beta) \leq 1$ and $0 \leq \sum_g f_g^{comp}(\beta) \leq 1$. Invoking Polya's theorem with equation (24), $\sup_\beta |f_g(\beta) - f_g^{lim}(\beta)| \to 0$ in probability. Analogously, $\sup_\beta |f_g^{comp}(\beta, D_1, \ldots, D_n) - f_g^{lim}(\beta)|$ converges to zero in probability.

Consider a sequence $n \mapsto c_n := b^{-1}_{D_1,\ldots,D_n}(1 - \alpha/2)$ such that

$$\sum_g f_g^{comp}(c_n, D_1, \ldots, D_n) = 1 - \alpha/2.$$

Since $f_g^{comp}(\beta)$ converges to $f_g^{lim}(\beta)$ uniformly, we have that $\sum_g f_g^{lim}(c_n) \to 1 - \alpha/2$ in probability. Since $\beta \mapsto \sum_g f_g^{lim}(\beta)$ is strictly increasing and continuous, c_n converges in probability to the unique $c^0 \in \mathbb{R}$ with $\sum_g f_g^{lim}(c^0) = 1 - \alpha/2$. Thus, for all $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \leq b^{-1}_{D_1,\ldots,D_n}(1 - \alpha/2)] \\
&= \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \leq c_n] \\
&\leq \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \leq c^0 + \epsilon] + \mathbb{P}[|c_n - c^0| \geq \epsilon] \\
&= \sum_g f_g(c^0 + \epsilon) + o(1).
\end{aligned}$$

Now, letting $\epsilon \to 0$ and using that $\beta \mapsto f_g^{lim}(\beta)$ is continuous and $\sup_\beta |f_g(\beta) - f_g^{lim}(\beta)| \to 0$,

$$\lim_{n\to\infty} \sup \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)] \le \sum_g f_g^{lim}(c^0) = 1-\alpha/2. \quad (25)$$

Analogously,

$$\begin{aligned}
&\mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)] \\
&= \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le c_n] \\
&\ge \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le c^0 - \epsilon] + \mathbb{P}[|c_n - c^0| \ge \epsilon] \\
&= \sum_g f_g(c^0 - \epsilon) + o(1).
\end{aligned}$$

Thus,

$$\lim_{n\to\infty} \inf \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)] \ge \sum_g f_g^{lim}(c^0) = 1-\alpha/2. \quad (26)$$

Combining equation (25) and equation (26),

$$\lim_{n\to\infty} \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)] = 1-\alpha/2.$$

Analogously, it can be shown that

$$\lim_{n\to\infty} \mathbb{P}[\sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \ge b^{-1}_{D_1,\ldots,D_n}(\alpha/2)] = 1-\alpha/2.$$

Thus,

$$\lim_{n\to\infty} \mathbb{P}[b^{-1}_{D_1,\ldots,D_n}(\alpha/2) \le \sqrt{n}(\hat{\theta}_j^{(\bar{g})} - \theta_j^{(0)}) \le b^{-1}_{D_1,\ldots,D_n}(1-\alpha/2)] = 1-\alpha.$$

This completes the proof. □

B.5. Proof of Theorem 3

Proof. Let us first prove the result for the special case where $\delta^{(0)} = 0$. With probability exceeding $1 - \exp(-t)$,

$$\begin{aligned}
&\left|\frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \frac{n}{d}\|\delta^{(g)}\|_2^2 - \frac{1}{d}\|\tau^{(g)}\|_2^2\right| \\
&= \left|\frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)} - \delta^{(g)}\|_2^2 - \frac{1}{d}\|\tau^{(g)}\|_2^2 + \frac{2n}{d}(\hat{\theta}^{(g)} - \hat{\theta}^{(0)} - \delta^{(g)}) \cdot \delta^{(g)}\right| \\
&= \left|\frac{n}{d}\|\epsilon^{(g)} - \epsilon^{(0)}\|_2^2 - \frac{1}{d}\|\tau^{(g)}\|_2^2 + \frac{2n}{d}(\epsilon^{(g)} - \epsilon^{(0)}) \cdot \delta^{(g)}\right|
\end{aligned}$$

$$\leq s_\infty^2 C_1 \frac{t}{d} + s_\infty^2 C_1 \sqrt{\frac{t}{d}} + s_\infty C_2 \frac{b_\infty \sqrt{tn}}{\sqrt{d}},$$

for some constants C_1 and C_2. Here, we used sub-Gaussian and subexponential tail bounds, see for example Chapter 2 in Wainwright (2019). More precisely, we used that $\frac{n}{d}(\epsilon^{(g)} - \epsilon^{(0)}) \cdot \delta^{(g)}$ is sub-Gaussian with variance proxy $db_\infty^2 n \frac{4}{d^2} \max_{g,j}(\eta_j^{(g)})^2$ and that $n\|\epsilon^{(g)} - \epsilon^{(0)}\|_2^2 - (\tau^{(g)})^2$ is subexponential with parameter $64d \max_{g,j}(\eta_j^{(g)})^2$. Using a union bound, for every $\kappa > 0$ there exists a constant $C_3(\kappa)$, such that with probability exceeding $1 - \kappa$,

$$\sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \frac{n}{d}\|\delta^{(g)}\|_2^2 - \frac{1}{d}\|\tau^{(g)}\|_2^2 \right| \leq C_3 \left(s_\infty^2 \sqrt{\frac{\log G}{d}} + s_\infty^2 \frac{\log G}{d} + s_\infty b_\infty \frac{\sqrt{n \log G}}{\sqrt{d}} \right)$$

Since $\log(G)/d \leq c_1$, for all $\kappa > 0$ there exists a constant $C_4(c_1, \kappa)$ such that with probability exceeding $1 - \kappa/2$,

$$\sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \frac{n}{d}\|\delta^{(g)}\|_2^2 - \frac{1}{d}\|\tau^{(g)}\|_2^2 \right| \leq C_4 \left(s_\infty^2 \sqrt{\frac{\log G}{d}} + s_\infty b_\infty \frac{\sqrt{n \log G}}{\sqrt{d}} \right) \tag{27}$$

Analogously, it can be shown that there exists a constant C_5 such that with probability exceeding $1 - \kappa/2$,

$$\begin{aligned} &\sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2 - \frac{n}{d}\|\delta^{(g)}\|_2^2 - \frac{1}{d}\|\sigma^{(g)}\|_2^2 \right| \\ &= \sup_g \left| \frac{n}{d}\|\epsilon^{(g)}\|_2^2 - \frac{1}{d}\|\sigma^{(g)}\|_2^2 + \frac{2n}{d}\epsilon^{(g)} \cdot \delta^{(g)} \right| \\ &\leq C_5 \left(s_\infty^2 \sqrt{\frac{\log G}{d}} + s_\infty b_\infty \frac{\sqrt{n \log G}}{\sqrt{d}} \right). \end{aligned} \tag{28}$$

Combining equation (27) and equation (28), with probability $1 - \kappa$,

$$\begin{aligned} &\sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2 - \frac{n}{d}\hat{R}^{\text{mod}}(g) \right| \\ &= \sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2 - \max\left(\frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 - \frac{1}{d}\sum_{j=1}^d (\hat{\tau}_j^{(g)})^2, 0 \right) - \frac{1}{d}\sum_{j=1}^d (\hat{\sigma}_j^{(g)})^2 \right| \\ &\leq 2\iota_{n,d} + \sup_g \left| \frac{n}{d}\|\hat{\theta}^{(g)} - \theta^{(0)}\|_2^2 - \max\left(\frac{n}{d}\|\hat{\theta}^{(g)} - \hat{\theta}^{(0)}\|_2^2 \right.\right. \\ &\qquad \left.\left. - \frac{1}{d}\sum_{j=1}^d (\tau_j^{(g)})^2, 0 \right) - \frac{1}{d}\sum_{j=1}^d (\sigma_j^{(g)})^2 \right| \end{aligned}$$

$$\leq 2\iota_{n,d}+C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right)+\sup_g\left|\frac{1}{d}\sum_{j=1}^d(\sigma_j^{(g)})^2+\frac{n}{d}\|\theta^{(g)}-\theta^{(0)}\|_2^2\right.$$
$$\left.-\max\left(\frac{n}{d}\|\theta^{(g)}-\theta^{(0)}\|_2^2+\frac{1}{d}\sum_{j=1}^d(\tau_j^{(g)})^2-(\tau_j^{(g)})^2,0\right)-\frac{1}{d}\sum_{j=1}^d(\sigma_j^{(g)})^2\right|$$
$$\leq 2\iota_{n,d}+C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right),$$

for some constant C_6 that depends on c_1 and κ. Thus, on that event,

$$\left|\min_g\frac{n}{d}\|\hat\theta^{(g)}-\theta^{(0)}\|_2^2-\frac{n}{d}\min_g\hat R^{\mathrm{mod}}(g)\right|$$
$$\leq 2C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right)+4\iota_{n,d}.$$

Furthermore, on that event,

$$\left|\frac{n}{d}\hat R^{\mathrm{mod}}(\bar g)-\frac{n}{d}\|\hat\theta^{(\bar g)}-\theta^{(0)}\|_2^2\right|\leq C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right)+2\iota_{n,d}.$$

Thus,

$$\frac{n}{d}\|\hat\theta^{(\bar g)}-\theta^{(0)}\|_2^2$$
$$\leq\min_g\frac{n}{d}\|\hat\theta^{(g)}-\theta^{(0)}\|_2^2+3C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right)+6\iota_{n,d}.$$

Here, we used that $\min_g\hat R^{\mathrm{mod}}(g)=\hat R^{\mathrm{mod}}(\bar g)$. This proves the special case with $\delta^{(0)}=0$. For the more general case, we have just shown that

$$\frac{n}{d}\|\hat\theta^{(\bar g)}-\theta^{(0)}-\delta^{(0)}\|_2^2$$
$$\leq\min_g\frac{n}{d}\|\hat\theta^{(g)}-\theta^{(0)}-\delta^{(0)}\|_2^2+3C_6\left(s_\infty^2\sqrt{\frac{\log G}{d}}+s_\infty b_\infty\frac{\sqrt{n\log G}}{\sqrt{d}}\right)+6\iota_{n,d}. \tag{29}$$

Let $C''=\max_g\|\hat\theta^{(g)}-\theta^{(0)}\|_2+\|\delta^{(0)}\|_2$. Using that

$$\|\hat\theta^{(\bar g)}-\theta^{(0)}\|_2^2\leq\|\hat\theta^{(\bar g)}-\theta^{(0)}-\delta^{(0)}\|_2^2+C''\|\delta^{(0)}\|_2,$$

and

$$\|\hat\theta^{(g)}-\theta^{(0)}-\delta^{(0)}\|_2^2\leq\|\hat\theta^{(g)}-\theta^{(0)}\|_2^2+C''\|\delta^{(0)}\|_2$$

in equation (29) completes the proof. □

Acknowledgments

The author would like to thank Guillaume Basse, Guido Imbens, and Bin Yu for inspiring discussions. The author gratefully acknowledges support by the Dieter Schwarz Foundation, by the Chambers Foundation, and by David Huntington.

References

H. Akaike. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer, 1974. MR0423716

J. Angrist, G. Imbens, and D. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 2010. MR2602303

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 2016. MR3531135

S. Athey, R. Chetty, G. W. Imbens, and H. Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. MR3099122

R. Bowden and D. Turkington. *Instrumental variables*. Cambridge University Press, 1990. MR1113481

M. Brookhart and M. Van Der Laan. A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics & Data Analysis*, 50(2):475–498, 2006. MR2201874

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018. MR3769544

G. Claeskens and N. Hjort. The focused information criterion. *Journal of the American Statistical Association*, 2003. MR2041482

G. Claeskens and N. Hjort. *Model selection and model averaging*. Cambridge University Press, 2008. MR2431297

R. Crump, V. Hotz, G. Imbens, and O. Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.

Y. Cui and E. Tchetgen-Tchetgen. Selective machine learning of doubly robust functionals. *Biometrika*, 2024. MR4745579

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. 2014.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, 2001. MR1851606

R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995. MR1354456

F. Guo and E. Perković. Efficient least squares for estimating total effects under linearity and causal sufficiency. *Journal of Machine Learning Research*, 2022. MR4576689

M. Hernan and J. Robins. *Causal inference: What If*. Chapman & Hall, 2020.

P. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967. MR0216620

Z. M. Hussain, M. Oberst, M.-C. Shih, and D. Sontag. Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, 35:6161–6174, 2022.

S. Hyun, M. G'sell, and R. J. Tibshirani. Exact post-selection inference for the generalized lasso path. 2018. MR3777139

G. Imbens and J. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

M. Jakobsen and J. Peters. Distributional robustness of k-class estimators and the PULSE. *The Econometrics Journal*, 2022. MR4423408

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

A. Kapelner, J. Bleich, A. Levine, Z. D. C., R. DeRubeis, and R. Berk. Inference for the effectiveness of personalized medicine with software. *arXiv preprint arXiv:1404.7844*, 2014.

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes' estimates. *Univ. Calif. Publ. in Statist.*, 1:277–330, 1953. MR0054913

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 2016. MR3485948

H. Leeb and B. M. Pötscher. The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142, 2003. MR1965844

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005. MR2153856

H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? 2006. MR2291510

O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, pages 929–947, 1997. MR1447734

O. Lepskii. Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1991. MR1147167

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013. MR3086420

J. R. Loftus and J. E. Taylor. Selective inference in regression models with groups of variables. 2015.

A. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, pages 575–595, 1959. MR0120730

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2021. MR4259133

S. Powers, J. Qian, K. Jung, A. Schuler, N. Shah, T. Hastie, and R. Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 2018. MR3799840

J. Robins, A. Rotnitzky, and L. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 1994. MR1294730

C. Rolling and Y. Yang. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B*, 2014. MR3248675

P. Rosenbaum and D. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B*, 45(2):212–218, 1983.

E. T. Rosenman, G. Basse, A. B. Owen, and M. Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 2020. MR4680697

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *Journal of the Royal Statistical Society Series B*, 2021. MR4250274

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

A. Schuler, M. Baiocchi, R. Tibshirani, and N. Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2), 1978. MR0468014

J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993. MR1224373

J. Splawa-Neyman, D. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. *Statistical Science*, pages 465–472, 1990.

J. Taylor and R. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. MR3371123

H. Theil. *Economic forecasts and policy*. North-Holland Pub. Co., 1961.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007. MR2233926

M. Van der Laan and J. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003. MR1958123

M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011. MR2867111

A. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000. MR1652247

M. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019. MR3967104

P. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

F. Yang, R. Foygel Barber, P. Jain, and J. Lafferty. Selective inference for group-sparse linear models. *Advances in neural information processing systems*, 29, 2016.

Y. Yang. Consistency of cross validation for comparing regression procedures. 2007. MR2382654

Z. Zhan and Y. Yang. Profile electoral college cross-validation. *Information Sciences*, 586:24–40, 2022.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014. MR3153940

Y. Zhang and Y. Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015. MR3347297

Y. Zhao, X. Fang, and D. Simchi-Levi. Uplift modeling with multiple treatments and general response types. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017.