Contributions to Evolutionary Dynamics and Causal Inference

Contributions to Evolutionary Dynamics and Causal Inference

a dissertation presented
by
Lin L. Liu
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
April 2018

Dissertation advisors: Professor Franziska Michor; Professor James M. Robins Lin L. Liu

Contributions to Evolutionary Dynamics and Causal Inference

Abstract

In this dissertation, we investigate topics in two different quantitative disciplines, both of which have profound impact in biomedical sciences. The first area is evolutionary dynamical systems to model biological systems; the second area is causal inference, with the primary goal of drawing causal conclusions from experimental and observational studies.

In Chapter 1, we investigate the dynamical behavior of reprogramming of somatic cells to induced pluripotent stem cells (iPSCs). In order to define a unified framework to study and compare the dynamics of reprogramming under different conditions, we developed an *in silico* analysis platform based on evolutionary modeling. Our approach takes into account the variability in experimental results stemming from probabilistic growth and death of cells and potentially heterogeneous reprogramming rates. We found that reprogramming driven by the Yamanaka factors alone is a more heterogeneous process possibly due to cell-specific reprogramming rates, which can be homogenized by the addition of additional factors.

In Chapter 2, we study the problem of data-driven confounder selection in causal inference. The recently proposed Collaborative Targeted Minimum Loss Estimation (CTMLE) provides a framework of constructing doubly-robust estimators by selecting appropriate covariates into the propensity score model. We focus on the asymptotic (large-sample) theory of CTMLE, together with some other alternatives such as Focused Information Criterion (FIC) and the Lepski's method (or equivalently, hypothesis testing). The algebraic connections among these selection statistics are presented.

In Chapter 3, we investigate some practical issues in the application of higher-order influence functions (HOIFs) in the problem semi-/non-parametric functional estimation, which is closely connected to literature of estimating causal estimand of interest with modern machine learning technique. We discuss several ideas of stabilizing the finite-sample performance of HOIF-based estimators and demonstrate the superiority of these modifications to the original construction of HOIF-based estimator in simulation studies.

Contents

Acknowledgments

I owe my deepest gratitude to my thesis advisors, Professor Franziska Michor and Professor James Robins, for both teaching me how to become a good researcher and more importantly a better person. In particular, I want to thank Franziska for bringing me into the world of evolutionary dynamics and branching processes. As a biology major in college, I was always fascinated by the beauty and principle in evolution, but using mathematics to study evolution never came to my mind until I had the opportunity to work with Franziska, one of the best minds in this field. More importantly, I will be deeply influenced by Franziska because she always prioritizes her research to help cancer patients instead of only being a theoretician to just do theory. I want to thank Jamie for bringing me to the world of causal inference, a field blended with mathematics, statistics, and philosophy. The course I took from Jamie is life-changing because how mathematically elegant and beautiful single world intervention graph (SWIG) is. Without Jamie, I might have never been able to discover my ultimate love to mathematical statistics. More importantly, learning the way that Jamie approaches a research problem is probably my most important lesson in graduate school – starting from the simple examples, finding counterexamples, and trying to understand a problem from the first principle.

I would also like to thank Professor Lorenzo Trippa for being an extremely helpful thesis committee member and an amazing instructor for my very first course in advanced mathematical/statistical topics, Professor Alkes Price for being a tremendous instructor in population genetics and his selfless support, Professor Winston Hide for his tips on how to succeed in graduate school.

I want to mention some special thanks to several of my classmates in the Department of Biostatistics – Boyu Ren, Yuanyuan Shen and Fei Li for being my friends from day one. I also want to thank four stu-

dent alumni in the department – Roland Matsouaka for being a wonderful student mentor in my first year, Abhishek Chakrabortty and Rajarshi Mukherjee for really inspiring me to work on theoretical problems because of their amazing teaching skills in materials related to mathematical statistics and probability. I would also like to thank my collaborators, mentors and Michor lab members throughout my graduate school: Yong Zhang and Shirley Liu for being the first two people ever encouraging me to become a scientist; Subhajyoti De for our collaboration and for giving me many advices on how to identify important project and planning a research career; Philipp Altrock for the joyful collaboration on our review paper on mathematical oncology and teaching me the basic evolutionary game theory; Kornelia Polyak and Michalina Janiszewska for our collaboration on breast cancer data; Justin Brumbaugh, Zachary Smith, Alex Meissner and Konrad Hochedlinger for teaching me stem cell biology; Hua-Jun Wu, Jiantao Shi and Qiong Xu for all the helps in both my personal life and my research; Ollie McDonald for his proofreading on my first thesis paper and all the suggestions on my presentation skills; Shaon Chakrabarty for being an amazing colleague from whom I learnt many aspects of physics which I would otherwise have never known; Kimiyo Yamamoto, Akira Nakamura and Hiroshi Haeno for our collaboration on pancreatic cancer projects; and finally Yiwen Chen, Feng Zhou, Tianlei Xu, and Sheng'en Hu for our endless discussion on computational biology and life.

This dissertation is also dedicated to my family and friends back in China. My parents, Hongfang Hao and Baofeng Liu, for the best upbringing that they could possibly give. My grandparents, Xiujin Fu, Jinyi Hao, Fengge Yin, and Fuzhen Liu, for raising me up while my parents were busy at work. Nan Wang, Xiaopeng Cai, Kai Fu, Kang Gao, Wei Guo, Kailong Tommy He, Hailan Jin, Jiaheng Li, Congcong Li, Long Ma, Xu Wu, Qianlong Xie, Guangyou Xu, Liang Yang, You Zhou, Meng Zhu and many other friends for supporting me selflessly throughout these years. Fanyi Meng, who passed away in a devastating accident when he was on his way to collecting geographical data for his research project,

for being my brother from another mother and motivating me to put as much effort as I can in my life and work.

Last but not least, this dissertation would not have been possible without Ms. Katherine Yue Li, for giving me all her love and support, tolerating me to a fault when I am busy with research and making me a better person.

1

Probabilistic modeling of reprogramming to induced pluripotent stem cells[49]

1.1 Introduction

Somatic cells can be experimentally reprogrammed into induced pluripotent stem cells (iPSCs) through overexpression of the four transcription factors Oct3/4, Sox2, Klf4, and c-Myc (OSKM)[92,93,111]. The reprogramming process usually takes weeks, yielding iPSCs at extremely low efficiency[32,33,66,92,93,111].

Several efforts have improved the efficiency of the reprogramming process; for example, Hanna et al.[32] reported that inhibition of the p53/p21 pathway or overexpression of Lin28 resulted in acceleration of reprogramming by increasing cell proliferation, whereas Nanog overexpression improved reprogramming in a cell-division independent manner. Subsequently, reduction of the methyl-binding protein Mbd3 during reprogramming was also shown to ensure that almost all responding somatic lineages form iPSCs within 8 days, consistent with a deterministic process[66]. Similarly, another study argued that a subset of "privileged" somatic cells appear to acquire pluripotency in a deterministic manner, indicating a latent intrinsic heterogeneity within the starting population either prior to or following OSKM induction[30]. Induction of C/EBPα in B-cells expressing OSKM provides another approach to activate the Oct4-GFP transgene in the majority of responding cells within a few days[15]. Most recently, two different studies optimized extrinsic conditions that facilitate iPSC formation from somatic progenitor cells within one week, thus avoiding the need for additional genetic manipulation[5,109]. For example, exposing somatic cells expressing OSKM to ascorbic acid and a GSK3-β inhibitor (AGi) was demonstrated to result in synchronous and rapid reprogramming[5].

Mathematical modeling has been a valuable approach to better understand the reprogramming process. For example, Hanna et al.[32] used a simple death process model to explain the dynamics under different conditions of reprogramming. Cell cycle modeling previously used to describe isotype switching in immune system development, in particular B-cell development and lineage commitment[19], can also provide a good fit to experimental data in the induced reprogramming setting using Mbd3 knockdown[66]. In conditions using OSKM overexpression only, however, neither the cell-cycle model nor a model assuming deterministic reprogramming can explain the complex lineage histories that lead to iPSCs[66]. Alternatively, the iPSC dynamics can be explained with a phase-type model[66], assuming a finite number of intermediate phases between the initial somatic cell and the final iPSC state. In this type

of model, the number of parameters linearly depends on the number of phases and their values are difficult to select using underlying biological knowledge; this model also ignored the effects of proliferation and apoptosis of different cell types on the population dynamics. However, it is difficult to interpret the number of phases inferred from this type of model and more difficult to verify such result experimentally. Lastly, from a statistical physics perspective, Fokker-Planck equations were also employed to construct the probability density function of the latency time to reprogramming, and then an inverse problem was solved to estimate the parameters from experimental data[52]. Though these predictions led to a good fit to the data with out-of-sample validation, the choice of the functional form for the potential is quite ad hoc and not subject to experimental validation based on currently available technology.

The framework of continuous-time birth-death processes[59] provides an alternative perspective to describe cellular reprogramming, including essential elements of the dynamics such as cell growth, death, and cell fate change (i.e. transition). This modeling framework has also been widely applied in other biological system, such as cancer[23,50,1]. One advantage of the birth-death-transition process approach is that it appreciates probabilistic effects of division, death, and reprogramming on the final outcome, either represented by the distribution of first passage times or the percentage of iPSCs at a certain time point. Another advantage is that the birth-death-transition process helps us better understand the sources of the variation observed from the data. Here we designed a generalizable probabilistic model with simple and explicit interpretations of all parameters to explore alternative explanations of the dynamics of reprogramming. Using this approach, we explicitly modeled reprogramming dynamics to analyze the cell dynamic data from different experimental setups. We first utilized cell proliferation data from Bar-Nur et al.[5] to parameterize the probabilistic model. We found that the use of a low and heterogeneous reprogramming rate, in the context of our mathematical model, explained the OSKM data, while a high and homogeneous reprogramming rate recapitulated the OSKM + AGi results. Data from other

Table 1.1: Experimental data analyzed

Resource	Data Type	Biomarker
Bar-Nur et al.[5]	% Oct4-GFP+ cells/well (GMP)	Oct4-GFP reporter
Vidal et al.[109]	% Oct4-GFP cells/colony (MEF)	Oct4-GFP reporter
Rais et al.[66]	% Nanog-GFP+ wells (MEF)	Nanog-GFP reporter and mCherry marker
Hanna et al.[32]	% Nanog-GFP+ wells (B Cells)	Nanog-GFP reporter
Smith et al.[87]	% Cell counts for each colony (MEF)	Stained for Nanog, E-Cadherin, and Alkaline Phosphatase

sources[66,109] were then used to further validate our approach and test its ability to also recapitulate early phase reprogramming dynamics[32,87]. A summary of the data used in this paper is listed in Table 1.1. Our approach allows quantification of reprogramming dynamics using the widely variable experimental setups of different studies (Table 1.1). For example, Rais et al.[66] collected data on the first passage time of the percentages of Oct4-GFP signal in each well surpassing some threshold, whereas Bar-Nur et al.[5] recorded the percentages of Oct4-GFP-positive cells in each well at several time points. In order to obtain as much information as possible from these types of experiments, we recommend collecting the full time course of the reprogramming signal instead of the first passage time only.

Our flexible approach provides a theoretical framework for describing cellular reprogramming under any condition. Importantly, it also establishes a quantitative method to compare between reprogramming systems. From a practical perspective, our modeling approach provides a platform to determine both the rate and homogeneity of any given cell fate conversion. Quantitative assessment of these parameters is particularly important for large-scale mechanistic studies that demand large cell numbers or for the design of differentiation protocols generating therapeutic cell types. For example, global transcriptomic or proteomic analyses often require bulk cell culture; our modeling approach could be used to identify reprogramming systems or time points well suited for these applications based on the reprogramming rate and its uniformity. Alternatively, such a model could be employed as an empirical standard to quan-

tify the uniformity and kinetics of any given cell fate conversion under different conditions to optimize improved protocols or understand the contributions of specific growth factors. Thus, in addition to the more fundamental modeling role, we anticipate that our approach will be useful for mapping the precise molecular trajectories of somatic cells acquiring pluripotency and for identifying novel reprogramming intermediates.

1.2 Induced reprogramming can be modeled as a two-type continuous-time Markov process

We began to explore the kinetics of iPSC generation by analyzing previous data obtained from a doxycycline inducible, polycistronic reprogramming system[5]. In this study, granulocyte-macrophage progenitors (GMPs) were exposed to doxycycline for varying time periods before scoring for activation of an OCT4-GFP reporter[5]. Using this data set, we designed a two-type probabilistic logistic birth-death-transition process with a carrying capacity to model the dynamics of cellular reprogramming. Such a process describes the growth and death of individual cells while the population as a whole initially expands exponentially but then reaches a maximum cell number, the "carrying capacity" due to the resource limitation of the in vitro cell culture system. In this model, we ignore any spatial interactions between different cells[65]. The population of cells is composed of two different cell types – somatic cells and iPSCs, whose numbers at time are denoted by $X_1(t)$ and $X_2(t)$, respectively. Initially, somatic cells and iPSCs proliferate with rates λ_1 and λ_2 and die with rates φ_1 and φ_2 per day per cell, respectively, when population sizes are sufficiently small such that they are not yet impacted by the carrying capacity. The maximum total number of cells for each well is M, i.e. $X_1(t) + X_2(t) \leq M$ if the culture is not split after the exponential growth phase. Therefore, as the population of cells increases, the growth pattern of cells depreciates according to the logistic function. The reprogramming rate from somatic cells into iPSCs is given by γ per day per cell. In one infinitesimally small time interval, only the following events can

occur: one somatic cell may divide or die, one iPSC may divide or die, or one somatic cell may transition to one iPSC; all other events have very small probabilities of occurrence. Without a carrying capacity, the numbers of cells at day 8 in the OSKM + AGi and at day 12 in the OSKM conditions are predicted to be much larger than M, which is inconsistent with experimental results; therefore a carrying capacity was included in the model. All results considering carrying capacity shown in the main text are based on $M = 100,000$, but sensitivity analyses demonstrated that perturbations of this and other parameters did not significantly change the dynamics. Our probabilistic model explicitly distinguishes the effects of cell growth, death and fate change on the reprogramming dynamics.

Using this approach, we then aimed to predict the percentage of iPSCs at a certain time t. We approximated the expected proportion of iPSCs at a certain time point t as $\mathbb{E}[X_2(t)/(X_1(t)+X_2(t))] \approx \mathbb{E}[X_2(t)]/(\mathbb{E}[X_1(t)]+\mathbb{E}[X_2(t)])+r(\mathbb{E}[X_1(t)],\mathbb{E}[X_2(t)])$ followed from multivariate Taylor expansion, where the form of $r(\mathbb{E}[X_1(t)],\mathbb{E}[X_2(t)])$ can be found in (1.9). Using the probability generating function (PGF) for the process, we obtained a system of two coupled first order ordinary differential equations for the following quantities: $\kappa_1(t) = \mathbb{E}[X_1(t)], \kappa_2(t) = \mathbb{E}[X_2(t)], \kappa_3(t) = \mathbb{E}[X_1(t)^2], \kappa_4(t) = \mathbb{E}[X_2(t)^2], \kappa_5(t) = \mathbb{E}[X_1(t)X_2(t)]$ (see the SI for details and derivations). We then obtain:

$$\begin{aligned} \frac{d\kappa_1(t)}{dt} &= (\lambda_1 - \varphi_1 - \gamma)\kappa_1(t) - \frac{\lambda_1}{M}(\kappa_3(t) + \kappa_4(t)), \\ \frac{d\kappa_2(t)}{dt} &= \gamma\kappa_1(t) + (\lambda_2 - \varphi_2)\kappa_2(t) - \frac{\lambda_2}{M}(\kappa_4(t) + \kappa_5(t)). \end{aligned} \tag{1.1}$$

where at time $t = 0$ (i.e., the start of the experiment), we have initial conditions $\kappa_1(0) = 1, \kappa_2(0) = 0, \kappa_3(0) = 1, \kappa_4(0) = 0, \kappa_5(0) = 0$. This system of differential equations was approximated using the moment closure approximation[55,56] followed by Euler's method to solve the approximate system of differential equations numerically[86]; the complete formula for this system of differential equations involving higher-order moments as well as the R code for solving such systems can be found in Liu

et al.[49]. To demonstrate the utility of this analytical approximation and the numerical method, we examined the consistency between the analytical approximation and exact numerical computer simulations of the process and concluded that the analytical approximation is sufficiently accurate to be used in our setting. The utility of this approximation is to aid in our parameter estimation procedure (Section 1.3.4). Unfortunately, no approximation of the variance of the iPSC proportion is available, and therefore this quantity was investigated entirely based on computer simulations.

1.3 A two-type stochastic logistic process model for reprogramming dynamics

1.3.1 Notations

We designed a stochastic model to predict the dynamics of $\{Y(t) = (X_1(t), X_2(t))\}$, where $X_1(t)$ denotes the number of somatic cells (in our case, granulocyte-macrophage progenitors, GMPs) at time t, and $X_2(t)$ denotes the number of induced pluripotent stem cells (iPSCs) at time t. We define an invariant mapping $g : (X_1(t), X_2(t)) \to (S(t), prop(t))$ through $S(t) = X_1(t) + X_2(t)$ and $prop(t) = \frac{X_2(t)}{S(t)}$ so that our model prediction corresponds to a readout from the experimental data, specifically the percentage of Oct4-GFP+ wells. Denote the percentage of Oct4-GFP+ wells at time t by $prop(t)$. Our model considers a carrying capacity constraint M such that $S(t) \leq M \, \forall t$, where $t \in \mathbb{R}^+ \cup \{0\}$. Though outside the scope of this work, this assumption can be relaxed[14]. We designed the underlying dynamical process as a two-species stochastic version of the Verhulst logistic growth model by extending the model in one dimension proposed by Tan & Piantadosi[94]. Within this stochastic logistic process, the parameters λ_1 and λ_2 denote the initial proliferation rates per day per cell for GMPs and iPSCs, respectively (i.e., the "cell-intrinsic" rates for population sizes sufficiently small such that they are not yet impacted by the carrying capacity). The parameters φ_1 and φ_2 denote the apoptosis rates per day for GMPs and iPSCs, respectively, and γ represents the reprogramming rate from GMPs to iPSCs per day. The parameters can be arbitrary

functions of time t. We assume that only the proliferation rates λ are affected by the presence of the carrying capacity. Note that in derivations in later sections, we treat all these cell-intrinsic rate parameters such as the proliferation rate, apoptosis rate and dedifferentiation rate as time-dependent variables to make our derivations more general, since the system with time-independent variables represents a special case of the system with time-dependent variables. Moreover, in Result II of the main text, we suggest that a random reprogramming rate might explain the variability observed in the experimental data.

1.3.2 A one-type stochastic logistic process

We first review the stochastic logistic process in the one-type case[94] as a building block for our two-dimensional extension. This process is also essential for estimating the proliferation and apoptosis rates for GMPs in both growth conditions. The process $\{X(t), t \geq 0\}$ is a Markov process. Given j individuals at time t, the infinitesimal transition probabilities at time $t + \Delta t$ are given by

$$
\begin{aligned}
&P(X(t+\Delta t) = n + j | X(t) = n) \\
&= \begin{cases}
\lambda(t) n (1 - \frac{n}{M}) \Delta t + o(\Delta t) & \text{if } j = 1, \\
\varphi(t) n \Delta t + o(\Delta t) & \text{if } j = -1, \\
1 - \lambda(t) n (1 - \frac{n}{M}) \Delta t - \varphi(t) n \Delta t + o(\Delta t) & \text{if } j = 0, \\
o(\Delta t) & \text{else,}
\end{cases}
\end{aligned}
\tag{1.2}
$$

where $\lambda(t), \varphi(t) > 0 \; \forall t \geq 0, j \leq M$, and $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$ for all $j = 0, 1, \cdots, M$. Then the master equation governing the probability mass function of $X(t)$ for the one-type stochastic logistic process is[95]:

$$
\begin{aligned}
&\frac{\partial P(X(t) = n, t)}{\partial t} \\
&= P(X(t) = n-1, t) \cdot \lambda(t) \cdot (n-1) \cdot \left(1 - \frac{n-1}{M}\right)
\end{aligned}
\tag{1.3}
$$

$$
\begin{aligned}
&+ P(X(t) = n+1, t) \cdot \varphi(t) \cdot (n+1) \\
&- P(X(t) = n, t) \cdot \left\{ \ \lambda(t) \cdot n \cdot \left(1 - \frac{n}{M}\right) + \varphi(t) \cdot n \ \right\}.
\end{aligned}
$$

Define $Q(u,z;t_0,t_1) = \sum_{v=0}^{M} z^v P_{uv}(t_0,t_1)$ to be the probability generating function (PGF) of $X(t_1)$ given $X(t_0) = u$, where $t_1 > t_0$ and $P_{uv}(t_0,t_1) = Pr\{X(t_1) = v | X(t_0) = u\}$ as a short-hand notation. The equation for PGF derived from the master equation[94] is given by

$$
\frac{\partial}{\partial t_1} Q(u,z;t_0,t_1) = (z-1)\cdot\left[\left\{z\cdot\left(1-\frac{1}{M}\right)\lambda(t)-\varphi(t)\right\}\cdot\frac{\partial}{\partial z}Q(u,z;t_0,t_1)-z\left\{\frac{z}{M}\lambda(t)-\varphi(t)\right\}\cdot\frac{\partial^2}{\partial z^2}Q(u,z;t_0,t_1)\right] \tag{1.4}
$$

Using 1.4, the differential equations for the first and second moments are given by

$$
\begin{aligned}
\frac{\partial}{\partial t}\kappa^{(1)} &= (\lambda(t) - \varphi(t))\kappa^{(1)} + \frac{\lambda(t)}{M}\kappa^{(2)}, \\
\frac{\partial}{\partial t}\kappa^{(2)} &= (\lambda(t) + \varphi(t))\kappa^{(1)} + \left\{2(\lambda(t) - \varphi(t)) - \frac{\lambda(t)}{M}\right\}\kappa^{(2)} - \frac{2\lambda(t)}{M}\kappa^{(3)},
\end{aligned} \tag{1.5}
$$

where $\kappa^{(k)} = \mathbb{E}[X^k]$. This system of two coupled differential equations is not solvable, but the first and second moments can be analytically approximated using the moment closure approximation[55,56] by setting the higher centered moments as zero. The system of differential equations can be solved using Euler's method[86].

Although throughout the paper, we assume a carrying capacity in the sense that $X(t) \leq M \, \forall t$, where $t \in \mathbb{R}^+ \cup \{0\}$, this limitation can be relaxed by replacing the carrying capacity penalization $1 - 1/M$ by other functional forms, for instance e^{-M}, where $1 - 1/M$ can be interpreted as the first order Taylor series expansion of e^{-M}. It can be further generalized to e^{-cM} for different carrying capacity constraints.

1.3.3 A two-type stochastic logistic process

Describing two-type stochastic logistic process

Two-type stochastic logistic process. Our two-type stochastic logistic process is a continuous-time Markovian process – suggesting that (1) events can happen at any point in time (i.e., continuous time) and (2) the future state of the system is independent of the past when conditioning on the present (i.e. Markovian property). Our model contains two types of cells (somatic cells and iPSC), and each can divide and die with a certain proliferation and apoptosis rate, respectively. Furthermore, a somatic cell state can transition to an iPSC state, which then cannot change back into a somatic cell. The two-type stochastic logistic process is defined using infinitesimal transition probabilities. At time t, with somatic cells and iPSCs in the system, the following possible events may occur during the next infinitesimally small time interval:

1. With probability $\lambda_* \cdot (1 - (X_1(t) + X_2(t))/M) \cdot X_*(t) \cdot \Delta t + o(\Delta t)$, one of the type-$*$ cells (where $*$ refers to either somatic cells or iPSCs) divides into two, where λ_* is the per-cell intrinsic proliferation rate when population sizes are sufficiently small such that they are not yet impacted by the carrying capacity. If the number of somatic cells is large then the probability of one somatic cell dividing is also large and this probability increases if the time interval becomes longer. The term $(1-(X_1(t)+X_2(t))/M)$ penalizes the proliferation dynamics such that the total number of somatic cells and iPSCs does not exceed M. The term $o(\Delta t)$ is an extremely small quantity compared to Δt;

2. With probability $\varphi_* \cdot X_*(t) \cdot \Delta t + o(\Delta t)$, one of the type-$*$ cells dies and the population size decreases by one;

3. With probability $\gamma \cdot X_1(t) \cdot \Delta t + o(\Delta t)$, one of the somatic cells transitions to an iPSC and the size of the population stays constant;

4. The probability of no events in next Δt time interval is the complement of the sum of the above

probabilities;

5. The probability of all other possible events is of much smaller order than Δt.

With the infinitesimal transition probabilities outlined above as a building block, we can derive important quantities such as the master equation, the probability-generating function, moment-generating function, sojourn time, and others[95]. Note that all rate parameters can in principle be time-dependent and random variables instead of constants.

Technical derivations

We extended the one-dimensional probabilistic logistic model to the two-type case using the notations introduced in 1.3.1. Although extensions to a larger number of types, i.e. m-type ($m \in \mathbb{Z}_{>0}$) generalized birth-death processes is possible, we decided not to further pursue such a generalization in this paper in order to not deviate too far from the subject matter. This process is defined by the following infinitesimal transition probabilities:

$$P(\boldsymbol{X}(t+\Delta t) = (n+j, m+k) | \boldsymbol{X}(t) = (n,m)) \tag{1.6}$$

$$= \begin{cases} \lambda_1(t) n \left(1 - \dfrac{n+m}{M}\right) \Delta t + o(\Delta t) & \text{if } j = 1, k = 0, \\ \varphi_1(t) n \Delta t + o(\Delta t) & \text{if } j = -1, k = 0, \\ \lambda_2(t) m \left(1 - \dfrac{n+m}{M}\right) \Delta t + o(\Delta t) & \text{if } j = 0, k = 1, \\ \varphi_2(t) m + o(\Delta t) & \text{if } j = 0, k = -1, \\ \gamma(t) n + o(\Delta t) & \text{if } j = -1, k = 1, \\ 1 - (\lambda_1(t) n + \lambda_2(t) m) \left(1 - \dfrac{n+m}{M}\right) \Delta t - (\varphi_1(t) n + \varphi_2(t) m) \Delta t - \gamma(t) n \Delta t + o(\Delta t) & \text{if } j = 0, k = 0, \\ o(\Delta t) & \text{else,} \end{cases}$$

where $\lambda_1(t), \lambda_2(t), \varphi_1(t), \varphi_2(t) \geq 0$, $n+m \leq M$ and $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$ for all $m, n = 0, 1, \cdots, M$. Then the master equation governing the joint probability mass function of $\mathbf{X}(t)$ for the two-type stochastic logistic process is[95]:

$$\begin{aligned}
&\frac{\partial P(X(t)=(n,m),t)}{\partial t} \\
&= P(X(t)=(n-1,m),t)\cdot\lambda_1(t)\cdot(n-1)\cdot\left(1-\frac{n-1+m}{M}\right) \\
&\quad + P(X(t)=(n,m-1),t)\cdot\lambda_2(t)\cdot(m-1)\cdot\left(1-\frac{n+m-1}{M}\right) \\
&\quad + P(X(t)=(n+1,m),t)\cdot\varphi_1(t)\cdot(n+1) + P(X(t)=(n,m+1),t)\cdot\varphi_2(t)\cdot(m+1) \\
&\quad + P(X(t)=(n-1,m+1),t)\cdot\gamma(t)\cdot(n-1) \\
&\quad - P(X(t)=(n,m),t)\cdot\left\{\begin{array}{c}\lambda_1(t)\cdot n\cdot\left(1-\frac{n+m}{M}\right)+\lambda_2(t)\cdot m\cdot\left(1-\frac{n+m}{M}\right)\\ +\varphi_1(t)\cdot n+\varphi_2(t)\cdot m+\gamma(t)\cdot n\end{array}\right\}.
\end{aligned} \tag{1.7}$$

Define $Q(u_1,u_2,z_1,z_2;t_0,t_1) = \sum_{v_1,v_2=0}^{v_1+v_2\leq M} z_1^{v_1} z_2^{v_2} P_{(u_1,u_2),(v_1,v_2)}(t_0,t_1)$ to be the probability generating function (PGF) of $\mathbf{X}(t_1) = (v_1,v_2)$ given $\mathbf{X}(t_0) = (u_1,u_2)$, where $t_1 > t_0$ and $P_{(u_1,u_2),(v_1,v_2)}(t_0,t_1) = Pr\{\mathbf{X}(t_1)=(v_1,v_2)|\mathbf{X}(t_0)=(u_1,u_2)\}$ as a short-hand notation.

We then determined the probability generating function using the Kolmogorov forward equation approach[95]:

$$\begin{aligned}
&\frac{\partial}{\partial t}Q(u_1,u_2,z_1,z_2;t_0,t_1) \\
&= \left[(z_1-1)\left\{z_1\lambda_1(t)\left(1-\frac{1}{M}\right)-\varphi_1(t)\right\}-(z_1-z_2)\gamma(t)\right]\cdot\frac{\partial}{\partial z_1}Q(u_1,u_2,z_1,z_2;t_0,t_1) \\
&\quad + \left[(z_2-1)\left\{z_2\lambda_2(t)\left(1-\frac{1}{M}\right)-\varphi_2(t)\right\}\right]\cdot\frac{\partial}{\partial z_2}Q(u_1,u_2,z_1,z_2;t_0,t_1)
\end{aligned} \tag{1.8}$$

$$
\begin{aligned}
&- \left[z_1(z_1-1)z_2\frac{\lambda_1(t)}{M} + z_1z_2(z_2-1)\frac{\lambda_2}{M}\right]\cdot\frac{\partial^2}{\partial z_1\partial z_2}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&- \left[z_1^2(z_1-1)\frac{\lambda_1}{M}\right]\cdot\frac{\partial^2}{\partial z_1^2}Q(u_1,u_2,z_1,z_2;t_0,t_1)\\
&- \left[z_2^2(z_2-1)\frac{\lambda_2^2}{M}\right]\cdot\frac{\partial^2}{\partial z_2^2}Q(u_1,u_2,z_1,z_2;t_0,t_1).
\end{aligned}
$$

Using 1.8, we can derive a system of coupled differential equations for the first and second moments of $\mathbf{X}(t)$:

$$
\begin{aligned}
\frac{\partial m_1^{(1)}(t)}{\partial t} &= (\lambda_1-\varphi_1-\gamma)m_1^{(1)}(t) - \frac{\lambda_1}{M}(m_{11}^{(2)}(t)+m_{12}^{(2)}(t)) \qquad (1.9)\\
\frac{\partial m_2^{(1)}(t)}{\partial t} &= \gamma m_1^{(1)}(t) + (\lambda_2-\varphi_2)m_2^{(1)}(t) - \frac{\lambda_2}{M}(m_{12}^{(2)}(t)+m_{22}^{(2)}(t))\\
\frac{\partial m_{11}^{(2)}(t)}{\partial t} &= (\lambda_1+\mu_1+\gamma)m_1^{(1)}(t) + \left(2\lambda_1-2\mu_1-2\gamma-\frac{\lambda_1}{M}\right)m_{11}^{(2)}(t)\\
&\quad - \frac{\lambda_1}{M}m_{12}^{(2)}(t) - \frac{2\lambda_1}{M}(m_{111}^{(3)}(t)+m_{112}^{(3)}(t))\\
\frac{\partial m_{12}^{(2)}(t)}{\partial t} &= -\gamma(m_1^{(1)}(t)-m_{11}^{(2)}(t)) + (\lambda_1-\mu_1+\lambda_2-\mu_2-\gamma)m_{12}^{(2)}(t) - \frac{\lambda_1+\lambda_2}{M}(m_{112}^{(3)}(t)+m_{122}^{(3)}(t))\\
\frac{\partial m_{22}^{(2)}(t)}{\partial t} &= (\lambda_2+\mu_2)m_2^{(1)}(t) + \gamma m_1^{(1)}(t) + \left(2\lambda_2-2\mu_2-\frac{\lambda_2}{M}\right)m_{22}^{(2)}(t) + \left(2\gamma-\frac{\lambda_2}{M}\right)m_{12}^{(2)}(t)\\
&\quad - \frac{2\lambda_2}{M}(m_{222}^{(3)}(t)+m_{122}^{(3)}(t))\\
\frac{\partial m_{111}^{(3)}(t)}{\partial t} &= (\lambda_1-\mu_1-\gamma)m_1^{(1)}(t) + \left(3\lambda_1+3\mu_1+3\gamma-\frac{\lambda_1}{M}\right)m_{11}^{(2)}(t) - \frac{\lambda_1}{M}m_{12}^{(2)}(t)\\
&\quad + 3\left(\lambda_1-\mu_1-\gamma-\frac{\lambda_1}{M}\right)m_{111}^{(3)}(t) - \frac{3\lambda_1}{M}m_{112}^{(3)}(t) - \frac{3\lambda_1}{M}m_{1111}^{(4)}(t) - \frac{3\lambda_1}{M}m_{1112}^{(4)}(t)\\
\frac{\partial m_{112}^{(3)}(t)}{\partial t} &= \gamma(m_1^{(1)}(t)-2m_{11}^{(2)}(t)) + (\lambda_1+\mu_1+\gamma)m_{12}^{(2)}(t) + \gamma m_{111}^{(3)}(t)\\
&\quad + \left(2\lambda_1-2\mu_1+\lambda_2-\mu_2-2\gamma-\frac{\lambda_1}{M}\right)m_{112}^{(3)}(t) - \frac{\lambda_1}{M}m_{122}^{(3)}(t) - \frac{2\lambda_1+2\lambda_2}{M}(m_{1112}^{(4)}(t)+m_{1122}^{(4)}(t))\\
\frac{\partial m_{122}^{(3)}(t)}{\partial t} &= -\gamma(m_1^{(1)}(t)-m_{11}^{(2)}(t)) + (\lambda_2+\mu_2-2\gamma)m_{12}^{(2)}(t) + \left(2\lambda_2-2\mu_2+\lambda_1-\mu_1-\gamma-\frac{\lambda_2}{M}\right)m_{122}^{(3)}(t)\\
&\quad + \left(2\gamma-\frac{\lambda_2}{M}\right)m_{112}^{(3)}(t) - \frac{2\lambda_1+2\lambda_2}{M}(m_{1222}^{(4)}(t)+m_{1122}^{(4)}(t))
\end{aligned}
$$

$$\frac{\partial m_{222}^{(3)}(t)}{\partial t} = (\lambda_2 - \mu_2) m_2^{(1)}(t) + \left(3\lambda_2 + 3\mu_2 - \frac{\lambda_2}{M}\right) m_{22}^{(2)}(t) - \frac{\lambda_2}{M} m_{12}^{(2)}(t)$$
$$+ 3\left(\lambda_2 - \mu_2 - \frac{\lambda_2}{M}\right) m_{222}^{(3)}(t) - \frac{3\lambda_2}{M} m_{122}^{(3)}(t) - \frac{3\lambda_2}{M} m_{2222}^{(4)}(t) - \frac{3\lambda_2}{M} m_{1222}^{(4)}(t)$$

where $m_i^{(1)}(t) = \mathbb{E}[X_i(t)]$, $m_{ij}^{(2)}(t) = \mathbb{E}[X_i(t)X_j(t)]$, $m_{ijk}^{(3)}(t) = \mathbb{E}[X_i(t)X_j(t)X_k(t)]$ and

$$m_{ijk\ell}^{(4)}(t) = \mathbb{E}[X_i(t)X_j(t)X_k(t)X_\ell(t)].$$

Again, we used the moment closure analytical approximation approach[55,56] by setting all the fourth central moments to zero, including $\mathbb{E}[(X_1(t)-\mathbb{E}[X_1(t)])^4]$, $\mathbb{E}[(X_1(t)-\mathbb{E}[X_1(t)])^3(X_2(t)-\mathbb{E}[X_2(t)])]$, $\mathbb{E}[(X_1(t)-\mathbb{E}[X_1(t)])^2(X_2(t)-\mathbb{E}[X_2(t)])^2]$, $\mathbb{E}[(X_1(t)-\mathbb{E}[X_1(t)])(X_2(t)-\mathbb{E}[X_2(t)])^3]$ and $\mathbb{E}[(X_2(t)-\mathbb{E}[X_2(t)])^4]$; as with the one-type logistic process, the two-type logistic process cannot be solved exactly[94]. Then, thanks to an anonymous reviewer, we used the following approximation to calculate the proportion of iPSCs at time t following the standard multivariate Taylor expansion[2]:

$$\mathbb{E}\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right]$$
$$\approx \frac{\mathbb{E}\left[X_2(t)\right]}{\mathbb{E}\left[X_1(t)+X_2(t)\right]} + \frac{\mathbb{E}\left[X_1(t)X_2(t)\right]\left(\mathbb{E}[X_2(t)] - \mathbb{E}[X_1(t)]\right) - \mathbb{E}[X_1(t)]\mathbb{E}[X_2(t)^2] + \mathbb{E}[X_1(t)^2]\mathbb{E}[X_2(t)]}{\left\{\mathbb{E}\left[X_1(t)+X_2(t)\right]\right\}^3} \tag{1.10}$$

The accuracy of the analytical approximation was validated using exact simulations based on Gillespie[27]. Together with the approximation (1.10), we also display the predictions of a much simpler formula

$$\mathbb{E}\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right] \approx \frac{\mathbb{E}\left[X_2(t)\right]}{\mathbb{E}\left[X_1(t)+X_2(t)\right]}. \tag{1.11}$$

The two approximations provide extremely similar results and interested readers can calculate either of

the two approximations by looking at the R code in the Supplementary Materials of Liu et al.[49].

Unfortunately, we did not find that this approach was useful for approximating the variance of the proportion of iPSCs Var $\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right]$. To approximate Var $\left[\frac{X_2(t)}{X_1(t)+X_2(t)}\right]$, one needs to approximate $\mathbb{E}\left[\left(\frac{X_2(t)}{X_1(t)+X_2(t)}\right)^2\right]$, which is a nonlinear functional of the distribution of $\{X_1(t), X_2(t)\}$. More theoretical investigations are necessary to obtain a reasonable approximation of this functional. In the interest of the theme of our paper, we decided not to pursue this goal further and use variances based on computer simulations instead.

1.3.4 Parameter estimation

To estimate the proliferation and apoptosis rates of somatic cells provided in Table 1.2, we first divided the real line into fixed size grids. We then searched within the grid to obtain a value of the proliferation rate that minimized the maximum squared difference over all measurements between the analytic approximation of the mean cell number trajectory predicted using the one-type probabilistic logistic process and the mean cell counts while assuming an apoptosis rate of 0. The mean cell counts were calculated from taking the product of the mean live cell counts and the mean live cell percentage from Table 1.2. Using this identified proliferation rate, we then chose the value of the apoptosis rate that minimized the maximum squared difference over all measurements between the analytic approximation and the mean live cell counts shown in Table 1.2. In particular, the proliferation rate estimator is of the form

$$\hat{\lambda}_* = \text{argmin}_{\lambda\in\mathbb{R}_+} \max_{k\in\{1,\cdots,K\}} \left(\widehat{E}[X_{*,\text{live}}(t_k)] - \widetilde{E}[X_{*,\text{live}}(t_k)]\right)^2$$

where $\widehat{E}[X_{*,\text{live}}(t_k)]$ is the average live cell count of type-$*$ cells at time t_k and $\widetilde{E}[X_{*,\text{live}}(t_k)]$ is the model-based mean cell count for type-$*$ cells at time t_k assuming no death rate. Here the initial cell count is set

Table 1.2: The number of live cells on day 1 and day 2, together with the percentage of live and dead cells, and the estimated proliferation and apoptosis rates for GMPs in the OSKM and OSKM + AGi conditions (data from Bar-Nur et al.[5]).

	OSKM		OSKM + AGi	
	Cell counts on day 1	Live cell counts on day 2	Cell counts on day 1	Live cell counts on day 2
Replicate 1	13,000	63,900	10,400	52,200
Replicate 2	11,700	66,600	10,100	58,200
Replicate 3	13,300	75,900	13,400	59,100
Mean	12,666.67	68,800	11,300	56,500
Standard deviation	850.49	6,295.24	1,824.83	3,751.00
Mean model prediction	69,223.00		55,927.98	
SD model prediction	4,451.80		8,619.29	
Proliferation rate λ_1	1.84		1.71	
Apoptosis rate φ_1	0.09		0.06	

as the average cell count at day 1. The death rate estimator is of a similar form by plugging in $\hat{\lambda}_*$:

$$\hat{\varphi}_* = \text{argmin}_{\varphi\in\mathbb{R}_+} \max_{k\in\{1,\cdots,K\}} \left(\widehat{E}[X_*(t_k)] - \widetilde{E}_{\hat{\lambda}_*}[X_*(t_k)]\right)^2$$

where $\widehat{E}[X_*(t_k)]$ is the average total cell count of type-$*$ cells at time t_k, and $\widetilde{E}_{\hat{\lambda}_*}[X_*(t_k)]$ is the model-based mean cell count of type-$*$ cells at time t.

To obtain confidence interval of these rates when the sample size is reasonably large (excluding the cell count data in Table 1.2), we employed the nonparametric bootstrap resampling approach Efron & Tibshirani[20] by sampling with replacement from the replicates and repeating the above procedures for $1,000$ bootstrap samples. Then the 95% confidence interval can be obtained from computing the 2.5%

and 97.5% quantile of the 1,000 bootstrap estimates.

Numerical modeling

All computer simulations[27] were performed using C and we used 1,000 replicates to obtain the summary statistics of the simulations. We used the open source R "deSolve"[88] function to numerically solve the differential equations with Euler methods[86], discretizing the time into 0.001-day unit intervals.

1.4 Mathematical modeling reveals different modes of reprogramming dynamics

We then utilized our mathematical model to analyze the time-course Oct4-GFP percent data from Bar-Nur et al.[5] with the goal of studying the dynamics of reprogramming under two growth conditions: somatic cells cultured in the presence of ascorbic acid and a GSK3-β inhibitor in addition to ectopic expression of the OSKM factors (the "OSKM + AGi" condition) and cells cultured with OSKM overexpression alone (the "OSKM" condition). We first obtained the parameter values for the proliferation and apoptosis rates of somatic cells under these two conditions from the proliferation data provided in Table 1.2; note that we do not provide a confidence interval for these estimates because the sample size is too small ($n = 3$). To this end, we counted the number of cells in wells of a 12-well dish at day 1 and day 2 as well as the percentage of live and dead cells. In particular, we used annexin staining with DAPI as a viability dye to determine cells that were apoptotic in order to directly estimate the apoptosis rate from the dead cell count. We then estimated proliferation and apoptosis rates together with the mean and standard deviation of cell counts at day 2 (Table 1.2). The net growth rate of iPSCs was calculated from an empirically derived iPSC doubling time of approximately 10.2 hours. However, since the cell doubling time might not be a very accurate way to estimate the proliferation rate, sensitivity analyses were conducted in the supplementary materials of Liu et al.[49]. The apoptosis rate of iPSCs was considered equal to that of somatic progenitor

cells. Sensitivity analyses to account for imprecise estimation showed that slight perturbations of the proliferation and apoptosis rates did not modify our results.

We then estimated the reprogramming rate γ from the experimental data by identifying the value that minimized the mean squared difference between the model-predicted mean percentage of iPSCs and the experimentally observed empirical mean of the percentage of cells with the Oct4-GFP signal. For the OSKM + AGi condition, we used the first measurement as the initial time point because only 8 out of 96 wells showed any signal. Using the estimation strategy detailed in Materials and Methods of Liu et al.[49], we identified $\gamma = 0.55\text{day}^{-1}$ (with 95% confidence interval [0.50, 0.61] day^{-1} obtained from a nonparametric bootstrap[20] in the OSKM + AGi condition. Next, we evaluated the consistency for the model prediction compared to the data using the maximum squared distance between model-predicted mean and sample average proportion of iPSCs over all six measurement occasions (0.0074), and found a correlation coefficient of $R^2 = 0.99$, suggesting consistency between the model predictions and the observed data. The relative overestimation of the model-predicted iPSC percentage on day 2 could potentially be explained by the results in Smith et al.[87]. Furthermore, to evaluate whether the model-based variability of the percentage of iPSCs at each time point was significantly different from the empirical variability, we calculated both the model-based and empirical Fano factors (defined as the ratio between the variance and mean) and performed a linear regression (adjusted $R^2 = 0.9386$), finding that the intercept of the linear regression output (-0.0177 with standard error 0.0122) was not significantly different from zero and the slope was not significantly different from one (0.833 with standard error 0.0947); we thus demonstrated that in the OSKM + AGi condition, the model prediction does not underestimate the variability of the observed data. These findings indicate that, even when assuming constant proliferation, apoptosis, and reprogramming rates across time and individual cells, the level of variability observed in this condition can be determined by the probabilistic nature of the model itself, and is not necessarily due

to any heterogeneous properties of the cells or reprogramming process themselves.

We then sought to utilize the same approach to analyze data from the OSKM condition. Using constant per-cell proliferation, apoptosis and reprogramming rates, we found that the reprogramming rate for the OSKM condition ($\gamma = 0.080\text{day}^{-1}$ with 95% confidence interval [0.073, 0.088] day^{-1} again computed from a nonparametric bootstrap) was significantly lower (p-value < 0.05) than for the OSKM + AGi condition ($\gamma = 0.56\text{day}^{-1}$ with 95% confidence interval [0.50, 0.61] day^{-1}), indicating that AGi exposure induces a dramatic increase in reprogramming efficiency. Similarly, we evaluated the consistency of the model prediction compared to the data using the maximum squared distance between the model-predicted mean and the average proportion of iPSCs over all eleven measurements (0.045, mainly driven by the fifth (day 20) and sixth (day 24) measurements during which the cell culture was split randomly; when removing these two points, the maximum squared distance was 0.0025), and correlation coefficients ($R^2 = 0.96$). We also found similar proliferation and apoptosis rates between the two conditions, which are thus unlikely to contribute significantly to the different reprogramming efficiencies between them (Table 1.2). Interestingly, the model-predicted variability did not provide as good a match to the data in the OSKM condition as in the OSKM + AGi condition. A visualization of Fano factors between the model prediction and the data demonstrates that only four time points out of eleven are localized on or below the 45-degree line. We decided not to evaluate the linear model between predicted and empirical Fano factors in this comparison because of lack of fit of linear regression (adjusted $R^2 = 0.06$). In addition, the average squared distance between model-based and data-based Fano factors in the OSKM condition is 0.0140, which is larger than that in the OSKM + AGi condition (0.006). There exist multiple explanations for the underestimated variability by the model. Measurement errors in the GFP read-out could be one possibility. However, to estimate the measurement errors, more experimental data obtained in different laboratories is necessary. Here we proposed another biologically plausible possibility – if

the reprogramming rate γ is a heterogeneous random variable instead of a homogeneous constant, the underestimation can also be compensated. As an example, considering a log-normal distribution of γ in the OSKM condition, we identified the parameters (a log-normal distribution with mean 0.08 and standard deviation 0.75) such that the variance of the model prediction based on 1000 simulations matched the empirical data with mean squared distance 0.007. The maximum squared distance between simulation-based and data-based mean % iPSCs was 0.035 (when not considering days 20 and 24, decreasing to 0.01). A similar Fano factor comparison showed that more than half of the data points were located below the 45-degree line, suggesting that a heterogeneous reprogramming rate can capture the variability observed in the data better than a homogeneous reprogramming rate.

It is possible that a heterogeneous proliferation and/or apoptosis rate can also contribute to the increased extent of variability observed in the experiments compared to the model prediction. We thus used the proliferation data (Table 1.2) and compared the model predictions, based on different assumptions about the variability of the proliferation and death rates, to the experimental data. These investigations indicate that the proliferation and/or apoptosis rates are not heterogeneous, hence supporting a heterogeneous reprogramming rate in order to explain the data if assuming that the additional variability is due to a heterogeneous property of the cells themselves. Together, these observations might suggest a heterogeneous reprogramming process in the OSKM condition but a homogeneous process during OSKM + AGi treatment when using GMPs as starting cells. However, other possibilities are still possible, such as measurement error or lineage priming. We also performed sensitivity analyses based on analytical approximations to test the robustness of our results; we obtained consistent results when considering data variability such as potential counting inaccuracies and insufficient data to estimate the iPSC apoptosis rate. Finally, we performed sensitivity analyses for the OSKM condition by changing the magnitude of proliferation and apoptosis rates of iPSCs but fixing the net growth rate of iPSCs to test whether that

approach would increase the intrinsic variability of the reprogramming dynamics when considering a homogeneous reprogramming rate.

1.5 The probabilistic two-type logistic process modeling reprogramming dynamics has predictive power

One criterion for evaluating the generalizability and utility of a quantitative model is to evaluate its out-of-sample predictive power[25]. To this end, we first used a subset of time points from the experiments in Bar-Nur et al.[5] to predict the iPSC trajectories, in an approach similar to that used in Morris et al.[52]. We then investigated whether the model predictions based on a subset of time points was similar to that based on all time points. In the OSKM + AGi condition, the estimated reprogramming rate based on only the first three out of seven time points (0.52day^{-1}) was similar to the estimate using all time points (0.55day^{-1}).

We next aimed to evaluate the model with an independent dataset[109] in which somatic cells were exposed to either OSKM overexpression alone or in combination with ascorbic acid treatment, TGF-β inhibition, and GSK3-β inhibition. There was insufficient data available for the OSKM experiment to evaluate the model fit; the other growth condition, however, was amenable for analysis. We thus compared this dataset with the model prediction using parameters obtained from the investigation of data from Bar-Nur et al.[5] and achieved an excellent fit ($R^2 = 0.96$). We also estimated the reprogramming rate (0.52 per day, with confidence interval [0.42, 0.61]) from this new dataset, which is very similar to the one estimated from the OSKM + AGi experiment. Our model thus has significant predictive power when applied to independent datasets. In addition, when comparing the Fano factors calculated from model predictions and the data using linear regression (adjusted $R^2 = 0.81$), we found again that the intercept was not significantly different from 0 (-0.02 with standard error 0.050) and the slope was not significantly

smaller than 1 (1.85 with standard error 0.40) respectively, indicating that a constant reprogramming rate can capture the variability of the observed data.

1.6 The probabilistic two-type birth-death process can model the first appearance time of the iPSC signal

Aside from collecting the time-series percentages of certain markers (such as Oct4-GFP or Nanog-GFP) representing the level of iPSC formation, another common approach is to measure the time of the first appearance of some signal of these markers across multiple replicates (wells or colonies)[32,66]. We thus also utilized the multi-type birth-death-transition process to analyze such datasets[32,66] to further demonstrate the generalizability of our approach. We did not consider a carrying capacity due to the frequent plate splitting in the experiments[32,66]. To find out the first passage time when the percentage of iPSCs reached a certain threshold (0.5%), we performed Monte Carlo simulations to generate 1000 replicates for a range of reprogramming rates and searched for the rate that minimized the maximum squared distance between the simulation and the observed data over all measurements.

We first studied the Mbd3 knock-down experiment[66], which was interpreted by the authors to lead to a relatively fast and deterministic transition. Assuming exponential growth, the proliferation rate (0.853 day^{-1}) for MEF cells was directly estimated from the raw cell-doubling time (19.5 hrs) shared by the authors. Unfortunately, no other information was available to estimate the apoptosis rate. We found that a delayed constant reprogramming rate explained the data ($R^2 = 0.98$ for both replicate experiments), where the delayed reprogramming rate was a step function equal to zero before day 1 and equal to 0.344 week^{-1} after day 1. Otherwise, without this delayed effect, the predicted percentage of wells with more than 5% iPSCs at day 2 is larger than zero. Here we again used the procedure described in Materials and Methods of Liu et al.[49] by identifying the reprogramming rate that minimizes the maximum squared

distance between the model prediction based on the simulation and the experimental data. Such delayed effects might be observed due to multiple reasons; it could be due to the detection sensitivity[32,66], or because cells in culture need to pass through unobserved intermediate states before dividing or reprogramming. Unfortunately, there was no higher-resolution time-series data available to address such questions. Furthermore, we found that our multi-type birth-death-transition process model without delayed reprogramming can explain the relatively low efficiency NGFP1 control experiment[66] (reprogramming rate is $8.57 \times 10^{-6}\text{week}^{-1}, R^2 = 0.99$) as well as the NGFP1-Nanog(OE) experiment performed by Hanna et al.[32] (reprogramming rate is $6.4 \times 10^{-4}\text{week}^{-1}, R^2 = 0.99$).

1.7 The probabilistic birth-death-transition process can model the colony cell count data

We then collected data of three distinct cell fate types defined by Smith et al.[87], in which cells were not selected for iPSC potency and were categorized into fast-dividing (FD), slowly-dividing (SD) and iPSC-forming lineages after doxycycline induction. We observed that the cellular growth patterns satisfied an exponential growth model without reaching confluence, and therefore used a linear birth-death process without a carrying capacity to model the cellular growth based on the cell count data described in Result IV. Since the cell count data over multiple time points for the three cell fates were measured retrospectively and conditional on lineage non-extinction, i.e. colony formation, we first calculated the theoretical mean and variance of cell counts at different time points conditional on population non-extinction (supplementary materials of Liu et al.[49]). We then used the empirical mean and variance computed from the data halfway to the end of follow-up to estimate the growth and death rates of the three cell types (Table 1.3). Based on these rates, we then compared the model prediction and the empirical data in terms of both mean and standard deviation of the cell count trajectory over time, demonstrating that our approach can also be used to model cellular growth data in this experimental setup. Finally, using the estimated birth

Table 1.3: Growth and death rates with their 95% confidence intervals (based on nonparametric bootstrapping) estimated for FD, SD, and iPSC fates (data from Smith et al.[87]).

	Proliferation rate (day^{-1})	Apoptosis rate (day^{-1})
FD cells	1.724 [1.612, 1.836]	0.553 [0.441, 0.665]
SD cells	0.964 [0.825, 1.103]	0.330 [0.191, 0.469]
iPSCs	1.567 [1.454, 1.680]	0.483 [0.370, 0.596]

and death rates for FD cells and iPSCs, and the estimated reprogramming rate for iPSCs (0.01 per day) from Pour et al.[65] and for FD ($\sim 10^{-8}$ per day) from Hanna et al.[32], we simulated the reprogramming dynamics for a mixture of FD cells and iPSC-forming lineages with the empirically determined mixture ratios of FD:iPSC = 6:58 and FD:iPSC = 6:19. Using this approach, we obtained lower predicted early-phase iPSC dynamics for admixtures as compared to homogeneous iPSC populations. This population admixture effect captured in the early phase of reprogramming in Smith et al.[87] and Pour et al.[65] might explain the overestimation of our model prediction for the percentage of iPSCs in the earliest measured time points of the OSKM + AGi condition in Bar-Nur et al.[5] and possibly also the overestimation of the model proposed in Hanna et al.[32] for the early phase Nanog-GFP+ well percentages.

1.8 Discussion

Here we designed a two-type probabilistic logistic process model to investigate the dynamics of induced reprogramming from somatic cells into iPSCs. We found that this birth-death-transition process with a constant (or homogeneous) reprogramming rate can recapitulate the dynamics of iPSCs after exposure to chemical supplements in addition to OSKM overexpression from two independent datasets[5,109]. For experiments with only ectopic expression of OSKM, the same process applies but with a heterogeneous instead of constant reprogramming rate. Our investigations thus reveal two different modes of cellular reprogramming dynamics: OSKM expression alone leads to heterogeneous reprogramming while OSKM plus certain other factors homogenizes the dynamics.

Unlike previous methods focusing on statistics such as the first passage time[32,52,66,112], our approach explicitly models the reprogramming rate and thus can be used to make direct computational inferences about the heterogeneity of cellular populations with regard to induced reprogramming. Furthermore, by carefully considering the effects of proliferation, apoptosis, reprogramming and the carrying capacity, we were able to identify differences in the reprogramming rate itself that resulted in the acceleration of reprogramming in the OSKM + AGi as compared to the OSKM condition. We further explored the source of variability leading to the increased variance observed in the OSKM data. However, due to lack of sufficiently many replicates and longer follow-up times when counting the cell numbers, further work is warranted to better assess the variability of cell growth and death in different conditions. It will also be necessary to conduct follow-up experiments to further address whether the additional variability comes from measurement error or heterogeneous cell population. In addition, the log-normal distribution of the reprogramming rate used in our paper is only one out of infinitely many possibilities based on the current data. A recent paper[96] also showed that combining ascorbic acid (AA) and 2i (MAP kinase and GSK inhibitors) can synergize reprogramming. Even though our modeling does not directly model the first passage time, it is not difficult to use our model to study such data. Since we can always transform the time course percent iPSC data into first passage time data, we argue for collecting time course percent iPSC whenever possible since such data allows for more detailed characterization of the reprogramming dynamics.

Although our current framework is promising for modeling induced reprogramming or more general cellular fate change phenomena, several caveats apply. First, we do not have enough information to distinguish between different OKSM systems. For instance, Hanna et al.[32] used OSKM while Bar-Nur et al.[5] used OKSM; however, we cannot directly compare the data because of different data collection processes employed by these two laboratories. As a result, we used the same terminology “OSKM” to

indicate the overexpression of Yamanaka factors. Second, we estimated the parameters in the two-type model by minimizing the squared distance between the model prediction and observed data; an alternative inference strategy would include likelihood-based methods to obtain the maximum likelihood estimator with good statistical properties[14]. Though some recent advances have been reported[35], the tools needed to make inferences about the reprogramming rate in the two-type case, however, are currently unavailable. Furthermore, likelihood-based methods such as the EM algorithm are usually computationally intensive when applied to situations with population sizes at the scale of millions[14]. More carefully designed experiments[16] and advanced technology to collect single cell as well as molecular data would also allow for better model design and parameterization. Another implication of our model is that there is a positive probability of acquiring pluripotency immediately after the start of the experiment, when AGi is added, which might suggest acceleration of the transition from an early population with a heterogeneous capacity of acquiring pluripotency towards a more deterministic or homogeneous process occurring later[9]. To delineate these possibilities and to retrace the early events in relatively fast regimes such as with addition of AGi, data needs to be collected frequently in the very early phases of the experiment. Also, when analyzing the data from Rais et al.[66], we observed time-delayed reprogramming rates, especially in the relatively slow reprogramming regimes. These results might be partly due to the use of different biomarkers for tracing reprogramming events (Table 1.1), thus emphasizing the need to standardize approaches and biomarker usage in the field to enable a quantitative comparison of results and processes. Furthermore, it is possible that the 'conversion to iPSC' does not represent the immediate acquisition of all iPSC characteristics but rather the symmetrical transmission of iPSC competence to all subsequent progeny – i.e. the switch to deterministic acquisition of pluripotency after an initially probabilistic event[9,60,64,63].

To robustly test the assumptions and the consequences of the multi-type birth-death-transition process model exploited in this paper, experiments from different laboratories will be necessary to account for

potential confounders such as batch effects of these cellular dynamic/kinetic experiments. Also, to test whether the assumption of the model listed in Materials and Methods of Liu et al.[49], one need cell count measurements for more time points instead of two time points, to test the relation between the population cell growth and the current population size. In addition, to test heterogeneity in reprogramming rate vs. measurement error, the same 96-well plates experiment repeated multiple times will be important to infer the well-to-well variability in different "batches". Our approach can further be extended to explicitly study the effects of cell cycle times on reprogramming dynamics. For instance, Guo et al.[30] reported that fast cycling cells tend to reprogram more efficiently than slow cycling ones. To directly test such a hypothesis in our system, data on cell division kinetics for both fast and slow cycling cells are required together with data for dissecting the time-ordering between reprogramming and proliferation, but unfortunately such data is currently not available.

Apart from the probabilistic birth-death-transition process framework, several studies have explored different modeling perspectives for studying reprogramming dynamics[19,32,52,66,111,112]. Most of these directly model the reprogramming latency time. In this paper, we also demonstrated that the current probabilistic logistic birth-death-transition process model can be applied to study the latency time distribution by calculating, at each time point, the fraction of wells surpassing a certain threshold. However, to our best knowledge, there is no available standard of choosing such as threshold, and therefore we suggest that experimentalists collect the iPSC percentage for all wells rather than discontinuing to follow the dynamics when the signal first appears.

In summary, we have developed a new two-type probabilistic logistic birth-death model to interrogate the dynamics of transcription factor-induced reprogramming of somatic cells into iPSCs following different genetic or environmental perturbations by independent laboratories. We anticipate that our methodology will be applicable to other reprogramming systems utilizing different transcription factor

combinations and cell fate conversion systems such as the reprogramming of epiblast stem cells into embryonic stem cells or cellular transdifferentiation. Likewise, our approach is useful for interrogating the dynamics of forward differentiation approaches using pluripotent stem cells.

2

Asymptotics of Confounder Selection under the Doubly-Robustness Framework

2.1 Introduction

Estimating the treatment effect is a central pillar in both statistical and causal inference literatures[81,78,70,71,61]. Even though not very often stated in practice[107], estimating treatment effect usually involves some model selection steps implicitly or explicitly. For a fairly conceptual and philosophical discussion on

model selection in causal inference, readers are referred to Robins & Greenland[72]. Fairly recently, an algorithm called *Collaborative Targeted Minimum Loss Estimation* (abbreviated as CTMLE) was developed and improved by van der Laan and his coworkers over the years. Many key results of CTMLE are included in two comprehensive textbooks[102,103], together with a large body of papers and ongoing works covering a variety of different grounds, including theoretical studies on the use of CTMLE for causal inference in both single and multiple-point treatments[98,99,91], together with some successful applications in epidemiology and genomics[29,90]. The goal of model selection when estimating causal or treatment effect is to tease out the confounders from all the measured covariates and balance the bias-variance trade-off: one do not want to adjust for more covariates than necessary, which can sometimes blow-up the variance of the estimator, or occasionally even introduce M-bias[82,62]; one do not want to leave out some confounders either, which will lead to huge bias of the estimator. For a more conceptual discussion on the definition of confounders, the readers are referred to VanderWeele & Shpitser[106] that offers a very comprehensive and insightful discussion on this subject matter, building upon the earlier effort in Robins & Morgenstern[75], which was among the first to explore a rigorous definition of confounder in the field of epidemiology. When dealing with model selection problem in practice, our view closely follows that of Robins & Greenland[72] to balance the bias-variance tradeoff: we would like a certain selection procedure to be able to adjust for a potential confounder unless the increase of uncertainty outweighs the gain in unbiasedness. Given such philosophy, we bring another selection procedure "*Focused Information Criterion*" (abbreviated as FIC)[12,13,107] into the story because the very idea of FIC is to construct consistent estimators (FICs) for the mean squared errors of all the candidate estimators and use FIC to select the final estimator, or perform model averaging technique to obtain the final estimator[34,13]. In contrast, CTMLE use the prediction risk estimated from cross-validation as the statistics for model selection. Apart from the use of different statistics as in the selection algorithm, CTMLE also incorpo-

rates a more complicated and unique algorithm which will be elaborated in Section 2.2 later. One other type of selection procedure that we will discuss in this paper is a modification by Spokoiny & Vial[89] on Lepski's method, which was named after its introduction by the celebrated papers by Oleg Lepski in 1990's[45,46]. Spokoiny & Vial[89]'s method (abbreviated as Spokoiny's method) is built upon the idea of hypothesis testing and multiple testing correction by fine tuning the cut-off for accepting or rejecting a covariate or a set of covariates.

In a nutshell, our aim in the present paper is to point out the mathematical structure and use the simplified mathematics in asymptotics to study the behavior of different confounder selection procedures mentioned above. To stress the doubly-robust[4,79] origin of the estimators considered in our paper, we call the "clever" covariates defined in van der Laan & Gruber[97] and van der Laan & Gruber[98] as DR (short for "doubly-robust") covariates. To fix idea, let's consider a data analysis to estimate the counterfactual mean of some outcome of interest Y were the subjects being treated with treatment $R \in \{0,1\}$. Together with the outcome and treatment assignment, the experimenter also measured a vector of covariates X for each subject that may or may not be a confounder. We are interested in estimating the target estimand $\Psi_0 := \mathbb{E}[Y(r=1)]$ (the counterfactual mean of Y were the subject being treated with $R=1$). The ideas developed in this paper can be applied to other interested causal functionals as well but we only focus on Ψ_0 for the ease of presentation. Then under the no unmeasured confounders, consistency and positivity assumptions[70,73], one can connect the counterfactual quantity with the quantity solely depends on observed data: $\Psi_0 := \mathbb{E}[Y(r=1)] = \mathbb{E}\left[\frac{RY}{Pr[R=1|X]}\right]$. Following the theory developed in Bickel et al.[7], the nonparametric efficient influence function for Ψ_0 is

$$\begin{aligned}\text{EIF}\left(\Psi_0\right) &= \left(\frac{RY}{Pr[R=1|X]} - \Psi_0\right) - \frac{R - Pr[R=1|X]}{Pr[R=1|X]}\mathbb{E}\left[Y|R=1,X\right] \\ &= \frac{R}{Pr[R=1|X]}\left(Y - \mathbb{E}\left[Y|R=1,X\right]\right) + \mathbb{E}\left[Y|R=1,X\right] - \Psi_0 \ .\end{aligned} \tag{2.1}$$

Equation (2.1) has mean zero, based on which one can construct an estimating equation for the target estimand Ψ_0. There are a number of strategies to construct estimators based on eq. (2.1) and the estimator so constructed is known to enjoy the doubly-robustness property, i.e. the estimator consistent to the target estimand in the union of the outcome regression model and the propensity score model[4,104]. CTMLE is based on one . In the very first step of CTMLE, super learner[100] is utilized to establish an initial estimator $b_0(X)$ for the conditional outcome regression model $\mathbb{E}\left[Y|R=1,X\right]$, possibly from an independent dataset. To enjoy the doubly-robustness property, CTMLE proceed by constructing estimators of the following form: $\mathbb{E}_n\left[b_0(X) + \hat{\epsilon}^T \cdot g(R,X)\right]$, where $\mathbb{E}_n$ is the empirical mean operator, g is the DR covariate, the form of which depends on the target estimand, and $\hat{\epsilon}$ is the estimated coefficient of the DR covariate. In this example, $g(R,X) = \frac{1}{Pr(R=1|X)}$. We call estimators of such form CTMLE-estimators. Any confounder selection procedure based on CTMLE-estimators will select which subset of X should enter the propensity score model from data. CTMLE algorithm offers one possible solution. As mentioned above, there exist other possibilities, e.g. FIC and Lepski-related method. For a more detailed description of the CTMLE selection process, one can refer to Chapter 19 in van der Laan & Gruber[98] and Schnitzer et al.[82]. We will also provide a brief review in Section 2.2. For other alternative methods, we will describe their procedures respectively in the corresponding sections. To study how different selection procedures behave in terms of bias-variance trade-off or the final mean squared error of the estimator after confounder selection, the data generating process (DGP) with $\sqrt{n}$ local alternative outcome regression model (or Pitman's alternative) provides a ideal setup as the bias and variance of each candidate estimator

are of the same order. Such data generating process has been frequently brought up in the model selection literature to demonstrate the difficulty of making valid statistical inference after model selection[12,42] because asymptotic bias is not dominated by asymptotic variance. We now describe such DGP below (Section 2.1).

Notations

Throughout this paper, we reserve $Z.$ to denote one element of a multivariate Gaussian random vector with zero mean, unitary variance, and certain covariance structure that will be made clear later. $\xrightarrow{d}$ denotes convergence in distribution. No unmeasured confounder is assumed to hold, implying that adjusting for all the measured potential confounders should suffice to produce an unbiased estimator of the targeted estimand. In this paper, we are interested in estimating the counterfactual mean outcome were subjects being treated $\Psi_0 := \mathbb{E}[Y(r = 1)]$, where $Y(r = 1)$ denotes the counterfactual outcome Y were the subjects being treated. We can also perceive $R = 1$ as the observable indicator and the target becomes the counterfactual mean were the subjects being observed. The data generating processes (DGP) that we are primarily investigating in this paper is the $\sqrt{n}$ local alternative (or Pitman's alternative) conditional outcome regression model (Definition 2.1.1). Such DGP has been widely used as a theoretical framework in classical papers on model selection, such as Claeskens & Hjort[12] and Leeb & Pötscher[42]. The DGP in Definition 2.1.1 is so constructed that the asymptotic bias and variance of the candidate estimators are of the same order $O\left(n^{-1/2}\right)$. If the perturbation is of smaller order than $\sqrt{n}$, then the mean squared error (M.S.E.) is dominated by the variance asymptotically; if larger than $\sqrt{n}$, then the M.S.E. is dominated by the bias. Then one can study how different procedures balance the bias-variance tradeoff through the lens of M.S.E., or some other metric such as the ratio between bias squared and variance after confounder selection.

Definition 2.1.1. *DGP with $\sqrt{n}$ local alternative conditional outcome regression model:*

1. The underlying data consists of n independent and identically distributed (iid) realizations of the data vector $(Y_i, X_i = \{X_{1i}, X_{2i}, \cdots, X_{di}\}, R_i), i = 1, \cdots, n$, in which the subvector X_i is a collection of all d measured covariates or "potential confounders", scalar Y_i is the outcome, and R_i is the indicator of treatment option. When no meaning is lost, we will suppress the subject index i. The set of all the measured covariates is denoted as $K_{full} := \{X_1, \cdots, X_d\}$. Similarly we use $K_\cdot$ to denote an arbitrary subset of all the measured covariates.

2. The true outcome regression model is assumed to be $\mathbb{E}[Y|X] \equiv b_0 + \frac{h(K_{d_Y})}{\sqrt{n}}$. For simplicity we assume that $h(\cdot)$ is exchangeable in its argument. The true propensity score is assume to be $\mathbb{E}[R|X] \equiv \mathbb{E}[R|K_{d_R}] := \pi_{K_{d_R}}$. For convenience, any other possible propensity score models controlling for some arbitrary set of measured covariates $K_\cdot$ is also abbreviated with the short-hand notation $\pi_{K_\cdot}$. Further, we denote $K_{d^\star}$ as the set of true confounders, i.e. $Y \perp\!\!\!\perp R \mid K_{d_\star}$. Since we assume no unmeasured confounders, $K_{d_Y} \subseteq K_{full}$ and $K_{d_R} \subseteq K_{full}$ and obviously $K_{d_\star} \subseteq K_{full}$. When the subscript in $K_\cdot$ is an integer, we use that integer to denote the number of covariates included in this set by default. To be concrete, K_1 is a singleton, K_2 is a doubleton, etc. We assume any marginal or conditional of the true propensity score is strictly bounded between 0 and 1. Under such assumption, we can rewrite $\mathbb{E}[Y(r=1)] \equiv \mathbb{E}\left[\frac{RY}{\pi_{K_{d_R}}}\right]$. Assume the true conditional variance of the outcome as $var[Y|X] = var[Y|K_{d_Y}] := v_{K_{d_Y}}$ and similarly we denote any other conditional variance controlling for some arbitrary covariate set $K_\cdot$ as $v_{K_\cdot}$. Further we define the following parameters related to the (co)variance of the estimators which will be useful in later sections. For two different covariate sets K_ℓ, K_m, define v_{K_ℓ, K_m} as follows, where $v_{\cdot,\cdot}$ is symmetric in its subscript arguments but when the two set of potential confounders are ordered, i.e. $K_\ell \subseteq K_m$, we always place the smaller set of potential confounders K_ℓ in the first argument and the larger set K_m in

the second argument for clarity:

$$v_{K_\ell,K_m} := \mathbb{E}\left[\frac{R \cdot var[Y|X]}{\pi_{K_\ell}\pi_{K_m}}\right] \equiv \mathbb{E}\left[\frac{\pi_{K_{d_R}} v_{K_{d_Y}}}{\pi_{K_\ell}\pi_{K_m}}\right] \tag{2.2}$$

Similarly, we also define the following parameters related to the asymptotic biases of the candidate estimators, for some covariate set K_m

$$p_{K_m} := \sqrt{n}\mathbb{E}\left[\left(\frac{R}{\pi_{K_m}} - 1\right)(Y - b_0)\right] \equiv \begin{cases} \sqrt{n}\mathbb{E}\left[\left(\frac{\pi_{K_{d_R}}}{\pi_{K_m}} - 1\right)(Y - b_0)\right] & \text{if } K_{d_R} \not\subseteq K_m, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

For convenience, we denote $q := \sqrt{n}\mathbb{E}[Y - b_0]$ because this parameter will be frequently used in later sections when we present the results on the asymptotic distributions of the statistics used in CTMLE.

3. For this paper, we assume that we have full knowledge about all the propensity score models, the conditional outcome variance, and the distribution of the all the measured potential confounders.

Remark 2.1.2. *It is worth making some comments on the variance-covariance parameters $v_{\cdot,\cdot}$ when the outcome Y is homoscedastic conditional on the covariates X, i.e. $var[Y|X] \equiv v_{K_{d_Y}} = v$ for some constant v not dependent on X. Without loss of generality, assume $v = 1$. Then if $K_\ell \subseteq K_m$, $v_{K_\ell,K_\ell} \equiv \mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right] \leq v_{K_m,K_m} \equiv \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]$ following from Jensen's inequality. Also, the covariance parameter $v_{K_\ell,K_m} = \mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]$. Hence homoscedasticity dramatically simplifies the parameter space.*

Organization of this paper

In this paper, we derive the asymptotic distributions of the statistics used as the criterion for confounder selection in different procedures, including the prediction risks and within-sample prediction risks employed in CTMLE[98], focused information criterion statistics to unbiasedly estimate the asymptotic

M.S.E. (A.M.S.E.) used in FIC[12], the statistics derived from a sample-splitting strategy as an unbiased estimator of A.M.S.E. proposed in Vansteelandt et al.[107], and the test-statistic used in the procedures related to Lepski's method[45,46,89]. The main tasks that we set out to do in this article is to investigate the performance of different procedures in large sample, propose some other reasonable procedures on the basis of the above existing methods, and discuss their pros and cons in different settings (i.e. different bias and variance-covariance parameter combinations). The major criterion is to compare the post-selection A.M.S.E.. In general, the analytic formula for post-selection A.M.S.E. is difficult to derive, and we have to resort to *Monte Carlo* simulations for approximation. The rest of our paper is organized as follows.

From Section 2.2 to Section 2.4, we develop the large-sample results for the key quantities used in CTMLE algorithm, including the estimated coefficients for the "clever covariates" (Section 2.2), the squared error (Section 2.2), the within-sample prediction risk (Section 2.3), the population-version prediction risk conditioning on the estimated coefficients (Section 2.3), and the prediction risk based on M-fold cross-validation (Section 2.4). To make our paper self-contained, we will follow the mathematical definitions of the above quantities with a brief review on the CTMLE algorithm in Section 2.2.

From Section 2.5 to Section 2.6, we shift our focus from CTMLE to other procedures, including focused information criterion[12], a sample-splitting strategy to consistently estimate the A.M.S.E.[107] for each candidate estimator, and Lepski's method of constructing a final estimator from a family of candidate estimators[45,46,89]. Similarly, we develop the asymptotic distribution for these statistics.

In Section 2.7, we will briefly summarize the algebraic relations of the asymptotic distributions among the key statistics in our paper, including prediction risk based on either infinite sample (as if we know the true DGP) or M-fold cross validation as $M \to \infty$, within-sample prediction risk, FIC, sample-splitting strategy to estimate A.M.S.E. as the fold $M \to \infty$, and Lepski-related statistics.

We conclude with a discussion (Section 2.8), including limitations of our analysis and potential future

research directions.

2.2 Large sample distribution of the DR coefficients in CTMLE and squared error loss of the candidate estimators

A vignette of CTMLE algorithm

We begin this section by defining the family of candidate estimators in the CTMLE algorithm when estimating the target of interest Ψ_0. A more detailed description of the procedure for estimating a general target of interest can be found in Chapter 19 of van der Laan & Rose[102]. To start, CTMLE uses super learner[100] to construct an initial estimator Q_0 for the outcome regression model. For simplicity, we assume that $Q_0 = b_0$ for the time being. Another important concept is the "clever covariate" in CTMLE (or DR covariate as dubbed in this paper to reflect the doubly-robustness property of estimators so constructed). For Ψ_0, the DR covariates are of the form $\frac{1}{\pi_{K_\cdot}}$, where $\pi_{K_\cdot}$ can be any propensity score model by marginalizing over or taking the conditional expectation of the true propensity score $\pi_{K_{d_R}}$. Then the first estimator $\hat{\Psi}_{K_0}$ of Ψ_0, not adjusting for any potential confounders, is constructed as follows. Let $\hat{\Psi}_{K_0} := \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_0}}{\pi_{K_0}} \right]$, where $K_0 = \emptyset$ ($\pi_{K_0} \equiv \mathbb{E}[R]$), and $\hat{\epsilon}_{K_0}$ is the solution of $\mathbb{P}_n \left[\frac{R(Y - b_0 - \epsilon_{K_0}/\pi_{K_0})}{\pi_{K_0}} \right] = 0$ i.e. $\hat{\epsilon}_{K_0} = \mathbb{P}_n \left[\frac{R}{\pi_{K_0}^2} \right]^{-1} \cdot \mathbb{P}_n \left[\frac{R(Y - b_0)}{\pi_{K_0}} \right]$. $\hat{\Psi}_{K_0}$ is so constructed that it only includes one DR covariate not adjusting for any potential confounders in the propensity score. In this paper, we call the coefficients for the DR covariates as DR coefficients for convenience.

Remark 2.2.1. *Simple calculations show that* $\hat{\Psi}_{K_0} \equiv \mathbb{P}_n[R]^{-1} \cdot \mathbb{P}_n[RY]$, *i.e. the empirical average outcome within the treated/observed subjects:*

$$\hat{\Psi}_{K_0} := \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_0}}{\pi_{K_0}} \right] = \mathbb{P}_n[b_0] + \mathbb{P}_n \left[\frac{1}{\pi_{K_0}} \right] \mathbb{P}_n \left[\frac{R}{\pi_{K_0}^2} \right]^{-1} \cdot \mathbb{P}_n \left[\frac{R(Y - b_0)}{\pi_{K_0}} \right]$$

$$= b_0 + \mathbb{P}_n[R]^{-1} \cdot \mathbb{P}_n[RY - Rb_0] = b_0 + \mathbb{P}_n[R]^{-1} \cdot \mathbb{P}_n[RY] - b_0 = \mathbb{P}_n[R]^{-1} \cdot \mathbb{P}_n[RY]$$

in which the key part is π_{K_0} and b_0 are constants not depending on any potential confounders X. One interesting implication is that if using $\hat{b}_0 := \mathbb{P}_n[R]^{-1} \cdot \mathbb{P}_n[RY]$ instead using b_0 as the initial point, then $\hat{\Psi}_{K_0}^{\hat{b}_0} = \hat{b}_0$ because $\hat{\epsilon}_{K_0}^{\hat{b}_0} = \mathbb{P}_n\left[\frac{R}{\pi_{K_0}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-\hat{b}_0)}{\pi_{K_0}}\right] = 0$. In the asymptotic theory developed below, we still assume b_0 as the initial point.

After constructing the estimator $\hat{\Psi}_{K_0}$ adjusting for no potential confounders, one can continue to construct estimators adjusting for only one potential confounders out of all d collected potential confounders. To proceed, define two possible forms of estimators at level K_1: $\hat{\Psi}_{K_1} := \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_1}}{\pi_{K_1}}\right]$ and $\hat{\Psi}_{K_0,K_1} := \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_0}}{\pi_{K_0}} + \frac{\hat{\epsilon}_{K_0;K_1}}{\pi_{K_1}}\right]$, in which K_1 can be one of the members in the following family of singletons of potential confounders: $\{\{X_1\}, \{X_2\}, \cdots, \{X_d\}\}$ (where $\pi_{K_1} \equiv \mathbb{E}[R|K_1]$). Apparently, $K_0 \subseteq K_1$, i.e. they are partially ordered sets with order defined using the subset relationship "$\subseteq, \supseteq$". Following the terminology in van der Laan & Rose[102], $\hat{\Psi}_{K_1}$ is called the "non-fluctuated" estimator adjusting for K_1 whereas $\hat{\Psi}_{K_0,K_1}$ is called the "fluctuated" estimator adjusting for K_1 from $\hat{\Psi}_{K_0}$. Similarly, $\hat{\epsilon}_{K_1}$ is the solution of $\mathbb{P}_n\left[\frac{R(Y-b_0-\epsilon_{K_1}/\pi_{K_1})}{\pi_{K_1}}\right] = 0$ i.e. $\hat{\epsilon}_{K_1} = \mathbb{P}_n\left[\frac{R}{\pi_{K_1}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-b_0)}{\pi_{K_1}}\right]$, and $\hat{\epsilon}_{K_0;K_1}$ is the solution of

$$\mathbb{P}_n\left[\frac{R(Y-b_0-\hat{\epsilon}_{K_0}/\pi_{K_0}-\epsilon_{K_0;K_1}/\pi_{K_1})}{\pi_{K_1}}\right] = 0$$

i.e. $\hat{\epsilon}_{K_0;K_1} = \mathbb{P}_n\left[\frac{R}{\pi_{K_1}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-b_0-\hat{\epsilon}_{K_0}/\pi_{K_0})}{\pi_{K_1}}\right] = \mathbb{P}_n\left[\frac{R}{\pi_{K_1}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-b_0)}{\pi_{K_1}}\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_1}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R}{\pi_{K_0}\pi_{K_1}}\right] \cdot \mathbb{P}_n\left[\frac{R}{\pi_{K_0}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-b_0)}{\pi_{K_0}}\right]$. In this step, CTMLE will construct $2d$ estimators including DR covariates adjusting for at most one potential confounder in the propensity score models.

Next we describe how CTMLE constructs estimators including DR covariates adjusting for at most two potential confounders in the propensity score models. In CTMLE, one of two potential confounders

in the set K_2 has to belong to K_1, so that $K_1 \subseteq K_2$, i.e. K_0, K_1 and K_2 are partially ordered. Now one has four different ways of constructing the estimator at level K_2: the "non-fluctuated" estimator $\hat{\Psi}_{K_2} := \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_2}}{\pi_{K_2}}\right]$, the estimator fluctuated from $\hat{\Psi}_{K_i}$ $\hat{\Psi}_{K_i,K_2} := \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_2}}{\pi_{K_2}}\right]$ for $i = 0$ or 1, and the estimator fluctuated from $\hat{\Psi}_{K_0,K_1}$

$$\hat{\Psi}_{K_0,K_1,K_2} := \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_0}}{\pi_{K_0}} + \frac{\hat{\epsilon}_{K_0;K_1}}{\pi_{K_1}} + \frac{\hat{\epsilon}_{K_0,K_1;K_2}}{\pi_{K_2}}\right].$$

Notice that as discussed above K_1 is not unique. Similarly, we solve $\mathbb{P}_n\left[\frac{R(Y - b_0 - \epsilon_{K_2}/\pi_{K_2})}{\pi_{K_2}}\right] = 0$ to get $\hat{\epsilon}_{K_2} = \mathbb{P}_n\left[\frac{R}{\pi_{K_2}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y-b_0)}{\pi_{K_2}}\right]$,

$$\mathbb{P}_n\left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_1}/\pi_{K_1} - \epsilon_{K_1;K_2}/\pi_{K_2})}{\pi_{K_2}}\right] = 0$$

to get $\hat{\epsilon}_{K_1;K_2} = \mathbb{P}_n\left[\frac{R}{\pi_{K_2}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_1}/\pi_{K_1})}{\pi_{K_2}}\right]$, and

$$\mathbb{P}_n\left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_0}/\pi_{K_0} - \hat{\epsilon}_{K_0;K_1}/\pi_{K_1} - \epsilon_{K_0,K_1;K_2}/\pi_{K_2})}{\pi_{K_2}}\right] = 0$$

to get $\hat{\epsilon}_{K_0,K_1;K_2} = \mathbb{P}_n\left[\frac{R}{\pi_{K_2}^2}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_0}/\pi_{K_0} - \hat{\epsilon}_{K_0;K_1}/\pi_{K_1})}{\pi_{K_2}}\right]$. Essentially, the set of potential confounders adjusted in any additional DR covariates included in a new estimator have to be partially ordered with the set of potential confounders adjusted in the DR covariates that were included in the original estimator. For example, an estimator including both $\frac{1}{\pi_{\{X_1\}}}$ and $\frac{1}{\pi_{\{X_2\}}}$ is not well-defined in the CTMLE procedure because neither $\{X_1\} \subseteq \{X_2\}$ nor $\{X_2\} \subseteq \{X_1\}$ is true. We also want to emphasize that one does not have to fluctuate by only one additional potential confounder, e.g. in $\hat{\Psi}_{K_0,K_2}$, one directly fluctuates from the empty set K_0 to the doubleton K_2.

Finally, repeat the above process until one reaches the step to include the DR covariate adjusting for $K_{\text{full}} = \{X_1, \cdots, X_d\}$. CTMLE will perform a relatively complicated procedure to select one of them as the final estimator of Ψ_0.

In general, for an estimator including propensity score model adjusting for at most the potential set of confounders K_m, there are many possible ways to construct it:

$$\text{Entirely nonfluctuated} : \hat{\Psi}_{K_m} = \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_m}}{\pi_{K_m}} \right]$$

$$\text{Fluctuated from } K_i, K_j, \cdots, K_\ell \text{ up to } K_m : \hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m} = \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} + \cdots + \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}} \right]$$

for any integers $0 \leq i < j < \ell < m \leq d$ and $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$ being partially ordered. Here $\hat{\epsilon}_{K_m}$ is the solution of $\mathbb{P}_n \left[\frac{R(Y - b_0 - \bar{\epsilon}_{K_m}/\pi_{K_m})}{\pi_{K_m}} \right] = 0$; $\hat{\epsilon}_{K_i}, \hat{\epsilon}_{K_i;K_j}, \cdots, \hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}$ are the solutions of iteratively solving

$$\begin{aligned}
\mathbb{P}_n \left[\frac{R(Y - b_0 - \bar{\epsilon}_{K_i}/\pi_{K_i})}{\pi_{K_m}} \right] &= 0, \\
\mathbb{P}_n \left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_i}/\pi_{K_i} - \bar{\epsilon}_{K_i;K_j}/\pi_{K_j})}{\pi_{K_j}} \right] &= 0, \\
&\vdots \\
\mathbb{P}_n \left[\frac{R(Y - b_0 - \hat{\epsilon}_{K_i}/\pi_{K_i} - \hat{\epsilon}_{K_i;K_j}/\pi_{K_j} - \cdots - \bar{\epsilon}_{K_i,\cdots,K_\ell;K_m}/\pi_{K_m})}{\pi_{K_m}} \right] &= 0.
\end{aligned}$$

For the true values of these DR coefficients, ϵ_{K_m} solves $\mathbb{E} \left[\frac{R(Y - b_0 - \bar{\epsilon}_{K_m}/\pi_{K_m})}{\pi_{K_m}} \right] = 0$; $\epsilon_{K_i}, \epsilon_{K_i;K_j}, \cdots,$ $\epsilon_{K_i,K_j,\cdots,K_\ell;K_m}$ are the solutions of iteratively solving

$$\begin{aligned}
\mathbb{E} \left[\frac{R(Y - b_0 - \bar{\epsilon}_{K_i}/\pi_{K_i})}{\pi_{K_m}} \right] &= 0, \\
\mathbb{E} \left[\frac{R(Y - b_0 - \epsilon_{K_i}/\pi_{K_i} - \bar{\epsilon}_{K_i;K_j}/\pi_{K_j})}{\pi_{K_j}} \right] &= 0,
\end{aligned}$$

$$\vdots$$

$$\mathbb{E}\left[\frac{R(Y - b_0 - \epsilon_{K_i}/\pi_{K_i} - \epsilon_{K_i;K_j}/\pi_{K_j} - \cdots - \bar{\epsilon}_{K_i,\cdots,K_\ell;K_m}/\pi_{K_m})}{\pi_{K_m}}\right] = 0.$$

To evaluate the deviation of an estimator from the truth Ψ_0, the mean squared error (M.S.E.) of the estimator i.e. $\text{M.S.E.}(\hat{\Psi}_\cdot, \Psi_0) := \mathbb{E}\left[(\hat{\Psi}_\cdot - \Psi_0)^2\right]$ is a natural choice because it can be decomposed into the variance component $\text{var}[\hat{\Psi}_\cdot] := \mathbb{E}\left[(\hat{\Psi}_\cdot - \mathbb{E}[\hat{\Psi}_\cdot])^2\right]$ and bias square component $\text{bias square}[\hat{\Psi}_\cdot] := \left(\mathbb{E}\left[\hat{\Psi}_\cdot - \Psi_0\right]\right)^2$ of the corresponding estimator. If instead one only aims at minimizing the bias of an estimator, then one will be likely to end up constructing the estimator adjusting for all but possibly redundant potential confounders, resulting in variance inflation[72]. To study M.S.E., we first introduce the squared error of some candidate estimator $\hat{\Psi}_\cdot$ standardized by sample size appropriately, in particular by $\sqrt{n}$ due to the $\sqrt{n}$-perturbative nature of the DGP in Definition 2.1.1:

Definition 2.2.2. *The squared error standardized by $\sqrt{n}$ for an estimator $\hat{\Psi}_\cdot$ is denoted by $\zeta_\cdot$ and defined as follows*

$$\zeta_\cdot := \left\{\sqrt{n}(\hat{\Psi}_\cdot - \Psi_0)\right\}^2 \equiv n\left(\hat{\Psi}_\cdot - \Psi_0\right)^2. \tag{2.4}$$

Then the standardized M.S.E. for $\hat{\Psi}_\cdot$ is just $\mathbb{E}[\zeta_\cdot]$. We denote the difference between two standardized squared errors as

$$\zeta_{\cdot,\cdot} := \zeta_{\cdot\cdot} - \zeta_{\cdot\cdot} \tag{2.5}$$

For example, $\zeta_{K_m} = n\left(\hat{\Psi}_{K_m} - \Psi_0\right)^2$ and $\zeta_{K_i,K_j,\cdots,K_\ell,K_m} = n\left(\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m} - \Psi_0\right)^2$.

Before we reveal the actual CTMLE procedure of constructing final estimator for the "target" Ψ_0, we need to define three more concepts: the within-sample prediction risk, the population prediction risk conditioning on the estimated coefficients (abbr. as conditional population prediction risk), and the prediction risk estimated using M-fold cross validation (abbr. as M-CV prediction risk). First, the definition

of within-sample prediction risk is given in Definition 2.2.3 below:

Definition 2.2.3. *After standardized by the sample size, the within-sample prediction risk for an estimator* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined to be*

$$\xi_{K_i,K_j,\cdots,K_\ell,K_m} := n\mathbb{P}_n\left[R\left(Y - b_0 - \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} - \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} - \cdots - \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}}\right)^2\right] \tag{2.6}$$

where $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$. *In addition, we denote the difference between two within-sample prediction risks as*

$$\xi_{\cdots_i\cdot} := \xi_{\cdots} - \xi_{\cdots\cdot} \tag{2.7}$$

This within-sample prediction risk is used in CTMLE as a criterion to decide whether in the next step the estimator should be fluctuated or not. We will elaborate on this when reviewing the CTMLE procedure later.

We next introduce the definition of conditional population prediction risk (Definition 2.2.4):

Definition 2.2.4. *After standardized by the sample size, the conditional population prediction risk of an estimator* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined as*

$$\begin{aligned}&\eta_{K_i,K_j,\cdots,K_\ell,K_m}\\ &:= n\mathbb{E}\left[R\left(Y - b_0 - \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} - \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} - \cdots - \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}}\right)^2 \middle| \hat{\epsilon}_{K_i}, \hat{\epsilon}_{K_i;K_j}, \cdots, \hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right]\end{aligned} \tag{2.8}$$

where $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$. *In addition, we denote the difference between two conditional population prediction risks as*

$$\eta_{\cdots_i\cdot} := \eta_{\cdots} - \eta_{\cdots\cdot} \tag{2.9}$$

This conditional population prediction risk is a conceptualization of the M-CV prediction risk, in which we use one part of the sample (training sample) to estimate the DR coefficients and evaluate the prediction risk using the rest of the sample (validation sample). In the conditional population prediction risk, one estimates the DR coefficients from the given sample but evaluates the prediction risk of this estimator as an oracle with either the knowledge of the true DGP or infinitely amount of independent samples to evaluate the integral exactly.

As conceptually interesting as the conditional population prediction risk, one cannot calculate such quantity with data in practice. Such quantity is usually estimated through nonparametric resampling methods, such as Jackknife and/or cross-validation[83]. We focus on cross-validation in this paper, which is also the method chosen in CTMLE[98]. One first splits the whole sample into M pieces, where M can be any positive integer less than the sample size n but greater than one. With $(M-1)/M$ of the whole dataset, one first estimates the DR coefficients and uses the rest $1/M$ of the dataset to evaluate the prediction risk, and then averages over all the M folds. The M-CV prediction risk is mathematically defined as below:

Definition 2.2.5. *After standardized by the sample size, the M-CV prediction risk for an estimator* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined to be*

$$\eta^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m} := \frac{n}{M}\sum_{t=1}^{M}\mathbb{P}^t_{n/M}\left[R\left(Y-b_0-\frac{\hat{\epsilon}^{\setminus t}_{K_i}}{\pi_{K_i}}-\frac{\hat{\epsilon}^{\setminus t}_{K_i;K_j}}{\pi_{K_j}}-\cdots-\frac{\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}}\right)^2\right] \tag{2.10}$$

where $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$ *and* $\hat{\epsilon}^{\setminus t}$ *indicates that we estimate* ϵ *from the sample not in the t-th group of the whole sample after we divide the whole sample into M non-overlapping groups, each with sample size* n/M*. In addition, we denote the difference between two M-CV prediction risks as*

$$\eta^{\dagger M}_{\cdots;\cdot} := \eta^{\dagger M}_{\cdots} - \eta^{\dagger M}_{\cdot}. \tag{2.11}$$

Large sample distribution of the DR coefficients in CTMLE

After reviewing the algorithm employed in CTMLE, we would like to understand the whole procedure better in a relatively simplified manner. In this paper, we focus on analyzing the large-sample or asymptotic behavior of the algorithm, together with some other alternative approaches such as FIC[12] and procedures related to Lepski's approach[45,46,89]. To achieve such goal, the very first thing that we need to characterize is the asymptotic distribution of the estimated DR coefficients.

As a side note, in Robins et al.[76], the authors mentioned another alternative estimator dubbed as "Joffe's estimator" which was first introduced by Marshall M. Joffee and also enjoys the double-robustness property. The "Joffe's estimators" for Ψ_0 has the following form, analogous to the CTMLE estimators:

$$\hat{\Psi}^{\text{Joffe}}_{K_i,K_j,\cdots,K_\ell,K_m} = \mathbb{P}_n\left[b_0 + \hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right]$$

with the "Joffe's" DR coefficients $\hat{\varepsilon}$'s. We will also use the superscript Joffeto denote any quantity used for "Joffe's" estimators. Similar to the CTMLE estimators, we also define "Joffe's" DR coefficients that can also be constructed using the strategy in CTMLE as follows:

$$\begin{aligned}
\hat{\varepsilon}_{K_i} &= \mathbb{P}_n\left[\frac{R}{\pi_{K_i}}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R}{\pi_{K_i}}(Y-b_0)\right], \\
\hat{\varepsilon}_{K_i;K_j} &= \mathbb{P}_n\left[\frac{R}{\pi_{K_j}}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R}{\pi_{K_j}}(Y-b_0-\hat{\varepsilon}_{K_i})\right] \\
&= \mathbb{P}_n\left[\frac{R}{\pi_{K_j}}\right]^{-1} \mathbb{P}_n\left[\frac{R}{\pi_{K_j}}(Y-b_0)\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_i}}\right]^{-1} \mathbb{P}_n\left[\frac{R}{\pi_{K_i}}(Y-b_0)\right], \\
&\vdots \\
\hat{\varepsilon}_{K_i,K_j,\cdots,K_k,K_\ell;K_m} &= \mathbb{P}_n\left[\frac{R}{\pi_{K_m}}\right]^{-1} \cdot \mathbb{P}_n\left[\frac{R}{\pi_{K_m}}\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots,K_k;K_\ell}\right)\right] \\
&= \mathbb{P}_n\left[\frac{R}{\pi_{K_m}}\right]^{-1} \mathbb{P}_n\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}}\right]^{-1} \mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right]
\end{aligned} \tag{2.12}$$

where the simplification follows from the same argument as in the case of the true CTMLE DR coefficients for Joffe's estimator even in finite sample. The same simplification also applies to the true DR coefficients for Joffe's estimator:

$$\begin{aligned}
\varepsilon_{K_i} &= \mathbb{E}\left[\frac{R}{\pi_{K_i}}\right]^{-1} \cdot \mathbb{E}\left[\frac{R}{\pi_{K_i}}\left(Y - b_0\right)\right] = \mathbb{E}\left[\frac{R}{\pi_{K_i}}\left(Y - b_0\right)\right], \\
\varepsilon_{K_i;K_j} &= \mathbb{E}\left[\frac{R}{\pi_{K_j}}\right]^{-1} \cdot \mathbb{E}\left[\frac{R}{\pi_{K_j}}\left(Y - b_0 - \varepsilon_{K_i}\right)\right] \\
&= \mathbb{E}\left[\frac{R}{\pi_{K_j}}\left(Y - b_0\right)\right] - \mathbb{E}\left[\frac{R}{\pi_{K_i}}\left(Y - b_0\right)\right], \\
&\vdots \\
\varepsilon_{K_i,K_j,\cdots,K_k,K_\ell;K_m} &= \mathbb{E}\left[\frac{R}{\pi_{K_m}}\right]^{-1} \cdot \mathbb{E}\left[\frac{R}{\pi_{K_m}}\left(Y - b_0 - \varepsilon_{K_i} - \varepsilon_{K_i;K_j} - \cdots - \varepsilon_{K_i,K_j,\cdots,K_k;K_\ell}\right)\right] \\
&= \mathbb{E}\left[\frac{R}{\pi_{K_m}}\left(Y - b_0\right)\right] - \mathbb{E}\left[\frac{R}{\pi_{K_\ell}}\left(Y - b_0\right)\right]
\end{aligned} \tag{2.13}$$

$$\tag{2.14}$$

Similarly, we also introduce $\xi_{\cdot}^{\text{Joffe}}$, $\eta_{\cdot}^{\text{Joffe}}$ and $\eta_{\cdot}^{\dagger M\ \text{Joffe}}$ for the standardized conditional population prediction risk and M-CV prediction risk respectively. According to Chapter 27 of van der Laan & Rose[102], the prediction-risk-related quantities for Joffe's estimators are defined as follows:

Definition 2.2.6. *After standardized by the sample size, the within-sample prediction risk for an estimator* $\hat{\Psi}^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined to be*

$$\begin{aligned}
&\xi^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m} \\
&:= n\mathbb{P}_n\left[\frac{R}{\pi_{K_m}}\left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2\right]
\end{aligned} \tag{2.15}$$

Similarly, the conditional population prediction risk for an estimator $\hat{\Psi}^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined to be*

$$\begin{aligned}&\eta^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m}\\ &:= n\mathbb{E}\left[\frac{R}{\pi_{K_m}}\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2\middle|\,\hat{\varepsilon}_{K_i},\hat{\varepsilon}_{K_i;K_j},\cdots,\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right]\end{aligned} \tag{2.16}$$

and the M-CV prediction risk for an estimator $\hat{\Psi}^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m}$ *is defined to be*

$$\eta^{\dagger M\ Joffe}_{K_i,K_j,\cdots,K_\ell,K_m} := \frac{n}{M}\sum_{t=1}^{M}\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_m}}\left(Y-b_0-\hat{\varepsilon}^{\setminus t}_{K_i}-\hat{\varepsilon}^{\setminus t}_{K_i;K_j}-\cdots-\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2\right] \tag{2.17}$$

where $K_i\subseteq K_j\subseteq\cdots\subseteq K_\ell\subseteq K_m$.

Remark 2.2.7. *The reason that one use different weights in the prediction risk-related quantities in Joffe's estimator compared to CTMLE estimator is that according to Chapter 27 of van der Laan & Rose[102], the derivative of the summand in the prediction risks should be proportional to the estimating equations used to estimate* $\hat{\varepsilon}$*'s.*

Then under Definition 2.1.1, together with the above algebraic simplifications, we derive the following asymptotic distributions for $\sqrt{n}\hat{\epsilon}_\cdot$ and $\sqrt{n}\hat{\varepsilon}_\cdot$:

Proposition 2.2.8. *As* $n\to\infty$*, under Definition 2.1.1, the CTMLE DR coefficients have the following asymptotic Gaussian distributions, for* $K_i\subseteq K_j\subseteq\cdots\subseteq K_\ell\subseteq K_m\subseteq K_{full}$

$$\begin{aligned}\sqrt{n}\hat{\epsilon}_{K_i} &\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]^{-1}\left\{v^{1/2}_{K_i,K_i}Z_{K_i}+p_{K_i}+q\right\}\\ \sqrt{n}\hat{\epsilon}_{K_i;K_j} &\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1}\left\{v^{1/2}_{K_j,K_j}Z_{K_j}+p_{K_j}-v^{1/2}_{K_i,K_i}Z_{K_i}-p_{K_i}\right\}\\ &\vdots\\ \sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m} &\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\left\{v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right\},\end{aligned} \tag{2.18}$$

and similarly, the "Joffe's" DR coefficients have the following asymptotic Gaussian distributions

$$\begin{aligned}
\sqrt{n}\hat{\varepsilon}_{K_i} &\xrightarrow{d} v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q \\
\sqrt{n}\hat{\varepsilon}_{K_i;K_j} &\xrightarrow{d} v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i} \\
&\vdots \\
\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m} &\xrightarrow{d} v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}
\end{aligned} \tag{2.19}$$

where $Z_{K_i}, Z_{K_j}, \cdots, Z_{K_\ell}, Z_{K_m}$ are multivariate Gaussian random variables with mean 0, unit variance, and covariance between Z_{K_r} and Z_{K_s} is $\frac{v_{K_r,K_s}}{v_{K_r,K_r}^{1/2} v_{K_s,K_s}^{1/2}}$ for any set of potential confounders K_r, K_s, throughout the paper.

Proof. The only difference between Joffe's DR coefficients and CTMLE DR coefficients is the factor $\mathbb{P}_n\left[\frac{R}{\pi_{K_m}}\right]^{-1}$ and $\mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1}$, and therefore the asymptotic distribution for the Joffe's DR coefficient is just the counterpart for the CTMLE DR coefficient multiplied by $\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]$. Now we first derive the asymptotic distribution of the kernel component $\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right]$ for some arbitrary K_h. Using central limit theorem, as $n \to \infty$

$$\sqrt{n}\left(\mathbb{P}_n\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right] - \mathbb{E}\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right]\right) \xrightarrow{d} \text{Normal}\left(0, \text{AsymVar}\left(\frac{R}{\pi_{K_h}}(Y-b_0)\right)\right).$$

Following the notations in Definition 2.1.1, $\text{AsymVar}\left(\frac{R}{\pi_{K_h}}(Y-b_0)\right)$ is defined as v_{K_h,K_h}, hence the distribution can be rewritten as $v_{K_h,K_h}^{1/2} Z_{K_h}$ for some standard normal random variable Z_{K_h} as a standardized version of the asymptotic distribution of $\sqrt{n}\left(\mathbb{P}_n\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right] - \mathbb{E}\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right]\right)$. Furthermore, since $\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right]$ is defined as $p_{K_h} + q$ of order $O(1)$ as in Definition 2.1.1, we can rearrange

the term to get

$$\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_h}}(Y-b_0)\right] \xrightarrow{d} v_{K_h,K_h}^{1/2} Z_{K_h} + p_{K_h} + q.$$

Then the asymptotic distribution of $\sqrt{n}\hat{\epsilon}_{K_i}$ directly follows from Slutsky's theorem[85]:

$$\begin{aligned}\sqrt{n}\hat{\epsilon}_{K_i} &= \mathbb{P}_n\left[\frac{R}{\pi_{K_i}^2}\right]^{-1} \cdot \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_i}}(Y-b_0)\right] \\ &\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]^{-1} \cdot \left\{v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q\right\}.\end{aligned}$$

Similarly, for the more general form $\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}$:

$$\begin{aligned}&\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m} \\ &= \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1} \\ &\quad\times\left\{\begin{array}{l}\mathbb{P}_n\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_i}}\right]\hat{\epsilon}_{K_i} - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_j}}\right]\hat{\epsilon}_{K_i;K_j} - \cdots - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_k}}\right]\hat{\epsilon}_{K_i,K_j,\cdots;K_k} \\ -\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right] + \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_i}}\right]\hat{\epsilon}_{K_i} \\ +\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_j}}\right]\hat{\epsilon}_{K_i;K_j} + \cdots \\ +\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_k}}\right]\hat{\epsilon}_{K_i,K_j,\cdots;K_k}\end{array}\right\} \\ &= \mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1}\left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right]\right) \\ &\quad -\mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1}\left\{\sqrt{n}\hat{\epsilon}_{K_i}\left(\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_i}}\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_i}}\right]\right)\right\} \\ &\quad -\mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1}\left\{\sqrt{n}\hat{\epsilon}_{K_i;K_j}\left(\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_i}}\right] - \mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_i}}\right]\right)\right\} - \cdots\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{P}_n\left[\frac{R}{\pi_{K_m}^2}\right]^{-1}\left\{\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots;K_k}\left(\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_k}}\right]-\mathbb{P}_n\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}^2}\right]^{-1}\mathbb{P}_n\left[\frac{R}{\pi_{K_\ell}\pi_{K_k}}\right]\right)\right\}\\
&\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{\left(v_{K_m,K_m}^{1/2}Z_{K_m}+p_{K_m}+q\right)-\left(v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}+p_{K_\ell}+q\right)\right\}\\
&\equiv\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{\left(v_{K_m,K_m}^{1/2}Z_{K_m}+p_{K_m}\right)-\left(v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}+p_{K_\ell}\right)\right\}.
\end{aligned}
$$

Finally, we derive the asymptotic covariance between $\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_r}}(Y-b_0)\right]$ and $\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_s}}(Y-b_0)\right]$:

$$
\begin{aligned}
&\operatorname{Cov}\left[\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_r}}(Y-b_0)\right],\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_s}}(Y-b_0)\right]\right]\\
&=\operatorname{Cov}\left[\frac{R}{\pi_{K_r}}(Y-b_0),\frac{R}{\pi_{K_s}}(Y-b_0)\right]\\
&=\mathbb{E}\left[\frac{R}{\pi_{K_r}\pi_{K_s}}(Y-b_0)\right]-\mathbb{E}\left[\frac{R}{\pi_{K_s}}(Y-b_0)\right]\mathbb{E}\left[\frac{R}{\pi_{K_r}}(Y-b_0)\right]\\
&\to\mathbb{E}\left[\frac{R}{\pi_{K_r}\pi_{K_s}}(Y-b_0)\right]
\end{aligned}
$$

because $\mathbb{E}\left[\frac{R}{\pi_{K_s}}(Y-b_0)\right]$ and $\mathbb{E}\left[\frac{R}{\pi_{K_r}}(Y-b_0)\right]$ are of order $O(n^{-1/2})$ due to the assumptions in Definition 2.1.1 and by definition $v_{K_r,K_s}:=\mathbb{E}\left[\frac{R}{\pi_{K_r}\pi_{K_s}}(Y-b_0)\right]$. □

Remark 2.2.9. *One can directly see that the asymptotic distribution of the estimated CTMLE DR coefficient is the same as its counterpart estimated "Joffe's" DR coefficient up to a scale transformation by the harmonic mean of the propensity score model. Using*

$$
\sqrt{n}\hat{\epsilon}_{K_s,K_r,\cdots,K_u,K_w}\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_w}}\right]^{-1}\left\{v_{K_w,K_w}^{1/2}Z_{K_w}+p_{K_w}-v_{K_u,K_u}^{1/2}Z_{K_u}-p_{K_u}\right\}
$$

with $K_s\subseteq K_r\subseteq K_u\subseteq K_w$ as an example, the scale factor only depends on the largest propensity score model being adjusted in the DR covariate, which is π_{K_w}. In addition, for the general form $\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}$, the asymptotic distribution only depends on K_ℓ and K_m, not on the other set of potential confounders.

This is essentially saying that, in asymptotics, the estimated DR coefficients are "memoryless" and it only matters what "ingredients" we are adding into the model and from which point we are adding "ingredients" into the model.

Large sample distributions for the standardized squared errors

An immediate consequence of Proposition 2.2.8 is that we can use plug-in the asymptotic distribution for the estimated DR coefficients to obtain the asymptotic distributions of the squared errors of the candidate CTMLE or Joffe's estimators for Ψ_0 in the following result (Proposition 2.2.10):

Proposition 2.2.10. *As $n \to \infty$, under Definition 2.1.1, the standardized squared errors (Definition 2.2.2) of all the CTMLE and Joffe's estimators including DR covariates adjusting for at most the set of potential confounders K_m have the same asymptotic distribution:*

$$\begin{aligned} \zeta_{K_m} &\xrightarrow{d} \tilde{\zeta}_{K_m} := \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2, \\ \zeta_{K_i,K_j,\cdots,K_\ell,K_m} &\xrightarrow{d} \tilde{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m} := \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2. \end{aligned} \tag{2.20}$$

and

$$\begin{aligned} \zeta^{Joffe}_{K_m} &\xrightarrow{d} \tilde{\zeta}^{Joffe}_{K_m} := \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2, \\ \zeta^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m} &\xrightarrow{d} \tilde{\zeta}^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m} := \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2. \end{aligned} \tag{2.21}$$

for any sets of potential confounders satisfying $K_1 \subseteq K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$. Since the asymptotic distribution of the above squared error losses is the same, we denote $\tilde{\zeta}^*_{K_m} := \tilde{\zeta}_{K_m} = \tilde{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m} = \tilde{\zeta}^{Joffe}_{K_m} = \tilde{\zeta}^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m}$ as the asymptotic distribution for squared errors in this equivalence class where the equivalence relation is defined as having the same asymptotic distribution of the squared error. When $K_m \supseteq K_{d^\star}$ ($K_{d^\star}$ is defined to be the true set of confounders), $\tilde{\zeta}^*_{K_m} = v_{K_m,K_m} Z^2_{K_m}$ i.e. the CTMLE/Joffe's

estimators in this equivalence class are asymptotically unbiased for Ψ_0.

Proof. We first derive the asymptotic distributions of ζ_{K_m} and $\zeta_{K_m}^{\text{Joffe}}$:

$$\begin{aligned}
\zeta_{K_m} &:= n\left(\hat{\Psi}_{K_m} - \Psi_0\right)^2 = n\left(\mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_m}}{\pi_{K_m}}\right] - \mathbb{E}[Y]\right)^2 \\
&= \left(\mathbb{P}_n\left[\frac{1}{\pi_{K_m}}\right]\sqrt{n}\hat{\epsilon}_{K_m} - \sqrt{n}\mathbb{E}[Y - b_0]\right)^2 \\
&\xrightarrow{d} \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} + q - q\right)^2 \\
&\equiv \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)^2 := \tilde{\zeta}_{K_m}, \\
\zeta_{K_m}^{\text{Joffe}} &:= n\left(\hat{\Psi}_{K_m}^{\text{Joffe}} - \Psi_0\right)^2 = n\left(\mathbb{P}_n\left[b_0 + \hat{\varepsilon}_{K_m}\right] - \mathbb{E}[Y]\right)^2 \\
&= \left(\sqrt{n}\hat{\varepsilon}_{K_m} - \sqrt{n}\mathbb{E}[Y - b_0]\right)^2 \\
&\xrightarrow{d} \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} + q - q\right)^2 \\
&\equiv \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)^2 := \tilde{\zeta}_{K_m}^{\text{Joffe}},
\end{aligned}$$

where the convergence in distribution steps follow from the results in Proposition 2.2.10 and Slutsky's theorem. For $\zeta_{K_i,K_j,\cdots,K_\ell,K_m}$ and $\zeta_{K_i,K_j,\cdots,K_\ell,K_m}^{\text{Joffe}}$, we again use a telescoping argument:

$$\begin{aligned}
&\zeta_{K_i,K_j,\cdots,K_\ell,K_m} \\
&:= n\left(\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m} - \Psi_0\right)^2 \\
&= n\left(\mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} + \cdots + \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_k;K_\ell}}{\pi_{K_\ell}} + \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}}\right] - \mathbb{E}[Y]\right)^2 \\
&= \left(\mathbb{P}_n\left[\frac{1}{\pi_{K_i}}\right]\sqrt{n}\hat{\epsilon}_{K_i} + \mathbb{P}_n\left[\frac{1}{\pi_{K_j}}\right]\sqrt{n}\hat{\epsilon}_{K_i;K_j} + \cdots + \mathbb{P}_n\left[\frac{1}{\pi_{K_m}}\right]\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m} - q\right)^2 \\
&\xrightarrow{d} \left(\begin{array}{l}
\left\{v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q\right\} + \left\{v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i}\right\} + \cdots \\
+ \left\{v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} + p_{K_\ell} - v_{K_k,K_k}^{1/2} Z_{K_k} - p_{K_k}\right\} + \left\{v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right\} - q
\end{array}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\equiv \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)^2 := \tilde{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m}, \\
&\overset{\text{Joffe}}{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m} \\
&:= n\left(\hat{\Psi}^{\text{Joffe}}_{K_i,K_j,\cdots,K_\ell,K_m} - \Psi_0\right)^2 \\
&= n\left(\mathbb{P}_n\left[b_0 + \hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots,K_k;K_\ell} + \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right] - \mathbb{E}[Y]\right)^2 \\
&= \left(\sqrt{n}\hat{\varepsilon}_{K_i} + \sqrt{n}\hat{\varepsilon}_{K_i;K_j} + \cdots + \sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m} - q\right)^2 \\
&\xrightarrow{d} \left(\begin{array}{l} \left\{v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q\right\} + \left\{v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i}\right\} + \cdots \\ + \left\{v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} + p_{K_\ell} - v_{K_k,K_k}^{1/2} Z_{K_k} - p_{K_k}\right\} + \left\{v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right\} - q \end{array}\right)^2 \\
&\equiv \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)^2 := \tilde{\zeta}^{\text{Joffe}}_{K_i,K_j,\cdots,K_\ell,K_m}
\end{aligned}
$$

where again the convergence in distribution steps follow from the results in Proposition 2.2.10 and Slutsky's theorem. □

Remark 2.2.11. *The most important part of Proposition 2.2.10 is that in the limit of $n \to \infty$ and in terms of the squared error, there is no difference among the different estimators constructed in CTMLE or Joffe's way including DR covariate adjusting for at most the set of potential confounders K_m, where K_m can be arbitrary. By taking expectations over the Gaussian random variable $Z_\cdot$'s, the standardized asymptotic M.S.E. (A.M.S.E.) is $v_{K_m,K_m} + p_{K_m}^2$, and the standardized asymptotic variance and bias square are v_{K_m,K_m} and $p_{K_m}^2$ respectively. Under homoscedasticity, and without loss of generality assuming $v_\cdot = 1$, if one adjust for more potential confounders than $K_{d^\star}$, then the asymptotic variance $v_{K_m,K_m} = \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right] \geq \mathbb{E}\left[\frac{1}{\pi_{K_{d^\star}}}\right]$ when $K_{d^\star} \subseteq K_m$. Hence over-adjustment results in variance inflation under homoscedasticity. Under general heteroscedastic conditional variance $v_{K_{d^\star}}$, there is no fixed ordering in terms of the asymptotic variance and one has to figure out the order case-by-case.*

Remark 2.2.12. *Another interesting observation is that if we focus on all the CTMLE estimators, i.e.*

whenever we add another DR covariate, the enlarged set of potential confounders have to contain the original set of potential confounders used to construct the old estimator. Under such scenario, in Section A.3, we prove that if instead of using the greedy estimation strategy proposed in CTMLE, every time we re-estimate the DR coefficients using a least-square approach, the asymptotic distributions for the squared errors are invariant to this change in estimation strategy. The key of such equivalence is using the Sherman-Morrison formula[84] or Woodbury identity[110].

When talking about asymptotics, we use A.M.S.E. for asymptotic M.S.E.. We define the oracle A.M.S.E. as the minimum A.M.S.E. out of all candidate estimators: Oracle A.M.S.E. $:=$ min{A.M.S.E....}. Similarly, we define the maximum A.M.S.E.as Max A.M.S.E. $:=$ max{A.M.S.E....}. After standardized by the sample size, define $\tilde{\zeta}^*_{\text{oracle}} := \min\{\tilde{\zeta}^*_{K_0}, \cdots, \tilde{\zeta}^*_{K_{\text{full}}}\}$ and $\tilde{\zeta}^*_{\max} := \max\{\tilde{\zeta}^*_{K_0}, \cdots, \tilde{\zeta}^*_{K_{\text{full}}}\}$.

Complexity of the family of CTMLE estimators

The above results have an interesting implication on the algorithmic aspects of the CTMLE procedure. Because regardless of the specific form, any CTMLE estimators sharing the same maximal set of potential confounders in the DR covariates also share the same asymptotic squared error loss, we can view these estimators as an equivalence class of estimators in terms of their asymptotic squared error loss. For instance, suppose we only care about the asymptotic squared error distribution, then $\hat{\Psi}_{K_\ell, K_m}$ and $\hat{\Psi}_{K_m}$ are indistinguishable. Similar to the notation in Section 2.2, we also add the superscript $*$ to denote the equivalence class of the estimators including at most the set of potential confounders K_m as $\hat{\Psi}^*_{K_m}$, indicating that their asymptotic distributions of the squared errors are equal. Apparently, if one only use statistic based on the shared quantities within each equivalence class just so defined to select estimator, the number of possible estimators will be much smaller. This actually serves as one motivation for using FIC or Lepski-related methods to select estimator, which will be discussed later in Section 2.5 and Section 2.6.

Now we will count the total number of CTMLE estimators concretely. Given the complicated procedures, an interesting digress is to count how many possible estimators that CTMLE procedure needs to survey. Let's first count the number of propensity score models one can fit in principle with d potential confounders, which should be the cardinality of the power set of X, equal to 2^d in total. Then according to the rule of constructing estimators based on the DR covariates $\frac{1}{\pi}$., then the grand total number of estimators that one can construct in principle would be 2^{2^d} if we do not restrict ourselves to the estimators constructed in CTMLE.

When counting all the CTMLE estimators, the only constraint is that the set of potential confounders adjusted in the DR covariates have to be partially ordered by the subset relation "$\subseteq$": for instance, for

$$\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m} = \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} + \cdots + \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}} \right],$$

we have to make sure $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$. Using the standard language in combinatorics, $K_i \to K_j \to \cdots \to K_\ell \to K_m$ forms a chain consisting partially ordered sets. Denote the total number of CTMLE estimators when one has collected d potential confounders as $N_{\text{CTMLE}}(d)$. Then if one starts from $K_0 \equiv \emptyset$, there will be

$$N_{\text{CTMLE}}(d) = N_{\text{CTMLE}}(d,1) + \cdots + N_{\text{CTMLE}}(d,d)$$

where $N_{\text{CTMLE}}(d,\ell)$ is the number of CTMLE estimators with ℓ DR covariates for $\ell = 1, \cdots, d$. Therefore, counting each $N_{\text{CTMLE}}(d,\ell)$ is mathematically equivalent to counting the number of all possible chains of length ℓ out of the power set of a set with cardinality d and $N_{\text{CTMLE}}(d)$ is just the total number of all possible chains from length 1 to length d.

Based on existing formula in the combinatorics literature for counting the total number of chains of

all the possible partially ordered sets from the power sets of $\{1, \cdots, d\}$[57], $N_{\text{CTMLE}}(d)$ has the following analytical representation

$$N_{\text{CTMLE}}(d) = 2\sum_{j=2}^{\infty} j^d 2^{-j}. \tag{2.22}$$

Example 2.2.13. *$N_{CTMLE}(1) = 3$, $N_{CTMLE}(2) = 11$, $N_{CTMLE}(3) = 51$, $N_{CTMLE}(4) = 299$, and $N_{CTMLE}(5) = 2163$.*

This can be easily evaluated in any programming language by truncating the series at a reasonably large number.

As a comparison, the total number of equivalence classes of estimators sharing the same asymptotic distributions of squared error is simply $O(2^d)$. When $d \geq 2$, $N_{\text{CTMLE}}(d) \gg 2^d$.

2.3 Large sample distributions for the standardized conditional population prediction risks and within-sample prediction risks

Since the CTMLE algorithm involves two key steps - one forward-selection step using the within-sample prediction risks and a final selection step using M-fold cross-validation with so-called M-CV prediction risks of the family of candidate estimators. In this section, we devote our efforts to derive the asymptotic distribution of the standardized within-sample prediction risks (Definition 2.2.3) and a "infinite-sample" conceptualization of M-CV prediction risks as if one knows the true DGP or equivalently has infinite amount of the validation sample to evaluate the integration in the definition of the prediction risks. To be concrete, we conceptualize the M-CV prediction risks into the conditional population prediction risks as defined in Definition 2.2.4. We first derive the asymptotic distribution for the standardized within-sample prediction risks, and then for the standardized conditional population prediction risks. We defer the derivation for the standardized M-CV prediction risk and its limit as $n \to \infty$ (with abuse of terminology,

we call it ∞-CV prediction risk) to the next section (Section 2.4). We also display the results for the Joffe's estimator in parallel to those for CTMLE estimators. Interestingly, unlike the squared errors, the two depart as the calculations below for the prediction risks will show.

Large sample distribution for the standardized within-sample prediction risks

In this section, we derive the asymptotic distribution of the standardized within-sample prediction risk difference between the "two nested estimators with only one degree of freedom difference". What we mean by "two nested estimators with only one degree of freedom difference" is the following: Consider an estimator $\hat{\Psi}_{K_i,K_j} := \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} \right]$, then $\hat{\Psi}_{K_i,K_j}$ and the estimator

$$\hat{\Psi}_{K_i,K_j,K_k} := \mathbb{P}_n \left[b_0 + \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} + \frac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}} + \frac{\hat{\epsilon}_{K_i,K_j;K_k}}{\pi_{K_k}} \right]$$

fluctuated one time from $\hat{\Psi}_{K_i,K_j}$ are nested estimators with only one degree of freedom difference. Their standardized within-sample prediction risk difference is $\xi_{K_i,K_j,K_k:K_i,K_j}$ as defined in Definition 2.2.3. For an estimator of the form $\hat{\Psi}_{K_i}$, we compare it to $\mathbb{P}_n[b_0]$ and denote their standardized within-sample prediction risk difference as $\xi_{K_i:} = n\mathbb{P}_n \left[R(Y - b_0 - \hat{\epsilon}_{K_i}/\pi_{K_i})^2 - R(Y - b_0)^2 \right]$.

Now we have the following result regarding the asymptotic distribution of the standardized within-sample prediction risk difference between the two nested estimators with only one degree of freedom difference. In parallel, we will show how Joffe's estimators differ from CTMLE estimators in terms of the prediction risk comparisons in the Remark 2.3.3.

Proposition 2.3.1. *As $n \to \infty$, under Definition 2.1.1, the standardized within-sample prediction risk difference between the two nested estimators with only one degree of freedom differenceare of the following*

form:

$$\begin{aligned}
\xi_{K_m:} \xrightarrow{d} \tilde{\xi}_{K_m:} &:= -\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} + q \right\}^2, \\
\xi_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} &\xrightarrow{d} \tilde{\xi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= -\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell} \right\}^2.
\end{aligned} \tag{2.23}$$

where $K_1 \subseteq K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$.

Proof. We first derive results for $\xi_{K_m:}$:

$$\begin{aligned}
\xi_{K_m:} &:= n\mathbb{P}_n \left[R\left(Y - b_0 - \hat{\epsilon}_{K_m}/\pi_{K_m}\right)^2 - R(Y-b_0)^2 \right] \\
&= \mathbb{P}_n \left[\frac{R}{\pi^2_{K_m}} \right] n\hat{\epsilon}^2_{K_m} - 2\sqrt{n}\hat{\epsilon}_{K_m}\sqrt{n}\mathbb{P}_n \left[\frac{R}{\pi_{K_m}}(Y-b_0) \right] \\
&= \mathbb{P}_n \left[\frac{R}{\pi^2_{K_m}} \right] n\hat{\epsilon}^2_{K_m} - 2\mathbb{P}_n \left[\frac{R}{\pi^2_{K_m}} \right] n\hat{\epsilon}^2_{K_m} \\
&= -\mathbb{P}_n \left[\frac{R}{\pi^2_{K_m}} \right] n\hat{\epsilon}^2_{K_m} \\
&\xrightarrow{d} -\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} + q \right\}^2
\end{aligned}$$

where the line for the convergence in distribution follows from Proposition 2.2.8 and Slutsky's theorem.

Then for the general $\xi_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$:

$$\begin{aligned}
&\xi_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= n\mathbb{P}_n \left[\begin{array}{l} R\left(Y - b_0 - \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} - \frac{\hat{\epsilon}_{K_i:K_j}}{\pi_{K_j}} - \cdots - \frac{\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}}{\pi_{K_\ell}} - \frac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell:K_m}}{\pi_{K_m}}\right)^2 \\ -R\left(Y - b_0 - \frac{\hat{\epsilon}_{K_i}}{\pi_{K_i}} - \frac{\hat{\epsilon}_{K_i:K_j}}{\pi_{K_j}} - \cdots - \frac{\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}}{\pi_{K_\ell}}\right)^2 \end{array} \right] \\
&= n\mathbb{P}_n \left[\frac{R}{\pi^2_{K_m}} \right] \hat{\epsilon}^2_{K_i,K_j,\cdots,K_\ell;K_m} - 2\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\mathbb{P}_n \left[\frac{R}{\pi_{K_m}}(Y-b_0) \right]
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\mathbb{P}_n\left[\frac{R}{\pi^2_{K_i}}\right]\hat{\epsilon}_{K_i}+\cdots+\mathbb{P}_n\left[\frac{R}{\pi^2_{K_\ell}}\right]\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}\right)\\
&\xrightarrow{d}\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\times\\
&\left\{\begin{array}{l}\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right)^2\\ -2\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\\ +2\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right)\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}+q\right)\end{array}\right\}\\
&=-\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)^2.
\end{aligned}
$$

where the line for the convergence in distribution follows from Proposition 2.2.8, Slutsky's theorem and a telescoping summation argument. □

Remark 2.3.2. *The standardized (asymptotic) within-sample prediction risk difference between the two nested estimators with only one degree of freedom difference is always negative, similar to the likelihood ratio statistic. Using the result above (Proposition 2.3.1), one can formulate the within-sample prediction risk difference between any two CTMLE estimators by taking the summation over the form given in eq.* (2.23)*. For example, $\tilde{\xi}_{K_i,K_j,K_\ell:K_i}$ with $K_i\subseteq K_j\subseteq K_\ell$ can be calculated as*

$$
\begin{aligned}
&\tilde{\xi}_{K_i,K_j,K_\ell:K_i}=\tilde{\xi}_{K_i,K_j,K_\ell:K_i,K_j}+\tilde{\xi}_{K_i,K_j:K_i}\\
&=-\mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]^{-1}\cdot\left\{v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}-v^{1/2}_{K_j,K_j}Z_{K_j}-p_{K_j}\right\}^2\\
&\quad-\mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1}\cdot\left\{v^{1/2}_{K_j,K_j}Z_{K_j}+p_{K_j}-v^{1/2}_{K_i,K_i}Z_{K_i}-p_{K_i}\right\}^2.
\end{aligned}
$$

Another example is to compare $\hat{\Psi}_{K_j}$ and $\hat{\Psi}_{K_i,K_j}$, as such comparison is executed in CTMLE to decide whether $\hat{\Psi}_{K_j}$ or $\hat{\Psi}_{K_i,K_j}$ survives as the candidate estimator adjusting for at most the set of potential con-

founders K_j:

$$\begin{aligned}
&\tilde{\xi}_{K_i,K_j:K_j} = \tilde{\xi}_{K_i,K_j:K_i} + \tilde{\xi}_{K_i:} - \tilde{\xi}_{K_j:} \\
=& -\mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1} \cdot \left\{ v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i} \right\}^2 - \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]^{-1} \cdot \left\{ v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q \right\}^2 \\
&+ \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1} \cdot \left\{ v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} + q \right\}^2 .
\end{aligned}$$

Another relevant example is to compare $\hat{\Psi}_{K_i,K_j}$ and $\hat{\Psi}_{K_i,K_j'}$, as such comparison is executed in CTMLE to decide whether one set of $K_i \subseteq K_j$ or the other set $K_i \subseteq K_j'$ should be further included to construct the candidate estimator adjusting for at most the set of potential confounders with $|K_j| = |K_j'|$ dimensions.

$$\begin{aligned}
&\tilde{\xi}_{K_i,K_j:K_i,K_j'} = \tilde{\xi}_{K_i,K_j:K_i} - \tilde{\xi}_{K_i,K_j':K_i} \\
=& -\mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1} \cdot \left\{ v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i} \right\}^2 \\
&+ \mathbb{E}\left[\frac{1}{\pi_{K_j'}}\right]^{-1} \cdot \left\{ v_{K_j',K_j'}^{1/2} Z_{K_j'} + p_{K_j'} - v_{K_i,K_i}^{1/2} Z_{K_i} - p_{K_i} \right\}^2 .
\end{aligned}$$

Remark 2.3.3. *One might wonder how the within-sample prediction risks for the Joffe's estimators look like compared to the CTMLE estimators, given that they share the same structure of the asymptotic distributions for the squared errors. We start from the simple calculation for $\xi_{K_m:}^{Joffe}$:*

$$\begin{aligned}
&\xi_{K_m:}^{Joffe} := n\mathbb{P}_n \left[\frac{R}{\pi_{K_m}} (Y - b_0 - \hat{\varepsilon}_{K_m})^2 - R(Y - b_0)^2 \right] \\
=& \mathbb{P}_n \left[\frac{R}{\pi_{K_m}}\right] n\hat{\varepsilon}_{K_m}^2 - 2\sqrt{n}\hat{\varepsilon}_{K_m}\sqrt{n}\mathbb{P}_n \left[\frac{R}{\pi_{K_m}}(Y - b_0)\right] + n\mathbb{P}_n \left[R\left(\frac{1-\pi_{K_m}}{\pi_{K_m}}\right)(Y - b_0)^2 \right] \\
=& \left(\mathbb{P}_n \left[\frac{R}{\pi_{K_m}}\right] - 2 \right) n\hat{\varepsilon}_{K_m}^2 + n\mathbb{P}_n \left[R\left(\frac{1-\pi_{K_m}}{\pi_{K_m}}\right)(Y - b_0)^2 \right] .
\end{aligned}$$

Under the assumptions adopted in Definition 2.1.1, as in Proposition 2.3.1, the first term converges in distribution to $-\left\{v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} + q\right\}^2$. *For the second term, the expectation of the quantity within the empirical mean operator is*

$$\begin{aligned}
&\mathbb{E}\left[R\left(\frac{1-\pi_{K_m}}{\pi_{K_m}}\right)(Y-b_0)^2\right] \\
&= \mathbb{E}\left[\left(\frac{1-\pi_{K_m}}{\pi_{K_m}}\right)\mathbb{E}\left[R(Y-b_0)^2\middle| K_{full}\right]\right] \\
&= \mathbb{E}\left[\frac{(1-\pi_{K_m})\pi_{K_{d_R}}}{\pi_{K_m}}\mathbb{E}\left[(Y-b_0)^2|K_{d_Y}\right]\right] \\
&\overset{n\to\infty}{\to} \mathbb{E}\left[\frac{(1-\pi_{K_m})\pi_{K_{d_R}} v_{K_{d_Y}}}{\pi_{K_m}}\right] = O(1)
\end{aligned}$$

even under homoscedasticity because $\mathbb{E}[(Y-b_0)^2|K_{d_Y}] = v_{K_{d_Y}} + (\mathbb{E}[(Y-b_0)|K_{d_Y}])^2 = v_{K_{d_Y}} + O\left(n^{-1}\right) \overset{n\to\infty}{\to} v_{K_{d_Y}} = O(1)$. *Hence unless we further assume the conditional outcome variance shrinks to zero as sample size increases, the second term will blow up as* $n \to \infty$. *To find an appropriate standardizing factor, one notice that for the second term*

$$\begin{aligned}
&\frac{1}{n} n\mathbb{P}_n\left[R\left(\frac{1-\pi_{K_m}}{\pi_{K_m}}\right)(Y-b_0)^2\right] \\
&\overset{n\to\infty}{\to} \mathbb{E}\left[\frac{(1-\pi_{K_m})\pi_{K_{d_R}} v_{K_{d_Y}}}{\pi_{K_m}}\right] = O(1).
\end{aligned}$$

Therefore under the current framework, the within-sample prediction risk difference between the two nested estimators with only one degree of freedom difference for Joffe's estimators is converging to a constant and the selection procedure is asymptotically a deterministic selection process. To complete our calculation, we also derive the asymptotics for the re-standardized within-sample prediction risk

difference between $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *and* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell}$*:*

$$
\begin{aligned}
&\frac{1}{n}\xi^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
:= &\mathbb{P}_n \left[\begin{array}{l} \frac{R}{\pi_{K_m}} \left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} - \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -\frac{R}{\pi_{K_\ell}} \left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \end{array} \right] \\
= &\mathbb{P}_n \left[\frac{R}{\pi_{K_m}}\right] \left(\hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} + \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\
&- \mathbb{P}_n \left[\frac{R}{\pi_{K_\ell}}\right] \left(\hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \\
&- 2\mathbb{P}_n \left[\frac{R}{\pi_{K_m}} (Y - b_0)\right] \left(\hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} + \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right) \\
&+ 2\mathbb{P}_n \left[\frac{R}{\pi_{K_\ell}} (Y - b_0)\right] \left(\hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) + \mathbb{P}_n \left[\left(\frac{R}{\pi_{K_m}} - \frac{R}{\pi_{K_\ell}}\right) (Y - b_0)^2\right] \\
\xrightarrow{\mathcal{P}} &\mathbb{E}\left[\left(\frac{\pi_{K_{d_R}}}{\pi_{K_m}} - \frac{\pi_{K_{d_R}}}{\pi_{K_\ell}}\right) v_{K_{d_Y}}\right].
\end{aligned}
$$

The issue here is more severe because under homoscedasticity, the above asymptotic limit is zero and based on the formulation of the within-sample prediction risk comparison, one cannot distinguish between $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *and* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell}$ *asymptotically.*

Remark 2.3.4. *One might wonder what if using other formulations instead of following the CTMLE's recommendation which was originally designed for CTMLE estimators and using* $\frac{R}{\pi_{K_m}}$*, given that the problem of the above calculation arises from that one cannot cancel out the term with* $(Y-b_0)^2$*? Another natural potential statistics for evaluating the Joffe's estimator is to consider the "quasi"-within-sample prediction risk difference between* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ *and* $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell}$*, denoted as* $\xi^{quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$*:*

$$
\xi^{quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \tag{2.24}
$$

$$
\begin{aligned}
&:= n\mathbb{P}_n \left[\begin{array}{l} R\left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} - \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -R\left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \end{array} \right] \\
&= \mathbb{P}_n\left[R\right]\left(\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 - 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\mathbb{P}_n\left[R\left(Y-b_0\right)\right] \\
&\quad + 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\mathbb{P}_n\left[R\right]\hat{\varepsilon}_{K_i} + \cdots + \mathbb{P}_n\left[R\right]\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&= \mathbb{P}_n[R]\cdot\left\{\left(\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 - 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_0}}(Y-b_0)\right]\frac{\pi_{K_0}}{\mathbb{P}_n[R]}\right\} \\
&\quad + \mathbb{P}_n[R]\cdot 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\hat{\varepsilon}_{K_i} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&= \mathbb{P}_n[R]\cdot\left\{\left(\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 - 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\hat{\varepsilon}_{K_0}\frac{\pi_{K_0}}{\mathbb{P}_n[R]}\right\} \\
&\quad + \mathbb{P}_n[R]\cdot 2\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\hat{\varepsilon}_{K_i} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&\xrightarrow{d} \pi_{K_0}\cdot\left\{ \begin{array}{l} \left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} - p_{K_\ell}\right) \\ \times\left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} + v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} + p_{K_\ell} - 2v^{1/2}_{K_0,K_0}Z_{K_0} - 2p_{K_0}\right) \end{array} \right\} \\
&= \pi_{K_0}\cdot\left\{ \begin{array}{l} \left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m}\right)^2 - \left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} + p_{K_\ell}\right)^2 \\ -2\left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} - p_{K_\ell}\right)\left(v^{1/2}_{K_0,K_0}Z_{K_0} + p_{K_0}\right) \end{array} \right\} \qquad (2.25) \\
&\quad := \tilde{\zeta}^{quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \qquad (2.26)
\end{aligned}
$$

which is not necessarily less than zero between the two nested estimators with only one degree of freedom difference, hence is not a natural choice for within-sample prediction risk. For the sake of completeness, we also derive the asymptotic result for the "quasi"-within-sample prediction risk difference between

$\hat{\Psi}_{K_m}$ *and* $\mathbb{P}_n[b_0]$*:*

$$
\begin{aligned}
\xi^{quasi}_{K_m:} &:= n\mathbb{P}_n\left[R\left(Y-b_0-\hat{\varepsilon}_{K_m}\right)^2 - R(Y-b_0)^2\right] \\
&= \mathbb{P}_n[R]\, n\hat{\varepsilon}^2_{K_m} - 2\sqrt{n}\hat{\varepsilon}_{K_m}\sqrt{n}\mathbb{P}_n[R(Y-b_0)] \\
&= \mathbb{P}_n[R]\left(n\hat{\varepsilon}^2_{K_m} - 2\sqrt{n}\hat{\varepsilon}_{K_m}\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_0}}(Y-b_0)\right]\frac{\pi_{K_0}}{\mathbb{P}_n[R]}\right) \\
&= \mathbb{P}_n[R]\left(n\hat{\varepsilon}^2_{K_m} - 2\sqrt{n}\hat{\varepsilon}_{K_m}\sqrt{n}\hat{\varepsilon}_{K_0}\frac{\pi_{K_0}}{\mathbb{P}_n[R]}\right) \qquad (2.27) \\
&\xrightarrow{d} \pi_{K_0}\cdot\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-2v^{1/2}_{K_0,K_0}Z_{K_0}-2p_{K_0}-q\right)\right\} \\
&= \pi_{K_0}\cdot\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}\right)^2 - q^2 - 2\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(v^{1/2}_{K_0,K_0}Z_{K_0}+p_{K_0}\right)\right\} \\
&:= \tilde{\xi}^{quasi}_{K_m:}\,.
\end{aligned}
$$

Again this is not necessarily less than zero.

Large sample distributions for the conditional population prediction risks

In the last step of CTMLE, the final estimator will be selected based on the prediction risks from the candidate estimators filtered by within-sample prediction risk comparison, computed from M-fold cross-validation in practice. As mentioned earlier, an idealization of computation based on cross-validation is the conditional population prediction risk, where we can evaluate the prediction risk of the estimator using infinite amount of validation data or equivalently knowing the true DGP. Similar to within-sample prediction risk, below (Proposition 2.3.5) we derive the asymptotic distribution of the standardized conditional population prediction risk difference between the two nested estimators with only one degree of freedom difference, such as $\eta_{K_i,K_j:K_i}$ ($\hat{\Psi}_{K_i,K_j}$ vs. $\hat{\Psi}_{K_i}$) and $\eta_{K_i,K_j,K_k:K_i,K_j}$ ($\hat{\Psi}_{K_i,K_j,K_k}$ vs. $\hat{\Psi}_{K_i,K_j}$) (as defined in Definition 2.2.4).

Proposition 2.3.5. *As $n\to\infty$, under Definition 2.1.1, the standardized conditional population prediction*

risk difference between the two nested estimators with only one degree of freedom difference are of the following form:

$$\begin{aligned}
\eta_{K_m:} \xrightarrow{d} \tilde{\eta}_{K_m:} &:= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{v_{K_m,K_m} Z_{K_m}^2 - (p_{K_m}+q)^2\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)^2 - q^2 - 2p_{K_m}^2 - 2p_{K_m} q - 2p_{K_m} v_{K_m,K_m}^{1/2} Z_{K_m}\right\} \\
\eta_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} &\xrightarrow{d} \tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{v_{K_m,K_m} Z_{K_m}^2 - \left(v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} + p_{K_\ell} - p_{K_m}\right)^2\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{\begin{array}{l} \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 \\ -2\left(v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} + p_{K_\ell}\right)\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right) \end{array}\right\}
\end{aligned} \tag{2.28}$$

the asymptotic mean and variance of which are equal to

$$\begin{aligned}
\mathbb{E}\left[\tilde{\eta}_{K_m:}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{v_{K_m,K_m} - (p_{K_m}+q)^2\right\} \\
\mathbb{E}\left[\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{v_{K_m,K_m} - v_{K_\ell,K_\ell} - (p_{K_m}-p_{K_\ell})^2\right\}
\end{aligned} \tag{2.29}$$

and

$$\begin{aligned}
var\left[\tilde{\eta}_{K_m:}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-2} \cdot \left\{2v_{K_m,K_m}^2\right\} \\
var\left[\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-2} \cdot \left\{2v_{K_m,K_m}^2 + 2v_{K_\ell,K_\ell}^2 + 4(p_{K_m}-p_{K_\ell})^2 v_{K_\ell,K_\ell} - 4v_{K_\ell,K_m}^2\right\}
\end{aligned} \tag{2.30}$$

respectively, for $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$.

Proof.

$$
\begin{aligned}
\eta_{K_m:} &:= n\mathbb{E}\left[R\left(Y-b_0-\frac{\hat{\epsilon}_{K_m}}{\pi_{K_m}}\right)^2-R(Y-b_0)^2\middle|\hat{\epsilon}_{K_m}\right] \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]\left(\sqrt{n}\hat{\epsilon}_{K_m}\right)^2-2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\sqrt{n}\hat{\epsilon}_{K_m} \\
&\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)^2-2(p_{K_m}+q)\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{\begin{array}{l} v_{K_m,K_m}Z^2_{K_m}+(p_{K_m}+q)^2+2(p_{K_m}+q)v^{1/2}_{K_m,K_m}Z_{K_m} \\ -2(p_{K_m}+q)v^{1/2}_{K_m,K_m}Z_{K_m}-2(p_{K_m}+q)^2 \end{array}\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{v_{K_m,K_m}Z^2_{K_m}-(p_{K_m}+q)^2\right\}
\end{aligned}
$$

where the line for the convergence in distribution follows from Proposition 2.2.8 and Slutsky's theorem.

Next we derive the asymptotic distribution for $\eta_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$:

$$
\begin{aligned}
&\eta_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= n\mathbb{E}\left[\begin{array}{l} R\left(Y-b_0-\dfrac{\hat{\epsilon}_{K_i}}{\pi_{K_i}}-\dfrac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}}-\cdots-\dfrac{\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}}{\pi_{K_\ell}}-\dfrac{\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}}{\pi_{K_m}}\right)^2 \\ -R\left(Y-b_0-\dfrac{\hat{\epsilon}_{K_i}}{\pi_{K_i}}-\dfrac{\hat{\epsilon}_{K_i;K_j}}{\pi_{K_j}}-\cdots-\dfrac{\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}}{\pi_{K_\ell}}\right)^2 \end{array}\middle|\hat{\epsilon}_{K_i},\cdots,\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell,K_m}\right] \\
&= n\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]\hat{\epsilon}^2_{K_i,K_j,\cdots,K_\ell;K_m}-2\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right] \\
&\quad+2\sqrt{n}\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]\hat{\epsilon}_{K_i}+\cdots+\mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]\hat{\epsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\times \\
&\quad\left\{\begin{array}{l} \left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right)^2-2(p_{K_m}+q)\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right) \\ +2\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_m}-p_{K_\ell}\right)\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}+q\right) \end{array}\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{v_{K_m,K_m}Z^2_{K_m}-\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_m}+p_{K_\ell}\right)^2\right\}.
\end{aligned}
$$

where the line for the convergence in distribution follows from Proposition 2.2.8, Slutsky's theorem and a telescoping summation argument.

The mean and variance of $\tilde{\eta}_{K_m:}$ and $\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$ are straightforward application of the mean, variance, and covariance of correlated standard χ^2 random variables with one degree of freedom. □

Remark 2.3.6. *The asymptotic distribution $\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$ essentially only depends on K_ℓ and K_m. Similar to the within-sample prediction risk difference, one can also formulate the conditional population prediction risk difference between any two CTMLE estimators by taking the summation over the form given in eq.* (2.28)*.*

When $K_m = K_{full}$, $\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m} \cdot \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right] = \tilde{\zeta}^*_{K_m} - \tilde{\zeta}^*_{K_\ell}$. In this case, the asymptotic conditional population prediction risk difference is equivalent to the asymptotic squared error difference between $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ and $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell}$. Therefore we also motivate ourselves another version of the prediction-risk based selection procedure by reweighting the prediction risks with the inverse harmonic mean of the propensity score $\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]$. Such weighting will make a difference when not comparing the two nested estimators with only one degree of freedom difference, e.g. $\hat{\Psi}_{K_i}$ and $\hat{\Psi}_{K_i,K_j,K_\ell}$, with $K_i \subseteq K_j \subseteq K_\ell$:

$$\begin{aligned}
&\tilde{\eta}_{K_i,K_j,K_\ell:K_i} = \tilde{\eta}_{K_i,K_j,K_\ell:K_i,K_j} + \tilde{\eta}_{K_i,K_j:K_i} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]^{-1} \cdot \left\{ v_{K_\ell,K_\ell} Z^2_{K_\ell} - \left(v^{1/2}_{K_j,K_j} Z_{K_j} + p_{K_j} - p_{K_\ell} \right)^2 \right\} \\
&\quad + \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right]^{-1} \cdot \left\{ v_{K_j,K_j} Z^2_{K_j} - \left(v^{1/2}_{K_i,K_i} Z_{K_i} + p_{K_i} - p_{K_j} \right)^2 \right\}.
\end{aligned}$$

If we take the reweighted population conditional prediction risk (denoted by superscript "w")

$$\begin{aligned}
&\tilde{\eta}^w_{K_i,K_j,K_\ell:K_i} = \tilde{\eta}^w_{K_i,K_j,K_\ell:K_i,K_j} + \tilde{\eta}^w_{K_i,K_j:K_i} \\
&= \left\{ v_{K_\ell,K_\ell} Z^2_{K_\ell} - \left(v^{1/2}_{K_j,K_j} Z_{K_j} + p_{K_j} - p_{K_\ell} \right)^2 \right\} + \left\{ v_{K_j,K_j} Z^2_{K_j} - \left(v^{1/2}_{K_i,K_i} Z_{K_i} + p_{K_i} - p_{K_j} \right)^2 \right\}.
\end{aligned}$$

Remark 2.3.7. *One might be interested in the difference between conditional population prediction risk difference and within-sample prediction risk difference. Simple calculation shows that*

$$\begin{aligned}
&\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} - \tilde{\xi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&= 2\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot v^{1/2}_{K_m,K_m} Z_{K_m} \cdot \left\{ v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell} \right\},
\end{aligned} \tag{2.31}$$

the sign of which is not fixed in general. Their difference is a product of two correlated Gaussian random variables.

Remark 2.3.8. *For the sake of completeness, we also derive the asymptotic distributions for the conditional population prediction risks for Joffe's estimators. Notice that we need to reweight the standardized conditional population prediction risks by $1/n$ as in Remark 2.3.3 for heteroscedastic conditional outcome variance scenario:*

$$\begin{aligned}
\frac{1}{n}\eta^{Joffe}_{K_m:} &:= \mathbb{E}\left[\frac{R}{\pi_{K_m}} (Y - b_0 - \hat{\varepsilon}_{K_m})^2 - R(Y-b_0)^2 \middle| \hat{\varepsilon}_{K_m} \right] \\
&= (\hat{\varepsilon}_{K_m})^2 - 2\mathbb{E}\left[\frac{R}{\pi_{K_m}} (Y - b_0) \right] \hat{\varepsilon}_{K_m} + \mathbb{E}\left[\frac{1-\pi_{K_m}}{\pi_{K_m}} R (Y-b_0)^2 \right] \\
&\overset{n\to\infty}{\to} \mathbb{E}\left[\frac{(1-\pi_{K_m})\pi_{K_{d_R}} v_{K_{d_Y}}}{\pi_{K_m}} \right] = O(1).
\end{aligned}$$

Similarly

$$\begin{aligned}
&\frac{1}{n}\eta^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= \mathbb{E}\left[\begin{array}{l} \dfrac{R}{\pi_{K_m}} \left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} - \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m} \right)^2 \\ -\dfrac{R}{\pi_{K_\ell}} \left(Y - b_0 - \hat{\varepsilon}_{K_i} - \hat{\varepsilon}_{K_i;K_j} - \cdots - \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} \right)^2 \end{array} \middle| \hat{\varepsilon}_{K_i}, \cdots, \hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell,K_m} \right] \\
&= \mathbb{E}\left[\left(\frac{R}{\pi_{K_m}} - \frac{R}{\pi_{K_\ell}} \right) (Y-b_0)^2 \right] - 2\mathbb{E}\left[\left(\frac{R}{\pi_{K_m}} - \frac{R}{\pi_{K_\ell}} \right) (Y-b_0) \right] \left(\hat{\varepsilon}_{K_i} + \hat{\varepsilon}_{K_i;K_j} + \cdots + \hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell} \right)
\end{aligned}$$

$$
\begin{aligned}
&- 2\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m} - \left(\hat{\varepsilon}_{K_i}+\hat{\varepsilon}_{K_i;K_j}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \\
&+ \left(\hat{\varepsilon}_{K_i}+\hat{\varepsilon}_{K_i;K_j}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}+\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\
\overset{n\to\infty}{\to}\ &\mathbb{E}\left[\left(\frac{\pi_{K_{d_R}}}{\pi_{K_m}}-\frac{\pi_{K_{d_R}}}{\pi_{K_\ell}}\right)v_{K_{d_Y}}\right]
\end{aligned}
$$

However, for homoscedastic conditional variance scenario, one can use the original sample-size scaling factor. To see this, let's first consider the difference conditional population prediction risk difference between $\hat{\Psi}^{Joffe}_{K_m}$ *and* $\hat{\Psi}^{Joffe}_{K_\ell}$*:*

$$
\begin{aligned}
\eta^{Joffe}_{K_m:K_\ell} &:= \mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0-\hat{\varepsilon}_{K_m})^2 - \frac{R}{\pi_{K_\ell}}(Y-b_0-\hat{\varepsilon}_{K_\ell})^2\middle|\hat{\varepsilon}_{K_m},\hat{\varepsilon}_{K_\ell}\right] \\
&= \left(\sqrt{n}\hat{\varepsilon}_{K_m}\right)^2 - \left(\sqrt{n}\hat{\varepsilon}_{K_\ell}\right)^2 + n\mathbb{E}\left[\left(\frac{1}{\pi_{K_m}}-\frac{1}{\pi_{K_\ell}}\right)R(Y-b_0)^2\right] \\
&\quad - 2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_m} + 2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_\ell} \\
&\overset{hom}{=} \left(\sqrt{n}\hat{\varepsilon}_{K_m}\right)^2 - \left(\sqrt{n}\hat{\varepsilon}_{K_\ell}\right)^2 - 2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_m} + 2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_\ell} \\
&\overset{d}{\to} \left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}-p_{K_m}-q\right) \\
&\quad - \left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}+q\right)\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}-q\right) \\
&\equiv v_{K_m,K_m}Z^2_{K_m} - (p_{K_m}+q)^2 - v_{K_\ell,K_\ell}Z^2_{K_\ell} + (p_{K_\ell}+q)^2 \\
&\equiv \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]\tilde{\eta}_{K_m:} - \mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]\tilde{\eta}_{K_\ell:} \textit{ (due to Proposition 2.3.5)} \\
&\equiv \tilde{\eta}^{w}_{K_m:} - \tilde{\eta}^{w}_{K_\ell:}.
\end{aligned}
$$

Similarly

$$
\eta^{Joffe}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}
$$

$$
\begin{aligned}
&:= n\mathbb{E}\left[\begin{array}{l} \frac{R}{\pi_{K_m}}\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}-\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -\frac{R}{\pi_{K_\ell}}\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \end{array}\middle|\, \hat{\varepsilon}_{K_i},\cdots,\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell,K_m}\right] \\
&= n\mathbb{E}\left[\left(\frac{R}{\pi_{K_m}}-\frac{R}{\pi_{K_\ell}}\right)(Y-b_0)^2\right]-2n\mathbb{E}\left[\left(\frac{R}{\pi_{K_m}}-\frac{R}{\pi_{K_\ell}}\right)(Y-b_0)\right]\left(\hat{\varepsilon}_{K_i}+\hat{\varepsilon}_{K_i;K_j}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&\quad -2n\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}-n\left(\hat{\varepsilon}_{K_i}+\hat{\varepsilon}_{K_i;K_j}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \\
&\quad +n\left(\hat{\varepsilon}_{K_i}+\hat{\varepsilon}_{K_i;K_j}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}+\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\
&\overset{hom}{=} -\left(\sqrt{n}\hat{\varepsilon}_{K_\ell}\right)^2+\left(\sqrt{n}\hat{\varepsilon}_{K_m}\right)^2-2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_m}+2\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_\ell}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_\ell} \\
&\xrightarrow{d}\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}-p_{K_m}-q\right) \\
&\quad -\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}+q\right)\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}-q\right) \\
&\equiv v_{K_m,K_m}Z^2_{K_m}-(p_{K_m}+q)^2-v_{K_\ell,K_\ell}Z^2_{K_\ell}+(p_{K_\ell}+q)^2 \\
&\equiv \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]\tilde{\eta}_{K_m:}-\mathbb{E}\left[\frac{1}{\pi_{K_\ell}}\right]\tilde{\eta}_{K_\ell:}\ \textit{(due to Proposition 2.3.5)} \\
&\equiv \tilde{\eta}^w_{K_m:}-\tilde{\eta}^w_{K_\ell:}.
\end{aligned}
$$

Therefore under homoscedastic conditional outcome variance assumption, regardless of the form of the Joffe's estimator, the conditional population prediction risk difference has the same asymptotic distribution as the conditional population prediction risk difference between two CTMLE estimators adjusting for only one DR covariate, then reweighted by the inverse harmonic mean of the propensity scoreof the two corresponding DR covariates.

Remark 2.3.9. *Similar to Remark 2.3.4, we also try how things work for "quasi"-conditional population*

prediction risksfor Joffe's estimators:

$$\begin{aligned}
\eta^{quasi}_{K_m:} &:= n\mathbb{E}\left[R\left(Y-b_0-\hat{\varepsilon}_{K_m}\right)^2 - R(Y-b_0)^2 \middle| \hat{\varepsilon}_{K_m}\right] \\
&= \pi_{K_0}\left(\sqrt{n}\hat{\varepsilon}_{K_m}\right)^2 - 2\pi_{K_0}\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_0}}(Y-b_0)\right]\sqrt{n}\hat{\varepsilon}_{K_m} \\
&\xrightarrow{d} \pi_{K_0}\cdot\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)^2 - 2\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(p_{K_0}+q\right)\right\} \qquad (2.32) \\
&= \pi_{K_0}\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-2p_{K_0}-q\right) \\
&= \pi_{K_0}\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}\right)^2 - q^2 - 2p_{K_0}\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+q\right)\right\} := \tilde{\eta}^{quasi}_{K_m:}.
\end{aligned}$$

And similarly,

$$\begin{aligned}
&\eta^{quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= n\mathbb{E}\left[\begin{array}{l} R\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}-\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -R\left(Y-b_0-\hat{\varepsilon}_{K_i}-\hat{\varepsilon}_{K_i;K_j}-\cdots-\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right)^2 \end{array} \middle| \hat{\varepsilon}_{K_i},\cdots,\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell,K_m}\right] \\
&= \pi_{K_0}n\hat{\varepsilon}^2_{K_i,K_j,\cdots,K_\ell;K_m} - 2\pi_{K_0}\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\mathbb{E}\left[\frac{R}{\pi_{K_0}}(Y-b_0)\right] \\
&\quad + 2\pi_{K_0}\sqrt{n}\hat{\varepsilon}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{n}\left(\hat{\varepsilon}_{K_i}+\cdots+\hat{\varepsilon}_{K_i,K_j,\cdots;K_\ell}\right) \\
&\xrightarrow{d} \pi_{K_0}\cdot\left\{\begin{array}{l}\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)^2 \\ +2\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}-p_{K_0}\right)\end{array}\right\} \\
&= \pi_{K_0}\cdot\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}+v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}-2p_{K_0}\right) \\
&= \pi_{K_0}\cdot\left\{\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}\right)^2 - \left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}+p_{K_\ell}\right)^2 - 2p_{K_0}\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)\right\} \\
&:= \tilde{\eta}^{quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}. \qquad (2.33)
\end{aligned}$$

2.4 Large sample distribution for the M-CV prediction risks and ∞-CV prediction risks

In this section, we study the asymptotic distributions of the M-CV prediction risks and its limit ∞-CV prediction risks as $M \to \infty$. The large-sample regime that we are interested in is when $n/M \to \infty$ and $n(M-1)/M \to \infty$, i.e. both the training sample and validation sample sizes after sample splitting still tend to infinity as $n \to \infty$. A typical example is $M = log(n)$ when $M \to \infty$ while $n \to \infty$ but with the above two conditions hold. Such asymptotic limit can dramatically simplify the expression of the asymptotic distribution for the ∞-CV prediction risks. Moreover, we will see that the mean of the ∞-CV prediction risk is equal to the mean of its corresponding conditional population prediction risk, whereas this equality does not hold for finite M. In other words, the "finite fold" cross-validation does not render an unbiased estimator for the true population marginal prediction risk.

To derive the asymptotic distributions of the prediction risk differences between the two nested estimators with only one degree of freedom difference as described in the previous section (Section 2.3), we first recall the definition of the M-CV prediction risk (Definition 2.2.5). In Definition 2.2.5, one needs to split the sample into M different pieces indexed by $t = 1, \cdots, M$. Then for each $t = 1, \cdots, M$, one first estimate the DR coefficients from all the samples but those falling into the t^{th} piece, e.g. $\hat{\epsilon}_{K_i}^{\setminus t} = \mathbb{P}_{n(M-1)/M}^{\setminus t}\left[\frac{R}{\pi_{K_i}}\right]^{-1} \cdot \mathbb{P}_{n(M-1)/M}^{\setminus t}\left[\frac{R(Y-b_0)}{\pi_{K_i}}\right]$ and similar for the more general DR coefficients such as $\hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m}^{\setminus t}$; then the prediction risk will be evaluated using the samples falling into the t^{th} piece. Then we average over all the $t = 1, \cdots, M$, hence getting the formula in eq. (2.10) standardized by sample size. Notice that for $\hat{\epsilon}_\cdot$, the appropriate scaling factor by sample size is $\sqrt{\frac{n(M-1)}{M}}$. Similarly, we define the DR coefficients estimated from the validation sample indexed by t as $\hat{\epsilon}_\cdot^t$ and its appropriate scaling factor is $\sqrt{\frac{n}{M}}$. Then following Proposition 2.2.8, we directly have

Proposition 2.4.1. *As $n/M \to \infty$ and $n(M-1)/M \to \infty$, under Definition 2.1.1, the CTMLE DR*

coefficients using the samples falling into the t^{th} piece and the samples outside of the t^{th} piece ($t = 1, \cdots, M$) have the following asymptotic Gaussian distributions

$$\begin{aligned}
\sqrt{\frac{n}{M}}\hat{\epsilon}^{t}_{K_i} &\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]^{-1}\left\{v^{1/2}_{K_i,K_i}Z^{t}_{K_i} + \sqrt{\frac{1}{M}}\left(p_{K_i}+q\right)\right\} \\
&\vdots \\
\sqrt{\frac{n}{M}}\hat{\epsilon}^{t}_{K_i,K_j,\cdots,K_\ell;K_m} &\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\left\{v^{1/2}_{K_m,K_m}Z^{t}_{K_m} + \sqrt{\frac{1}{M}}p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^{t}_{K_\ell} - \sqrt{\frac{1}{M}}p_{K_\ell}\right\},
\end{aligned} \tag{2.34}$$

and

$$\begin{aligned}
\sqrt{\frac{n(M-1)}{M}}\hat{\epsilon}^{\backslash t}_{K_i} &\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right]^{-1}\left\{v^{1/2}_{K_i,K_i}Z^{\backslash t}_{K_i} + \sqrt{\frac{M-1}{M}}\left(p_{K_i}+q\right)\right\} \\
&\vdots \\
\sqrt{\frac{n(M-1)}{M}}\hat{\epsilon}^{\backslash t}_{K_i,K_j,\cdots,K_\ell;K_m} &\xrightarrow{d} \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\left\{v^{1/2}_{K_m,K_m}Z^{\backslash t}_{K_m} + \sqrt{\frac{M-1}{M}}p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^{\backslash t}_{K_\ell} - \sqrt{\frac{M-1}{M}}p_{K_\ell}\right\}.
\end{aligned} \tag{2.35}$$

where $Z^{t}_{K_i}, Z^{t}_{K_j}, \cdots, Z^{t}_{K_\ell}, Z^{t}_{K_m}$ are multivariate Gaussian random variables with mean 0, unit variance, and covariance between $Z^{t}_{K_r}$ and $Z^{t}_{K_s}$ is $\frac{v_{K_r,K_s}}{v^{1/2}_{K_r,K_r}v^{1/2}_{K_s,K_s}}$ for any set of potential confounders K_r, K_s; similarly, $Z^{\backslash t}_{K_i}, Z^{\backslash t}_{K_j}, \cdots, Z^{\backslash t}_{K_\ell}, Z^{\backslash t}_{K_m}$ are multivariate Gaussian random variables with mean 0, unit variance, and covariance between $Z^{\backslash t}_{K_r}$ and $Z^{\backslash t}_{K_s}$ is also $\frac{v_{K_r,K_s}}{v^{1/2}_{K_r,K_r}v^{1/2}_{K_s,K_s}}$ for any set of potential confounders K_r, K_s. In addition, for any arbitrary subscripts, $Z^{t}_{\cdot}$ and $Z^{\backslash t}_{\cdot}$ are independent for any $t = 1, \cdots, M$; $Z^{t}_{\cdot}$ and $Z^{t'}_{\cdot}$ are also independent for any $t \neq t'$, $t, t' = 1, \cdots, M$. For the same subscript, $\sum_{t=1}^{M} Z^{t}_{\cdot} = \sqrt{M}Z_{\cdot}$ and $Z^{\backslash t}_{\cdot} = \sqrt{\frac{M}{M-1}}\left(Z_{\cdot} - \sqrt{\frac{1}{M}}Z^{t}_{\cdot}\right)$.

Then using Proposition 2.4.1, in Section A.2, we derive the following asymptotic distributions for the *M*-CV prediction risks difference between the two nested estimators with only one degree of freedom

difference:

Proposition 2.4.2. *As $n \to \infty$, under Definition 2.1.1, the standardized M-CV prediction risk difference between the two nested estimators with only one degree of freedom difference are of the following form, for $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m$:*

$$
\begin{aligned}
&\eta^{\dagger M}_{K_m:} \xrightarrow{d} \tilde{\eta}^{\dagger M}_{K_m:} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ \frac{2M-1}{(M-1)^2} \sum_{t=1}^{M} v_{K_m,K_m} Z^{t\,2}_{K_m} - \frac{M^2}{(M-1)^2} v_{K_m,K_m} Z^2_{K_m} - 2v^{1/2}_{K_m,K_m}(p_{K_m}+q) Z_{K_m} - (p_{K_m}+q)^2 \right\}, \\
&\eta^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&\xrightarrow{d} \tilde{\eta}^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \times \\
&\left\{ \begin{array}{l}
\frac{1}{(M-1)^2} \sum_{t=1}^{M} \left[(2M-1) v^{1/2}_{K_m,K_m} Z^t_{K_m} + v^{1/2}_{K_\ell,K_\ell} Z^t_{K_\ell}\right] \left[v^{1/2}_{K_m,K_m} Z^t_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z^t_{K_\ell}\right] \\
-\frac{M}{(M-1)^2} \left[M v^{1/2}_{K_m,K_m} Z_{K_m} - (M-2) v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell}\right] \left[v^{1/2}_{K_m,K_m} Z_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell}\right] \\
-2v^{1/2}_{K_m,K_m}(p_{K_m}-p_{K_\ell}) Z_{K_m} + 2v^{1/2}_{K_\ell,K_\ell}(p_{K_m}-p_{K_\ell}) Z_{K_\ell} - (p_{K_m}-p_{K_\ell})^2
\end{array} \right\}
\end{aligned}
\tag{2.36}
$$

the expectations of which are

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\eta}^{\dagger M}_{K_m:}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ \frac{M}{M-1} v_{K_m,K_m} - (p_{K_m}+q)^2 \right\}, \\
\mathbb{E}\left[\tilde{\eta}^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ \frac{M}{M-1} \left(v_{K_m,K_m} - v_{K_\ell,K_\ell}\right) - (p_{K_m}-p_{K_\ell})^2 \right\}.
\end{aligned}
\tag{2.37}
$$

Next as mentioned in the beginning of this section, we study the limiting behavior when $M \to \infty$ at a slower rate than n. A typical example of such scenario is to set $M = log(n)$. We have the following corollary.

Corollary 2.4.3. *If we also let $M \to \infty$ but at a slower rate than $n \to \infty$ so we still have $n/M \to \infty$ and $n(M-1)/M \to \infty$, for $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$:*

$$\begin{aligned}
&\tilde{\eta}^{\dagger\, M}_{K_m:} \overset{M\to\infty}{\to} \tilde{\eta}^{\dagger}_{K_m:} \\
&:= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ -\left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} + q \right)^2 + 2v_{K_m,K_m} \right\} \\
&= \tilde{\xi}_{K_m:} + 2\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} v_{K_m,K_m}, \\
&\tilde{\eta}^{\dagger\, M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \overset{M\to\infty}{\to} \tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&:= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ -\left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell} \right)^2 + 2v_{K_m,K_m} - 2v_{K_m,K_\ell} \right\} \\
&= \tilde{\xi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} + \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot (2v_{K_m,K_m} - 2v_{K_m,K_\ell})
\end{aligned} \tag{2.38}$$

the expectations and variances of which are

$$\begin{aligned}
\mathbb{E}\left[\tilde{\eta}^{\dagger}_{K_m:}\right] \equiv \mathbb{E}\left[\tilde{\xi}_{K_m:}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ v_{K_m,K_m} - (p_{K_m} + q)^2 \right\} \\
\mathbb{E}\left[\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] &\equiv \mathbb{E}\left[\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{ v_{K_m,K_m} - v_{K_\ell,K_\ell} - (p_{K_m} - p_{K_\ell})^2 \right\}
\end{aligned} \tag{2.39}$$

and

$$\begin{aligned}
var\left[\tilde{\eta}^{\dagger}_{K_m:}\right] \equiv var\left[\tilde{\xi}_{K_m:}\right] &= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-2} \cdot \left\{ 2v^2_{K_m,K_m} + 4v_{K_m,K_m}(p_{K_m} + q)^2 \right\} \\
var\left[\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] &\equiv var\left[\tilde{\xi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-2} \cdot \left\{ \begin{array}{l} 2(v_{K_\ell,K_\ell} + v_{K_m,K_m} - 2v_{K_\ell,K_m})^2 \\ +4(p_{K_m} - p_{K_\ell})^2(v_{K_\ell,K_\ell} + v_{K_m,K_m} - 2v_{K_\ell,K_m}) \end{array} \right\}.
\end{aligned} \tag{2.40}$$

Proof. We only prove eq. (2.39) in detail here. To see this

$$
\begin{aligned}
&\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \times \\
&\left\{ \begin{array}{l}
\frac{1}{(M-1)^2}\sum_{t=1}^{M}\left[(2M-1)v^{1/2}_{K_m,K_m}Z^t_{K_m}+v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}\right]\left[v^{1/2}_{K_m,K_m}Z^t_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}\right] \\
-\frac{M}{(M-1)^2}\left[Mv^{1/2}_{K_m,K_m}Z_{K_m}-(M-2)v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right]\left[v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right] \\
-2v^{1/2}_{K_m,K_m}(p_{K_m}-p_{K_\ell})Z_{K_m}+2v^{1/2}_{K_\ell,K_\ell}(p_{K_m}-p_{K_\ell})Z_{K_\ell}-(p_{K_m}-p_{K_\ell})^2
\end{array} \right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \times \\
&\left\{ \begin{array}{l}
\frac{2M-1}{(M-1)^2}\sum_{t=1}^{M}v_{K_m,K_m}Z^{t\,2}_{K_m}-\frac{2M-1}{(M-1)^2}\sum_{t=1}^{M}v^{1/2}_{K_m,K_m}Z^t_{K_m}v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}-\frac{1}{(M-1)^2}\sum_{t=1}^{M}v_{K_\ell,K_\ell}Z^{t\,2}_{K_\ell} \\
+\frac{1}{(M-1)^2}\sum_{t=1}^{M}v^{1/2}_{K_m,K_m}Z^t_{K_m}v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}-\frac{M^2}{(M-1)^2}v_{K_m,K_m}Z^2_{K_m}+\frac{2M(M-1)}{(M-1)^2}v^{1/2}_{K_m,K_m}Z_{K_m}v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} \\
-\frac{M(M-2)}{(M-1)^2}v_{K_\ell,K_\ell}Z^2_{K_\ell}-2v^{1/2}_{K_m,K_m}(p_{K_m}-p_{K_\ell})Z_{K_m}+2v^{1/2}_{K_\ell,K_\ell}(p_{K_m}-p_{K_\ell})Z_{K_\ell}-(p_{K_m}-p_{K_\ell})^2
\end{array} \right\} \\
&\xrightarrow[M\to\infty]{P} \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\cdot\left\{2v_{K_m,K_m}-2v_{K_\ell,K_m}-\left(v^{1/2}_{K_m,K_m}Z_{K_m}+p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-p_{K_\ell}\right)^2\right\}
\end{aligned}
$$

where $\frac{1}{M}\sum_{t=1}^{M}Z^{t\,2}_{\cdot} \xrightarrow[M\to\infty]{P} 1$ and $\frac{1}{M}\sum_{t=1}^{M}Z^t_{K_m}Z^t_{K_\ell} \xrightarrow[M\to\infty]{P} \frac{v_{K_\ell,K_m}}{v^{1/2}_{K_\ell,K_\ell}v^{1/2}_{K_m,K_m}}$. □

Remark 2.4.4. *Here we also establish an asymptotic algebraic relation between the ∞-CV prediction risk and the within-sample prediction risk. First they have the same variance because they are only different up to a constant. Second, from eq.* (2.38)*, under homoscedasticity, since $v_{K_m,K_m} \geq v_{K_\ell,K_\ell}$, $\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$ is always larger than or equal to $\tilde{\xi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$.*

Remark 2.4.5. *By looking at the expression in eq.* (2.38)*, when comparing $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ vs. $\hat{\Psi}_{K_i,K_j,\cdots,K_k,K_\ell}$, the criterion of selecting the former rather than the latter (or in other words, rejecting the null hypothesis*

$H_0 : \hat{\epsilon}_{K_i,K_j,\cdots,K_\ell;K_m} = 0$*) based on* ∞*-CV prediction risk is*

$$
\begin{aligned}
& - \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + 2v_{K_m,K_m} - 2v_{K_m,K_\ell} < 0 \\
\Leftrightarrow & \frac{\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2}{v_{K_m,K_m} - v_{K_m,K_\ell}} > 2 \\
\overset{H_0}{\Leftrightarrow} & \frac{\left(v_{K_m,K_m}^{1/2} Z_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell}\right)^2}{v_{K_m,K_m} - v_{K_m,K_\ell}} > 2 \\
\overset{H_0}{\Leftrightarrow} & \frac{\left(v_{K_m,K_m}^{1/2} Z_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell}\right)^2}{v_{K_m,K_m} + v_{K_\ell,K_\ell} - 2v_{K_m,K_\ell}} > 2(v_{K_m,K_m} - v_{K_\ell,K_m}).
\end{aligned}
$$

The RHS of the above inequality is essentially a χ^2 *random variable with one degree of freedom. Therefore there exists an equivalence between the* ∞*-CV prediction risk comparison and the hypothesis testing when comparing between the two nested estimators with only one degree of freedom difference with some cutoff depending on the tested hypothesis. However, when not comparing the two nested estimators with only one degree of freedom difference, the* ∞*-CV prediction risks are added up and there is no way to rewrite the* ∞*-CV prediction risk comparison into a form of hypothesis testing based on* χ_1^2 *statistic. This phenomenon also motivates us to study the performance of a related procedure developed by Oleg Lepski, Vladimir Spokoiny and their colleagues*[45,46,89]*. The statistic used in Lepski's method can be rewritten as the form of hypothesis testing with* χ_1^2 *statistic when comparing any two candidate estimators. One simplification of Lepski's method is that the CTMLE estimator and Joffe's estimator, just as the case for squared error, share the same asymptotic distribution for the Lepski's statistic. For more details, see Section 2.6.*

Remark 2.4.6. *Given the similarity between the quasi-conditional population prediction risks for Joffe's estimator and the conditional population prediction risks for CTMLE estimators, we also put effort on deriving the asymptotic distribution for the quasi-M-CV prediction risks and quasi-*∞*-CV prediction*

risks for the Joffe's estimators in Section A.2. Interestingly, as in the case of CTMLE estimators, we found that the quasi-∞-CV prediction risks for Joffe's estimators can be decomposed into one piece corresponding to the quasi-within-sample prediction risks for Joffe's estimators derived in Remark 2.3.4 and another piece which is a constant. The asymptotic distributions $\tilde{\eta}^{\dagger\ quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$ and $\tilde{\eta}^{\dagger\ quasi}_{K_m:}$ for $\eta^{\dagger\ quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}$ and $\eta^{\dagger\ quasi}_{K_m:}$ are shown below:

$$\tilde{\eta}^{\dagger\ quasi}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} = \pi_{K_0} \cdot \left\{ \begin{array}{l} \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2 - \left(v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} + p_{K_\ell}\right)^2 + 2v_{K_0,K_m} - 2v_{K_0,K_\ell} \\ -2\left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell}\right)\left(v^{1/2}_{K_0,K_0} Z_{K_0} + p_{K_0}\right) \end{array} \right\} \tag{2.41}$$

And

$$\tilde{\eta}^{\dagger\ quasi}_{K_m:} = \pi_{K_0} \cdot \left\{ \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2 - q^2 - 2\left(v^{1/2}_{K_0,K_0} Z_{K_0} + p_{K_0}\right)\left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} + q\right) + 2v_{K_0,K_m} \right\}. \tag{2.42}$$

Penalized ∞-CV prediction risk

In van der Laan & Rose[102], the authors proposed to add additional terms to the M-CV prediction risk (called cvRSS in van der Laan & Rose[102]) to penalize the prediction risk by terms corresponding to the "mean squared error". In the book, the authors named these terms as "cvVar" and square of "cvBias", which we define below for readers' convenience: The cvBias and cvVar of estimator $\hat{\Psi}_{K_m}$ are

$$\text{cvBias}^2_{K_m} := n\left(\frac{1}{M}\sum_{t=1}^{M} \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}^{\setminus t}_{K_m}}{\pi_{K_m}}\right] - \mathbb{P}_n\left[b_0 + \frac{\hat{\epsilon}_{K_m}}{\pi_{K_m}}\right]\right)^2 \tag{2.43}$$

and

$$\text{cvVar}_{K_m} := \frac{n}{M}\sum_{t=1}^{M}\mathbb{P}_{n/M}^{t}\left[\left\{\frac{R}{\pi_{K_m}}\left(Y-b_0-\frac{\hat{\epsilon}_{K_m}^{\setminus t}}{\pi_{K_m}}\right)+\frac{\hat{\epsilon}_{K_m}^{\setminus t}}{\pi_{K_m}}-\mathbb{P}_n\left[\frac{1}{\pi_{K_m}}\right]\hat{\epsilon}_{K_m}\right\}^2\right]. \tag{2.44}$$

We would like to comment on the cvBias first: it is not the bias of the estimator compared to the truth but rather the bias of cross-validation compared to the estimator using the whole sample. It is not hard to see that cvBias is asymptotically zero. This is true regardless of whether or not the outcome regression model is local alternative. To see this, we only need to consider the following comparison:

$$\begin{aligned}
&\mathbb{P}_n\left[\frac{1}{\pi_{K_m}}\right]\left(\frac{1}{M}\sum_{t=1}^{M}\sqrt{n}\hat{\epsilon}_{K_m}^{\setminus t}-\sqrt{n}\hat{\epsilon}_{K_m}\right)\\
&=\mathbb{P}_n\left[\frac{1}{\pi_{K_m}}\right]\left\{\sqrt{\frac{M}{M-1}}\frac{1}{M}\sum_{t=1}^{M}\sqrt{\frac{n(M-1)}{M}}\left(\hat{\epsilon}_{K_m}^{\setminus t}-\epsilon_{K_m}\right)-\sqrt{n}\left(\hat{\epsilon}_{K_m}-\epsilon_{K_m}\right)\right\}\\
&\xrightarrow{d}\sqrt{\frac{M}{M-1}}\frac{1}{M}\sum_{t=1}^{M}v_{K_m,K_m}^{1/2}Z_{K_m}^{\setminus t}-v_{K_m,K_m}^{1/2}Z_{K_m}\\
&=v_{K_m,K_m}^{1/2}\left\{\frac{1}{M}\sum_{t=1}^{M}\frac{M}{M-1}\left(Z_{K_m}-\sqrt{\frac{1}{M}}Z_{K_m}^{t}\right)-Z_{K_m}\right\}\\
&=v_{K_m,K_m}^{1/2}\left(\frac{MZ_{K_m}}{M-1}-\frac{Z_{K_m}}{M-1}-Z_{K_m}\right)=0.
\end{aligned}$$

For the cvVar difference, it can be shown to dominate the order of the M-CV prediction risk difference (and of course, the cvBias difference) if we do not rescale eq. (2.44) by n^{-1}.

In the next two sections (Section 2.5 and Section 2.6), we will discuss some other alternative approaches to constructing the final estimator from a family of candidate adjusting for propensity score models with different complexity levels. One alternative (Section 2.5) is the focused information criterion (FIC)[12], since FIC is essentially a consistent estimator of the A.M.S.E.. In some sense, FIC is "more targeted" towards selecting estimators with good M.S.E. compared to prediction-risk based procedures

such as CTMLE. However, as one can imagine, a consistent estimator of the mean squared error has its own specific asymptotic structure, and the difference between prediction risk and FIC is quite subtle in terms of the properties of the estimator obtained after model selection. In the most original version of FIC[12], one constructs a consistent estimator of the A.M.S.E. of this candidate estimator using the whole sample. Another alternative statistic similar to FIC was proposed in Vansteelandt et al.[107], also aiming at constructing a consistent estimator of the A.M.S.E. of the candidate estimators. However, instead of using the whole sample, Vansteelandt et al.[107] employed a sample-splitting method (dividing the whole sample into M pieces) and computes the squared error distance between the candidate estimator estimated from the training sample and the estimator controlling for every measured potential confounders estimated from the validation sample. Under the assumption of no unmeasured confounder, such statistic will also be a consistent estimator of the A.M.S.E. as $M \to \infty$ because the estimator controlling for all the measured potential confounders is a consistent estimator of Ψ_0. We will briefly discuss this approach in Section 2.5 and show its asymptotic equivalence with FIC. In terms of the selection procedure, there exists multiple strategies to build up the final estimator for the target Ψ_0 using the FIC-based statistics. The most straightforward strategy is to select the candidate estimator with the minimum FIC[12] or the sample-splitting statistics. Other strategies, including model averaging[34,13], are also possible but will be left to future investigation. One other strategy that we will discuss in Section 2.6 is inspired from the so-called Lepski's method[44,45,46,89]. The basic idea behind Lepski's method is to perform pair-wise comparison between the candidate estimators in the framework of hypothesis testing or optimal adaptive estimation.

2.5 Focused Information Criterion for CTMLE estimators

Large sample distribution of FIC

FIC was first proposed in Claeskens & Hjort[12]. Since FIC is essentially constructing a consistent estimator of A.M.S.E., we recall that for candidate estimators of Ψ_0 including DR covariates adjusting for at most the set of potential confounders K_m, the asymptotic standardized A.M.S.E. is the same (this is also true for candidate "Joffee's" estimators): $\mathbb{E}\left[\tilde{\zeta}^*_{K_m}\right] = v_{K_m,K_m} + p^2_{K_m}$. Therefore, when trying to construct a consistent estimator of the A.M.S.E., one does not have to distinguish among the estimators of the forms $\hat{\Psi}^{(\text{Joffe})}_{K_m}$ and $\hat{\Psi}^{(\text{Joffe})}_{K_i,K_j,\cdots,K_\ell,K_m}$ for $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{\text{full}}$. Following Claeskens & Hjort[12], the construction of the consistent estimator of A.M.S.E. is to estimate the asymptotic bias square $p^2_{K_m}$ consistently, where $p^2_{K_m} = n\left(\mathbb{E}\left[\left(\frac{R}{\pi_{K_m}} - 1\right)(Y - b_0)\right]\right)^2 = n\left(\mathbb{E}\left[\frac{R}{\pi_{K_m}}(Y - b_0)\right] - \mathbb{E}\left[Y - b_0\right]\right)^2$. The most straightforward estimator of $p^2_{K_m}$ is then to first replace the difference between two expectations by the difference between two empirical means $\left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi_{K_m}}(Y - b_0)\right] - \sqrt{n}\mathbb{P}_n\left[Y - b_0\right]\right)^2 = \left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_m}}\right]\cdot\hat{\epsilon}_{K_m} - \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_{\text{full}}}}\right]\cdot\hat{\epsilon}_{K_{\text{full}}}\right)^2$. However, since this expression converges in distribution to $\left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - v^{1/2}_{K_{\text{full}},K_{\text{full}}}Z_{K_{\text{full}}}\right)^2$, with mean $v_{K_m,K_m} + v_{K_{\text{full}},K_{\text{full}}} - 2v_{K_m,K_{\text{full}}} + p^2_{K_m}$, one should replace $p^2_{K_m}$ by

$$\left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_m}}\right]\cdot\hat{\epsilon}_{K_m} - \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_{\text{full}}}}\right]\cdot\hat{\epsilon}_{K_{\text{full}}}\right)^2 - v_{K_{\text{full}},K_{\text{full}}} - v_{K_m,K_m} + 2v_{K_m,K_{\text{full}}}.$$

Then the standardized FIC (by $\sqrt{n}$) for an estimator including the DR covariates adjusting for at most the set of potential confounders K_m is defined as

Definition 2.5.1.

$$FIC_{K_m} := -v_{K_{full},K_{full}} + 2v_{K_m,K_{full}} + \left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_m}}\right]\cdot\hat{\epsilon}_{K_m} - \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_{full}}}\right]\cdot\hat{\epsilon}_{K_{full}}\right)^2 \tag{2.45}$$

and the FIC difference between two estimators with standardized A.M.S.E. $\tilde{\zeta}^*_{K_m}$ and $\tilde{\zeta}^*_{K_\ell}$ respectively as

$$\begin{aligned} FIC_{K_m:K_\ell} := {} & 2v_{K_m,K_{full}} - 2v_{K_\ell,K_{full}} + \left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_m}}\right]\cdot\hat{\epsilon}_{K_m} - \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_{full}}}\right]\cdot\hat{\epsilon}_{K_{full}}\right)^2 \\ & - \left(\sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_\ell}}\right]\cdot\hat{\epsilon}_{K_\ell} - \sqrt{n}\mathbb{P}_n\left[\frac{R}{\pi^2_{K_{full}}}\right]\cdot\hat{\epsilon}_{K_{full}}\right)^2 \end{aligned} \tag{2.46}$$

Then replacing the asymptotic distribution of $\sqrt{n}\hat{\epsilon}_{K_m}$ and $\sqrt{n}\hat{\epsilon}_{K_{\text{full}}}$, one can easily show the following result.

Proposition 2.5.2. *As $n \to \infty$, for candidate estimators of Ψ_0 of the forms $\hat{\Psi}_{K_m}$ and $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ where $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$, the FIC as defined in eq.* (2.45) *converges to the following distribution:*

$$FIC_{K_m} \xrightarrow{d} \widetilde{FIC}_{K_m} := \left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - v^{1/2}_{K_{full},K_{full}}Z_{K_{full}}\right)^2 - v_{K_{full},K_{full}} + 2v_{K_m,K_{full}}. \tag{2.47}$$

Since we are also interested in the difference of FICs between two candidate estimators, let's denote the difference as $\widetilde{FIC}_{K_m:K_\ell}$ for $K_\ell \subseteq K_m$ as

$$\begin{aligned} FIC_{K_m:K_\ell} \xrightarrow{d} \widetilde{FIC}_{K_m:K_\ell} := {} & \left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m}\right)^2 - \left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} + p_{K_\ell}\right)^2 + 2v_{K_m,K_{full}} - 2v_{K_\ell,K_{full}} \\ & - 2v^{1/2}_{K_{full},K_{full}}Z_{K_{full}}\left(v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell} - p_{K_\ell}\right). \end{aligned} \tag{2.48}$$

We also derive the variance of $\widetilde{FIC}_{K_m:K_\ell}$ *as the following:*

$$
\begin{aligned}
& var\left[\widetilde{FIC}_{K_m:K_\ell}\right] \\
= & \, 2v^2_{K_m,K_m} + 2v^2_{K_\ell,K_\ell} + 4p^2_{K_m} v_{K_m,K_m} + 4p^2_{K_\ell} v_{K_\ell,K_\ell} + 4(p_{K_m} - p_{K_\ell})^2 v_{K_{full},K_{full}} \\
& + 4\left(v_{K_{full},K_{full}} v_{K_m,K_m} + v_{K_{full},K_{full}} v_{K_\ell,K_\ell} + v^2_{K_m,K_{full}} + v^2_{K_\ell,K_{full}}\right) \\
& - 4v^2_{K_\ell,K_m} - 8v_{K_m,K_m} v_{K_m,K_{full}} + 8v_{K_\ell,K_m} v_{K_m,K_{full}} + 8v_{K_\ell,K_m} v_{K_\ell,K_{full}} - 8v_{K_\ell,K_\ell} v_{K_\ell,K_{full}} \\
& - 8(p_{K_m} - p_{K_\ell}) p_{K_m} v_{K_m,K_{full}} + 8(p_{K_m} - p_{K_\ell}) p_{K_\ell} v_{K_\ell,K_{full}} - 8p_{K_m} p_{K_\ell} v_{K_\ell,K_m} \\
& - 8v_{K_{full},K_{full}} v_{K_\ell;K_m} - 8v_{K_\ell,K_{full}} v_{K_m,K_{full}}.
\end{aligned}
\tag{2.49}
$$

Estimating the A.M.S.E. using a sample-splitting strategy

Another statistic similar to the role of FIC was proposed in Vansteelandt et al.[107]. In Vansteelandt et al.[107], the authors employ a sample-splitting strategy to estimate the A.M.S.E.. First, the whole sample is divided into M pieces, indexed by $t = 1, \cdots, M$. Then for each $t = 1, \cdots, M$, the sample not in the t^{th} piece of the data, is used to estimate the target Ψ_0 using estimators of the form $\hat{\Psi}^{\setminus t}_{K_m}$ or of the form $\hat{\Psi}^{\setminus t}_{K_i,K_j,\cdots,K_\ell,K_m}$, where $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{\text{full}}$. Next for the same t, one estimate the target with the estimator, adjusting all the measured potential confounders with the sample in the t^{th} piece of the data and of the form $\hat{\Psi}^{t}_{K_{\text{full}}}$ or of the form $\hat{\Psi}^{t}_{K_i,\cdots,K_{\text{full}}}$ where $K_i \subseteq \cdots \subseteq K_{\text{full}}$. Then one estimates the A.M.S.E. by taking the ℓ_2-distance between the above two estimators across all $t = 1, \cdots, M$. Therefore we define the sample-splitting based estimator of standardized A.M.S.E. for the estimator with standardized A.M.S.E. $\tilde{\zeta}^*_{K_m}$.

Definition 2.5.3.

$$FIC^{\dagger M}_{K_m} = \frac{n}{M}\sum_{t=1}^{M}\left(\mathbb{P}^{\setminus t}_{n(M-1)/M}\left[b_0 + \frac{\hat{\epsilon}^{\setminus t}_{K_m}}{\pi_{K_m}}\right] - \mathbb{P}^{t}_{n/M}\left[b_0 + \frac{\hat{\epsilon}^{t}_{K_{full}}}{\pi_{K_{full}}}\right]\right)^2. \tag{2.50}$$

We denote the difference between $FIC^{\dagger M}_{K_m}$ and $FIC^{\dagger M}_{K_\ell}$ as $FIC^{\dagger M}_{K_m:K_\ell} := FIC^{\dagger M}_{K_m} - FIC^{\dagger M}_{K_\ell}$.

Then as derived in Section A.4, we obtain the following result:

Proposition 2.5.4. *As $n \to \infty$, $n/M \to \infty$, and $n(M-1)/M \to \infty$, under Definition 2.1.1 the limiting distribution of $FIC^{\dagger M}_{K_m}$ is*

$$\begin{aligned} FIC^{\dagger M}_{K_m} &\xrightarrow{d} \widetilde{FIC}^{\dagger M}_{K_m} \\ &:= \sum_{t=1}^{M}\left(\frac{1}{M-1}v^{1/2}_{K_m,K_m}Z^t_{K_m} + v^{1/2}_{K_{full},K_{full}}Z^t_{K_{full}}\right)^2 + \frac{M^2-2M}{(M-1)^2}v_{K_m,K_m}Z^2_{K_m} \\ &\quad - \frac{2M}{M-1}v^{1/2}_{K_m,K_m}v^{1/2}_{K_{full},K_{full}}Z_{K_m}Z_{K_{full}} - 2\left(v^{1/2}_{K_{full},K_{full}}Z_{K_{full}} - v^{1/2}_{K_m,K_m}Z_{K_m}\right)p_{K_m} + p^2_{K_m}. \end{aligned} \tag{2.51}$$

And apparently $FIC^{\dagger M}_{K_m:K_\ell} \xrightarrow{d} \widetilde{FIC}^{\dagger M}_{K_m:K_\ell} := \widetilde{FIC}^{\dagger M}_{K_m} - \widetilde{FIC}^{\dagger M}_{K_\ell}$.

Because confounder selection based on $\text{FIC}^{\dagger M}_{\cdot}$ is similar to FIC and relying on sample-splitting, we dub such selection strategy as FIC-CV throughout. In the original paper[107], the authors proposed heuristic searching algorithms such as stochastic search or deletion/addition/substitution algorithm mainly because of the computational complexity when the dimension of the covariate is relatively large. However, in our limiting distribution analyses, we decided to avoid such complication and just choose the estimator with the minimum $\widetilde{\text{FIC}}^{\dagger M}_{\cdot}$.

Similar to Corollary 2.4.3, we also provide the asymptotic distribution under $M \to \infty$ but at a slower rate than $n \to \infty$. We consider the difference of FIC-CV statistics $\widetilde{\text{FIC}}^{\dagger\, M}_{K_m:K_\ell} \overset{M\to\infty}{\to} \widetilde{\text{FIC}}^{\dagger}_{K_m:K_\ell}$ between models with standardized A.M.S.E. $\tilde{\zeta}^*_{K_m}$ and $\tilde{\zeta}^*_{K_\ell}$:

Corollary 2.5.5. *With the same conditions in Proposition 2.5.4, as $M \to \infty$ but at a slower rate than $n \to \infty$*

$$\begin{aligned}\widetilde{FIC}^{\dagger\, M}_{K_m:K_\ell} \xrightarrow{d} \widetilde{FIC}^{\dagger}_{K_m;K_\ell} = &\left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2 - \left(v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} + p_{K_\ell}\right)^2 + 2v_{K_m,K_{full}} - 2v_{K_\ell,K_{full}} \\ &- 2v^{1/2}_{K_{full},K_{full}} Z_{K_{full}} \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell}\right).\end{aligned} \tag{2.52}$$

Interestingly, $\widetilde{\text{FIC}}^{\dagger}_{K_m:K_\ell} \equiv \widetilde{\text{FIC}}_{K_m:K_\ell}$, i.e. as $M \to \infty$, the sample-splitting methods of estimating the A.M.S.E. is essentially equivalent to FIC in asymptotics.

2.6 An overview of the procedure related to Lepski's method

There exists a large body of literature on Lepski's method for non or semiparametric adaptive minimax estimations[44,45,46,47,43,8,54]. Similar strategies can be adopted to study the problem of interest in our paper, namely selecting a final estimator of some target Ψ_0 from a family of estimators, motivated from the discussion in Remark 2.4.5 about the existence of an equivalence between prediction risk comparison and hypothesis testing when comparing the two nested estimators with only one degree of freedom difference within the family of CTMLE estimators. For the sake of simplicity, we closely follow the approach and statistic taken in Spokoiny & Vial[89]. In the method described in Spokoiny & Vial[89], one needs to compare every pair of estimators in a specific order. Unlike the approach taken in CTMLE using prediction risk as selection criterion, every comparison can be written in the form of testing a hypothesis using χ^2 statistic. As a side note, since the same argument from Proposition 2.2.10 applies here, we do not have to distinguish between candidate estimators of the form $\hat{\Psi}_{K_m}$ and $\hat{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m}$ or between CTMLE estimators and Joffe's estimators when we are only interested in the asymptotic distributions. Now we define Lepski's statistic as in Spokoiny & Vial[89] when comparing estimators from the two equivalence classes $\hat{\Psi}^*_{K_m}$ and $\hat{\Psi}^*_{K_\ell}$, where $K_\ell, K_m \subseteq K_{\text{full}}$.

Definition 2.6.1. *The L-statistic used in Spokoiny & Vial*[89] *between estimators* $\hat{\Psi}^*_{K_m}$ *and* $\hat{\Psi}^*_{K_\ell}$*, where* $K_\ell, K_m \subseteq K_{full}$*, is defined as*

$$L_{K_m:K_\ell} := \frac{n\left(\hat{\Psi}_{K_m} - \hat{\Psi}_{K_\ell}\right)^2}{max\left(v_{K_m,K_m}, v_{K_\ell,K_\ell}\right)}. \tag{2.53}$$

Notice that in Spokoiny & Vial[89], the authors used $2\cdot\max\left(v_{K_m,K_m}, v_{K_\ell,K_\ell}\right)$ instead of $\max\left(v_{K_m,K_m}, v_{K_\ell,K_\ell}\right)$. We keep the current form to make it look closer to a usual χ^2 test statistic.

Then it is very straightforward to show the following result

Proposition 2.6.2. *As* $n \to \infty$*, under Definition 2.1.1, the Lepski's statistic between* $\hat{\Psi}_{K_m}$ *and* $\hat{\Psi}_{K_\ell}$ *eq.* (2.53) *when* $K_\ell, K_m \subseteq K_{full}$ *is*

$$L_{K_m:K_\ell} \xrightarrow{d} \tilde{L}_{K_m:K_\ell} := \frac{\left\{v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - p_{K_\ell}\right\}^2}{max\{v_{K_m,K_m}, v_{K_\ell,K_\ell}\}}. \tag{2.54}$$

Remark 2.6.3. *We make the following comments on Proposition 2.6.2:*

(1) As in FIC, using Lepski's statistic, any pairwise comparison can be applied to the asymptotic distribution derived in Proposition 2.6.2, not only restricted to the two nested estimators with only one degree of freedom difference.

(2) Assume $v_{K_m,K_m} > v_{K_\ell,K_\ell}$*. If one multiplies eq.* (2.53) *by* $\frac{v_{K_m,K_m}}{v_{K_m,K_m} - v_{K_\ell,K_m}}$*, then* $\frac{v_{K_m,K_m}}{v_{K_m,K_m} - v_{K_\ell,K_m}} \cdot \tilde{L}_{K_m:K_\ell} > 2$ *is equivalent to* $\tilde{\eta}^{\dagger}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} < 0$ *for* $K_i \subseteq K_j \subseteq \cdots \subseteq K_\ell \subseteq K_m \subseteq K_{full}$*. Under homoscedasticity,* $v_{K_m,K_m} - v_{K_\ell,K_m} = v_{K_m,K_m} + v_{K_\ell,K_\ell} - 2v_{K_\ell,K_m}$ *because* $v_{K_\ell,K_m} = v_{K_\ell,K_\ell}$*.*

Connection with usual hypothesis testing

In its core, Lespki-related method is an estimator selection tool based on hypothesis testing, with potential consideration of multiple testing correction when choosing the thresholding cutoff, such as the algorithm

proposed in Spokoiny & Vial[89]. However, it is not difficult to realize that if we treat $\sqrt{n}\left(\hat{\Psi}_{\underline{K_m}} - \hat{\Psi}_{\underline{K_\ell}}\right)$ as a single random variable and test whether it is significantly different from 0 or not, the usual test statistic is, under the null ($p_{K_m} - p_{K_\ell} = 0$),

$$\tilde{T}^{H_0}_{K_m:K_\ell} = \frac{\left(v^{1/2}_{K_m,K_m} Z_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell}\right)^2}{v_{K_\ell,K_\ell} + v_{K_m,K_m} - 2v_{K_\ell,K_m}} \sim \chi^2_1$$

Therefore the standardization of the usual test statistic is different from the L-statistic proposed in Spokoiny & Vial[89] on its denominator. But the two can be made equivalent by using different acceptance/rejection cutoff.

Connection with F- or t-statistics

Next we will explore the connection between the test statistic with asymptotic F- or t-distribution and the L-statistic. This connection can be explicitly established through M-fold cross validation. To see this, we again split our n independent vectors of data into M separate folds. Similarly, we studied the regime where $n \to \infty, n/M \to \infty$, and $n(M-1)/M \to \infty$. Define $\hat{\Psi}^{*\,t}_{\cdot}$ as the subset sample average version of $\hat{\Psi}_{\cdot}$ within the t-th portion of the data. Then $\hat{\Psi}^{*}_{K_m} = \frac{1}{M}\sum_{t=1}^{M} \hat{\Psi}^{*\,t}_{K_m}$. Then we can construct a test statistic resembling a F-test statistic denoted by $L^{\dagger M}_{K_m:K_\ell}$ as

$$L^{\dagger M}_{K_m:K_\ell} = \frac{\frac{1}{M}\left\{\sum_{t=1}^{M} \sqrt{n}\left(\hat{\Psi}^{*\,t}_{K_m} - \hat{\Psi}^{*\,t}_{K_\ell}\right)/M\right\}^2}{\frac{1}{M-1}\sum_{t=1}^{M}\left\{\sqrt{n}\left(\hat{\Psi}^{*\,t}_{K_m} - \hat{\Psi}^{*\,t}_{K_\ell}\right) - \sum_{t'=1}^{M}\sqrt{n}\left(\hat{\Psi}^{*\,t'}_{K_m} - \hat{\Psi}^{*\,t'}_{K_\ell}\right)/M\right\}^2}. \tag{2.55}$$

Under the $H_0 : p_{K_m} = p_{K_\ell}$ this a $F_{1,M-1}$ random variable under the null, instead of Gaussian random variable as those in Section 2.6. Then it is easy to see that, by defining $\tilde{L}^{\dagger M}_{K_m:K_\ell}$ as the ratio between the

asymptotic distribution of the numerator and denominator of eq. (2.55):

$$\tilde{L}^{\dagger M}_{K_m:K_\ell} = \frac{(M-1)\left\{\sum_{t=1}^{M}\left(v^{1/2}_{K_m,K_m}Z^t_{K_m} + \sqrt{\frac{1}{M}}p_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell} - \sqrt{\frac{1}{M}}p_{K_\ell}\right)/\sqrt{M}\right\}^2}{\sum_{t=1}^{M}\left\{\sqrt{M}\left(v^{1/2}_{K_m,K_m}Z^t_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}\right) - \sum_{t'=1}^{M}\left(v^{1/2}_{K_m,K_m}Z^{t'}_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^{t'}_{K_\ell}\right)/\sqrt{M}\right\}^2} > \mathcal{C}_{K_m,K_\ell}M, \tag{2.56}$$

where $\mathcal{C}_{K_m,K_\ell}$ can be tuned and will reject if $\tilde{L}^{\dagger M}_{K_m;K_\ell}$ exceeds $\mathcal{C}_{K_m,K_\ell}$.

Remark 2.6.4. *To make connections with the M-CV prediction risks in CTMLE algorithm, we can rewrite Equation* (2.56) *as*

$$\begin{aligned}&\frac{\mathcal{C}_{K_m,K_\ell}}{M-1}\sum_{t=1}^{M}\left(v^{1/2}_{K_m,K_m}Z^t_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z^t_{K_\ell}\right)^2 - \left(1+\frac{\mathcal{C}_{K_m,K_\ell}}{M-1}\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)^2\\ &-2\left(v^{1/2}_{K_m,K_m}Z_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)(p_{K_m}-p_{K_\ell}) - (p_{K_m}-p_{K_\ell})^2 < 0\end{aligned} \tag{2.57}$$

To help us compare Equation (2.57) *with the risk difference between two adjacent estimators, we copied Equation* (2.36) *here:*

$$\begin{aligned}&\frac{1}{(M-1)^2}\sum_{t=1}^{M}\left[(2M-1)v^{1/2}_{K_m,K_m}Z^t_{K_m} + v^{1/2}_{K_{m-1},K_{m-1}}Z^t_{K_{m-1}}\right]\left[v^{1/2}_{K_m,K_m}Z^t_{K_m} - v^{1/2}_{K_{m-1},K_{m-1}}Z^t_{K_{m-1}}\right]\\ &-\frac{M}{M-1}\left[v^{1/2}_{K_m,K_m}Z_{K_m} - \frac{M-2}{M}v^{1/2}_{K_{m-1},K_{m-1}}Z_{K_{m-1}}\right]\left[v^{1/2}_{K_m,K_m}Z_{K_m} - v^{1/2}_{K_{m-1},K_{m-1}}Z_{K_{m-1}}\right]\\ &-2\left(v^{1/2}_{K_m,K_m}Z_{K_m} - v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)(p_{K_m}-p_{K_\ell}) - (p_{K_m}-p_{K_\ell})^2 < 0\end{aligned} \tag{2.58}$$

When setting $\mathcal{C}_{K_m,K_\ell} = \frac{2M-1}{M-1}$ *and* $M \to \infty$, *Equation* (2.57) *and Equation* (2.58) *have the same limit up to a constant difference because asymptotically (in the sense of* $M \to \infty$*) the usual test statistic uses the correct asymptotic variance of the numerator whereas the* ∞*-CV prediction risk uses a slightly different*

"asymptotic variance" term:

$$eq.\ (2.57) \overset{M\to\infty}{\to} -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + 2\left(v_{K_m,K_m} + v_{K_\ell,K_\ell} - 2v_{K_\ell,K_m}\right) < 0$$

$$eq.\ (2.58) \overset{M\to\infty}{\to} -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + 2\left(v_{K_m,K_m} - v_{K_\ell,K_m}\right) < 0.$$

2.7 Summary of algebraic relations among the proposed confounder selection procedures

As a result of the effort of building up the asymptotic distributions in the above sections, in this section we lay out the algebraic connections among different confounder selection procedures.

We summarize the algebraic equivalences below:

1. The reweighted conditional population prediction risk difference is equivalent to the squared error difference when K_m contains all the confounders.

2. The marginal reweighted M-CV prediction risk difference is equivalent to the mean squared error difference when K_m contains all the confounders and $M \to \infty$ (i.e. equivalent to the marginal reweighted ∞-CV prediction risk difference).

3. The mean difference in FIC is equivalent to the mean squared error difference. And the difference in the sample-splitting statistic proposed in Vansteelandt et al.[107] (sample-splitting FIC) is equivalent to the FIC difference as $M \to \infty$, hence the mean difference is equivalent to the mean squared error difference as $M \to \infty$.

4. The conditional population prediction risk difference has the same mean as the M-CV prediction risk difference as $M \to \infty$, so does the ∞-CV prediction risk difference.

5. The conditional population prediction risk difference has the same mean as the FIC difference when K_m is the set of all collected potential confounders and the sample-splitting FIC difference when K_m is the set of all collected potential confounders as $M \to \infty$.

6. The within-sample prediction risk difference is equivalent to the M-CV prediction risk difference as $M \to \infty$, i.e. the ∞-CV prediction risk difference, up to a constant depending on K_m.

7. The square of the within-sample prediction risk difference is the numerator of the Lepski's statistic constructed as an asymptotic Normal distribution (Lepski's Z-statistic) under the null hypothesis (no bias).

8. As $M \to \infty$, the M-CV prediction risk difference or the ∞-CV prediction risk difference is equivalent to the FIC difference or the sample-splitting FIC difference as $M \to \infty$ when K_m contains all the collected potential confounders.

9. As $M \to \infty$, the comparison of the M-CV prediction risk difference or the ∞-CV prediction risk difference larger than or smaller than zero is equivalent to the comparison of the Lepski's Z-statistic larger than or smaller than 2 with a specific denominator and cutoff combination that depends on K_m.

10. For finite M, the comparison of the M-CV prediction risk difference larger than or smaller than zero is equivalent to the Lepski's F-statistic larger than or equal to $(2M-1)/(M-1)$ with the denominator as in eq. (2.55).

11. The within-sample prediction risk difference is equivalent to the M-CV prediction risk difference as $M \to \infty$, i.e. the ∞-CV prediction risk difference, up to a constant depending on both K_m and K_ℓ.

12. For finite M, there is no direct connection between the comparison of the M-CV prediction risk difference larger than or smaller than zero and the Lepski's F-statistic larger than or equal to $(2M-1)/(M-1)$ with the denominator as in eq. (2.55).

Apart from the connections, it is noteworthy to point out the difference between the Lepski's Z-statistic and the reweighted ∞-CV prediction risk comparison if the two estimators are not the two nested estimators with only one degree of freedom difference. Consider comparing the estimators $\hat{\Psi}_{K_\ell,K_k,K_m}$ and $\hat{\Psi}_{K_\ell}$

for $K_\ell \subseteq K_k \subseteq K_m$, the comparison based on the asymptotic distribution of Lepski's Z-statistic is

$$\tilde{L}_{K_m,K_k} = \frac{\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2}{v_{K_m,K_m} - v_{K_\ell,K_m}} \overset{?}{\lesseqgtr} C$$

whereas the comparison based on the reweighted ∞-CV prediction risk difference is

$$\begin{aligned}\tilde{\eta}^{\dagger w}_{K_\ell,K_k,K_m:K_\ell} = & -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_k,K_k}^{1/2} Z_{K_k} - p_{K_k}\right)^2 - \left(v_{K_k,K_k}^{1/2} Z_{K_k} + p_{K_k} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 \\ & + 2(v_{K_m,K_m} - v_{K_k,K_m}) + 2(v_{K_k,K_k} - v_{K_\ell,K_k}) \overset{?}{\gtreqless} 0\end{aligned}$$

Under homoscedasticity, $v_{K_\ell,K_k} = v_{K_\ell,K_m} = v_{K_\ell,K_\ell}$ and $v_{K_k,K_m} = v_{K_k,K_k}$. Therefore

$$\begin{aligned}\tilde{L}_{K_m,K_k} &= \frac{\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2}{v_{K_m,K_m} - v_{K_\ell,K_\ell}} \overset{?}{\lesseqgtr} C \\ \Leftrightarrow & -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + C(v_{K_m,K_m} - v_{K_\ell,K_\ell}) \overset{?}{\gtreqless} 0\end{aligned}$$

and

$$\begin{aligned}\tilde{\eta}^{\dagger w}_{K_\ell,K_k,K_m:K_\ell} = & -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_k,K_k}^{1/2} Z_{K_k} - p_{K_k}\right)^2 - \left(v_{K_k,K_k}^{1/2} Z_{K_k} + p_{K_k} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 \\ & + 2(v_{K_m,K_m} - v_{K_\ell,K_\ell}) \overset{?}{\gtreqless} 0.\end{aligned}$$

Then taking the difference of the LHS's:

$$\begin{aligned}& -\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + \left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} - v_{K_k,K_k}^{1/2} Z_{K_k} - p_{K_k}\right)^2 \\ & + \left(v_{K_k,K_k}^{1/2} Z_{K_k} + p_{K_k} - v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} - p_{K_\ell}\right)^2 + (C-2)(v_{K_m,K_m} - v_{K_\ell,K_\ell}) \\ = & \, 2\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)\left(v_{K_\ell,K_\ell}^{1/2} Z_{K_\ell} + p_{K_\ell}\right) - 2\left(v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m}\right)\left(v_{K_k,K_k}^{1/2} Z_{K_k} + p_{K_k}\right)\end{aligned}$$

$$
\begin{aligned}
&+2\left(v_{K_k,K_k}^{1/2}Z_{K_k}+p_{K_k}\right)^2-2\left(v_{K_k,K_k}^{1/2}Z_{K_k}+p_{K_k}\right)\left(v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}+p_{K_\ell}\right)+(C-2)(v_{K_m,K_m}-v_{K_\ell,K_\ell})\\
&=2\left(v_{K_m,K_m}^{1/2}Z_{K_m}+p_{K_m}-v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}-p_{K_\ell}\right)\left(v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}+p_{K_\ell}-v_{K_k,K_k}^{1/2}Z_{K_k}+p_{K_k}\right)\\
&\quad+(C-2)(v_{K_m,K_m}-v_{K_\ell,K_\ell}).
\end{aligned}
$$

When $p_{K_m}=p_{K_\ell}=p_{K_k}=0$, this difference becomes

$$
2\left(v_{K_m,K_m}^{1/2}Z_{K_m}-v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}\right)\left(v_{K_\ell,K_\ell}^{1/2}Z_{K_\ell}-v_{K_k,K_k}^{1/2}Z_{K_k}\right)+(C-2)(v_{K_m,K_m}-v_{K_\ell,K_\ell})
$$

2.8 Discussion

On a higher level we view our work as another complementary analysis of the CTMLE algorithm which has been quite widely accepted by practitioners who need to perform model selection when trying to estimate some causal effect. We open the possibilities that the algorithm could be extended to the current implementations by employing some other strategies, such as re-weighting the risks, using FIC-based metric, or incorporating other similar yet different selection strategies including Lepski's method. In a recent paper[82] the authors were more concerned about variance inflation issues when covariates include instrumental variables or can induce M-bias, which could be also fit into the proposed large sample $\sqrt{n}$-perturbation framework. A more systematic study similar to this paper but on small sample behaviors would be still vital to obtain a better sense on how different methods perform in settings closer to practice. Other possible extensions of our paper include generalizing our results to time-varying treatment or confounder cases[99,90] and higher-order bias correction cases[10]. Another potential area will fall into confounder selection in high dimension settings (e.g. $d=O(n)$ or $d\gg n$ where d is feature dimension), which is a very popular research area recently, including some important contributions such as Farrell[22], Belloni et al.[6], Athey et al.[3], Chernozhukov et al.[11].

3

Moving higher-order influence functions towards being practical – several directions of attempts

3.1 Introduction

Estimation of statistical functionals of data generating distribution is a fundamental problem in a variety of scientific disciplines, including (bio)statistics, epidemiology, economics, and signal processing. Examples include average treatment effect of a specific medical intervention in epidemiology and

medicine[81,70], effect of a specific education regime change in economics and political sciences[24] and estimating a mean value under data coarsening[26]. Understanding functional estimation for infinite-dimensional statistical models is especially important in modern era of scientific research because in most situations one wants to impose as few assumptions as possible on the data generating distribution to protect against potential model misspecification bias. Under the i.i.d. (independent and identically distributed) statistical model, one is often aiming at constructing an estimator to be of order $o_p(1/\sqrt{n})$ ($\sqrt{n}$-consistent), because standard error under i.i.d. model is of order $O(1/\sqrt{n})$ and with $\sqrt{n}$-consistent estimator one can build confidence interval with valid large-sample coverage probability.

A sequence of papers since 2008[68,74] developed a strategy to construct rate-optimal estimators for statistical functional based on higher-order U-statistic derived from the theory of higher-order influence functions[68]. It is also demonstrated[53] that such higher-order estimators are $\sqrt{n}$-consistent under further conditions which are proved to be minimal/necessary/tight[69,58].

To fix ideas, we consider the following data generating process: we observe n i.i.d. data vector $(Y_i \in \mathbb{R}, X_i \in \mathbb{R}^d)_{i=1}^n \sim \mathbb{P}_B$, where the statistical model/probability measure $\mathbb{P}_B$ is indexed by the parameter $B(\cdot) := \mathbb{E}[Y|X = \cdot]$. One can think of Y as some outcome of interest and X as a d-dimensional covariates. Instead of directly estimating B, we are actually interested in estimating the following functional of $\mathbb{P}_B$: $\Psi_0 := \mathbb{E}[\text{Var}[Y|X]] = \mathbb{E}[(Y - B(X))^2]$ where $\mathbb{E}[\cdot]$ is the expectation with respect to the true probability measure $\mathbb{P}_B$. This functional is important in many aspects and we will offer two motivations: (1) $\mathbb{E}[(Y - B(X))^2]$ has the same mathematical structure as square of L_2 norm or a quadratic functional[18]; (2) in causal inference, in the estimation of variance weighted average treatment effect under no unmeasured confounding, the quantity of interest can be equivalently written as $\frac{\mathbb{E}[\text{cov}[Y,A|X]]}{\mathbb{E}[\text{var}[Y|X]]}$, where A is the treatment indicator and $\mathbb{E}[(Y-B(X))^2]$ is exactly its denominator. To estimate Ψ_0, we have to estimate the unknown nuisance function $B(\cdot)$, which is the expectation of the outcome Y conditioning on the covariates X. $B(\cdot)$

is nuisance in the sense that we are not directly interested in $B(\cdot)$ but Ψ_0. One simple idea to estimate Ψ_0 is to use the conventional plug-in estimator: first one estimate the function $B(\cdot)$ from a set of sample to get the estimated function $\hat{B}(\cdot)$ and then one simply take sample average from another set of sample $\hat{\Psi}_{\text{plug-in}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{B}(X_i))^2$. To remark, this sample-splitting strategy regains its popularity in recent years with the advancement of nonparametric/machine learning methods of estimating regression functions to relax the Donsker-type assumptions in semiparametric inference[11,77,80] and to make valid distribution-free inference after model selection[67]. As straightforward as the plug-in approach is, because the estimator for the nuisance function $B(\cdot)$ is not directly designed for a better estimation of the parameter of interest Ψ_0, the plug-in approach often suffers from sub-optimal convergence rate, i.e. one can easily construct an estimator that attains faster rate than $\hat{\Psi}_{\text{plug-in}}$. One possibility is to use the theory of influence function originated from robust statistics[31] and later developed in semiparametric inference[7,101] to conduct bias correction[105]. Influence function (of first order, by default) is essentially the first-order gradient of the parameter of interest Ψ_0 along the direction of the first-order tangent space of the statistical model $\mathbb{P}_B$. Accidentally, the estimator $\hat{\Psi}_1$ after bias-correction by the influence function of Ψ_0 is equivalent to the plug-in estimator because the influence function is

$$\text{IF}(\Psi_0) = (y - B(x))^2 - \Psi_0.$$

Since $\mathbb{E}[\text{IF}(\Psi_0)] = 0$, then a reasonable estimator of Ψ_0 is simply $\frac{1}{n}\sum_{i=1}^{n}(Y_i - B(X_i))^2$ and similarly we replace the unknown $B(\cdot)$ by its estimator $\hat{B}(\cdot)$ from a training sample to obtain $\hat{\Psi}_1 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{B}(X_i))^2$. From here on, we will only call $\hat{\Psi}_{\text{plug-in}}$ as $\hat{\Psi}_1$ throughout.

Remark 3.1.1. *The equivalence between plug-in estimator and estimator after bias-correction by first-order influence function is in general not true. Examples can be found in the book Van der Laan &*

Robins[101].

We now demonstrate that $\hat{\Psi}_1$ can be suboptimal in terms of L_2 rate of convergence in a concrete example to motivate the usefulness of higher-order bias corrections using higher-order influence function. Suppose that the conditional expectation or regression function $B(\cdot)$ belongs to a Hölder space[28,21] with smoothness index β_b. Suppose $\beta_b/d = 1/4 + \varepsilon$, then the minimax rate of convergence of $\hat{B}(\cdot)$ to $B(\cdot)$ in L_2 norm is $\|\hat{B} - B\|_2 = O_p\left(n^{-\frac{\beta_b}{2\beta_b+d}}\right) = O_p(n^{-1/6+\varepsilon'})$. Then the bias of $\hat{\Psi}_1$ is $\mathbb{E}[(B(X) - \hat{B}(X))^2] = O_p(n^{-1/3+\varepsilon''})$. As a result, $\hat{\Psi}_1$ is not a $\sqrt{n}$-consistent estimator, so the usual α-level confidence interval based on large sample normal distribution theory will not have the valid coverage probability $1-\alpha$. However, the minimax rate of convergence result in Robins et al.[69] shows that $\beta_b > 1/4$ is the minimal condition for the existence of $\sqrt{n}$-consistent estimator of Ψ_0. Robins et al.[68] and Mukherjee et al.[53] explicitly constructed estimators achieving $\sqrt{n}$ rate of convergence using higher-order influence functions to perform higher-order bias correction.

3.2 Review of the statistical properties of empirical higher-order influence functions

For the sake of clarity, we will focus on second-order influence functions throughout this chapter and higher-order influence functions will be briefly discussed in Section 3.5. One easy interpretation of higher-order influence function is the following: consider the first-order bias of Ψ_0: $\mathbb{E}[(B(X) - \hat{B}(X))^2]$. Then second-order influence function can be viewed as a "data-driven" proxy of this first-order bias. Since this bias is impossible to be estimated because of the dependence on the unknown function $B(\cdot)$, one can instead project $B(x) - \hat{B}(x)$ onto a space spanned by a finite-dimensional (dim = k) basis functions $\{\Psi_\ell(x), \ell = 1, \ldots, k\}$, and this "projected first-order bias" is

$$\mathbb{E}[\Pi[Y - \hat{B}(X) | \{\Psi_\ell(x), \ell = 1, \ldots, k\}]^2]$$

which is exactly the population version of the second-order bias correction term based on second-order influence function. The second-order influence function based on k-dimensional basis $\{\Psi_\ell(x), \ell = 1, \ldots, k\}$ for Ψ_0 is nothing but

$$\mathbb{IF}_{2,2}\left(\vec{\Psi}\right) = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - B(X_{i_1}))\vec{\Psi}(X_{i_1})^\top \cdot \Omega \cdot \vec{\Psi}(X_{i_2})(Y_{i_2} - B(X_{i_2}))$$

where $\Omega = \Sigma^{-1}$ and $\Sigma = \mathbb{E}[\vec{\Psi}(X) \cdot \vec{\Psi}(X)^\top]$. The *bona fide* second-order influence function is thus

$$\hat{\mathbb{IF}}_{2,2}\left(\vec{\Psi}; \hat{\Omega}\right) = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \hat{B}(X_{i_1}))\vec{\Psi}(X_{i_1})^\top \cdot \hat{\Omega} \cdot \vec{\Psi}(X_{i_2})(Y_{i_2} - \hat{B}(X_{i_2}))$$

where $\hat{\Omega}$ is some estimator of Ω constructed from the training sample used to estimate $\hat{B}(\cdot)$. In Mukherjee et al.[53], the authors use the inverse of sample covariance matrix as an estimator of $\hat{\Omega}$ to construct the following empirical second-order influence function

$$\hat{\mathbb{IF}}_{2,2}\left(\vec{\Psi}; \left(\hat{\Sigma}_{\text{emp}}\right)^{-1}\right) = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \hat{B}(X_{i_1}))\vec{\Psi}(X_{i_1})^\top \cdot \left(\hat{\Sigma}_{\text{emp}}\right)^{-1} \cdot \vec{\Psi}(X_{i_2})(Y_{i_2} - \hat{B}(X_{i_2})).$$

Remark 3.2.1. *In the sequel, we will hide the dependence on the basis $\vec{\Psi}$. How to choose basis is definitely an important open problem to make higher-order influence functions but we choose not to focus on this huge issue in this chapter.*

The authors show that when the dimension of the finite-dimensional basis functions is k,

$$\begin{aligned}
\text{Var}\left[\hat{\mathbb{IF}}_{2,2}\left(\left(\hat{\Sigma}_{\text{emp}}\right)^{-1}\right)\right] &\asymp k/n^2 \\
\text{Bias}\left[\hat{\Psi}_1 + \hat{\mathbb{IF}}_{2,2}\left(\left(\hat{\Sigma}_{\text{emp}}\right)^{-1}\right)\right] &\lesssim \mathbb{E}[(B(X) - \hat{B}(X))^2] \cdot \|\hat{\Sigma}_{\text{emp}} - \Sigma\|_{\text{op}} \\
&= n^{-\frac{2\beta_b}{2\beta_b + d}} \cdot \sqrt{\frac{k \cdot \log(k)}{n}} + k^{-2\beta_b/d}
\end{aligned}$$

where $\|M\|_{\text{op}}$ is the operator norm of some matrix M. We use "$a \lesssim b$" and "$a \asymp b$" to denote $a \leq C_1 b$ with some constant $C_1 > 0$ and $a = C_2 b$ with some constant $C_2 > 0$.

Remark 3.2.2. *In the bias computation, the first term* $n^{-\frac{2\beta_b}{2\beta_b+d}} \cdot \sqrt{\frac{k \cdot log(k)}{n}}$ *is the so-called estimation bias*[68] *and the second term* $k^{-2\beta_b/d}$ *is the so-called approximation bias*[68] *induced by a finite-dimensional approximation of an infinite-dimensional space (in general a function space such as Hölder space, Sobolev space and etc.). Our philosophy is that such finite-dimensional approximation cannot be avoided in practice with the current computer system in the foreseeable future. The estimation bias is induced by estimating the function in the finite-dimensional approximated space of the original infinite-dimensional space, together with the estimation of covariance matrix.*

Remark 3.2.3. *Here* $k \cdot log(k)$ *needs to be at least* $o(n)$ *to make sure the upper bound on the bias of* $\hat{\mathbb{IF}}_{2,2}\left(\left(\hat{\Sigma}_{emp}\right)^{-1}\right)$ *to converge to zero. In Mukherjee et al.*[53]*, the authors choose* $k = n/log(n)^3$ *but this is not the unique choice.*

A natural question to ask is whether $\hat{\mathbb{IF}}_{2,2}\left(\vec{\Psi}; \left(\hat{\Sigma}_{\text{emp}}\right)^{-1}\right)$ is optimal or suboptimal. It turns out there exists another estimator with an improved upper bound on bias and we now study some of its statistical properties.

3.3 Leave-two-out second-order influence functions with faster rate of convergence

The idea of the to-be-proposed estimator is very simple: sample-splitting, which is motivated by a recent paper[58] using sample-splitting to bypass the need of higher-order influence functions for $\sqrt{n}$-estimable functionals under slightly stronger conditions than those proved to be necessary in Robins et al.[69]. One important observation is that for $\widehat{\mathbb{IF}}_{2,2}\left(\left(\hat{\Sigma}^{\text{tr}}_{\text{emp}}\right)^{-1}\right)$, the operator norm between the sample covariance matrix and true covariance matrix plays a key role in controlling the bias. If we use a sample-splitting

strategy to estimate the covariance matrix, it is plausible to speculate that we can obtain a better upper bound. This idea has also been used in many theoretical computation for rate of convergence type of problem. Consider the leave-two-out second-order influence functions, compared with $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$:

$$
\begin{aligned}
&\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo}}\right)^{-1}\right) \\
&= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \hat{B}(X_{i_1}))\vec{\Psi}_k(X_{i_1})^\top \cdot \left(\frac{1}{n-2} \sum_{j \neq i_1, i_2} \vec{\Psi}_k(X_j) \cdot \vec{\Psi}_k(X_j)^\top\right)^{-1} \cdot \vec{\Psi}_k(X_{i_2})(Y_{i_2} - \hat{B}(X_{i_2})), \\
&\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right) \\
&= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \hat{B}(X_{i_1}))\vec{\Psi}_k(X_{i_1})^\top \cdot \left(\frac{1}{n_{\text{tr}}} \sum_{\ell=1}^{n_{\text{tr}}} \vec{\Psi}_k(X_\ell) \cdot \vec{\Psi}_k(X_\ell)^\top\right)^{-1} \cdot \vec{\Psi}_k(X_{i_2})(Y_{i_2} - \hat{B}(X_{i_2})).
\end{aligned}
$$

We have the following result:

Proposition 3.3.1. *Under the same conditions as in Mukherjee et al.[53], we have*

$$
\begin{aligned}
Var\left[\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo}\right)^{-1}\right)\right] &\asymp k/n^2, \\
Bias\left[\hat{\Psi}_1 + \widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo}\right)^{-1}\right)\right] &\lesssim n^{-\frac{2\beta_b}{2\beta_b+d}} \cdot \frac{k \cdot log(k)}{n} + k^{-2\beta_b/d}.
\end{aligned}
$$

Proof. The result is a simple application of matrix Taylor expansion. □

Remark 3.3.2. *Under the condition that $k \cdot log(k) = o(n)$, $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo}\right)^{-1}\right)$ has a better bias upper bound than $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{tr}\right)^{-1}\right)$. However, there exists two major issues:*

1. Computation: in an unpublished manuscript, we showed that second order influence function, which is a second-order U-statistic, can be computed in linear time complexity as the sample size increases (quotient out the matrix multiplication part), if the kernel of the U-statistic is separable, in the following sense: for a bilinear function $h(x_{i_1}, x_{i_2})$ as the kernel of the U-statistic, then there exists two

functions f and g such that $h(x_{i_1}, x_{i_2}) = f(x_{i_1})g(x_{i_2})$. This result can be generalized to arbitrary higher-order influence functions but for the sake of clarity we do not state those results here. $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{tr}_{emp}\right)^{-1}\right)$ *satisfies this property but* $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{lo}_{emp}\right)^{-1}\right)$ *does not. Therefore* $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{lo}_{emp}\right)^{-1}\right)$ *is very hard to compute in practice when we even have moderate sample size n and relatively large k number of basis to make better approximation to the "true space".*

2. Leave-two-out estimator is quite unstable when k is large in our experience – one of the reasons is that for each pair i_1 and i_2, we have to obtain a new sample covariance matrix and then invert. This can be quite unstable when the dimension of the covariance matrix is large.

To address the above two concerns, we find an approximate version of $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo}}_{\text{emp}}\right)^{-1}\right)$ and call it $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo;quasi}}_{\text{emp}}\right)^{-1}\right)$. $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo;quasi}}_{\text{emp}}\right)^{-1}\right)$ is defined as the following:

$$\begin{aligned}
&\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo;quasi}}_{\text{emp}}\right)^{-1}\right) \\
&= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} (Y_{i_1} - \hat{B}(X_{i_1}))\vec{\Psi}_k(X_{i_1})^\top \cdot \left(\hat{\Sigma}^{-1} + \hat{\Sigma}^{-1} \cdot \hat{\Sigma}^{i_1,i_2} \cdot \hat{\Sigma}^{-1}\right) \cdot \vec{\Psi}_k(X_{i_2})(A_{i_2} - \hat{P}(X_{i_2}))
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{n} \sum_{j=1}^{n} \vec{\Psi}_k(X_j) \cdot \vec{\Psi}_k(X_j)^\top, \\
\hat{\Sigma}_{-i_1,-i_2} &= \frac{1}{n} \sum_{j \neq i_1, i_2} \vec{\Psi}_k(X_j) \cdot \vec{\Psi}_k(X_j)^\top, \\
\hat{\Sigma}_{i_1,i_2} &= \frac{1}{n}\vec{\Psi}_k(X_{i_1}) \cdot \vec{\Psi}_k(X_{i_1})^\top + \frac{1}{n}\vec{\Psi}_k(X_{i_2}) \cdot \vec{\Psi}_k(X_{i_2})^\top.
\end{aligned}$$

Remark 3.3.3. *It is easy to show that* $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{lo;quasi}_{emp}\right)^{-1}\right)$ *has separable U-statistic kernel and hence is easy to compute. The following result demonstrates that asymptotically* $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{lo;quasi}_{emp}\right)^{-1}\right)$ *has the*

same statistical property as $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo}\right)^{-1}\right)$.

Proposition 3.3.4. *Under the same conditions as in Mukherjee et al.*[53]*, we have*

$$Var\left[\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo;quasi}\right)^{-1}\right)\right] \asymp k/n^2,$$
$$Bias\left[\hat{\Psi}_1 + \widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{emp}^{lo;quasi}\right)^{-1}\right)\right] \lesssim n^{-\frac{2\beta_b}{2\beta_b+d}} \cdot \frac{k \cdot log(k)}{n} + k^{-2\beta_b/d}.$$

To demonstrate the superiority of $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ than $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$, we performed simulation study similar to that in Li et al.[48]. In the simulation, the sample size is $n = 2500$. We generate one-dimensional covariates X from a uniform distribution between $[0,1]$, and then generate a function $B(x)$ belonging to a Hölder space with smoothness index slightly greater than $1/4$. Then we assume $Y \sim B(X) + N(0,1)$ so Ψ_0 is simply 1. We report the simulation results in the following table:

Table 3.1: Finite sample performance comparison between $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ and $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$, with oracle *Monte Carlo* standard errors

# of basis	$\hat{\Psi}_1$	emp	s.e.	quasi-lo	s.e.	oracle	s.e.
512	1.20	-0.155	0.16	-0.102	0.024	-0.094	0.021
1024	1.20	Blow up	Blow up	-0.136	0.050	-0.15	0.029

In Table 3.1, the first-order estimator $\hat{\Psi}_1$ is 1.20, 0.20 away from the truth 1. The Daubechies dilated-and-translated father wavelets with six vanishing moments are chosen to be the basis functions $\vec{\Psi}_k$ truncated at two different levels: $k = 512$ and $k = 1,024$. In $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$, the true Σ^{-1} is only approximately true, which is computed by a huge sample size using sample covariance matrix ($n = 1,000,000$). We summarize the findings below: (1) When $k = 512$, $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ is -0.155, correcting roughly 77.5%

of the bias, which is higher than the bias correction of $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ $(0.102/0.2 = 51\%)$. But when compared to $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$ when we plug-in the "approximately true" inverse covariance matrix, the bias correction is roughly $0.094/0.2 = 47\%$. $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ is much closer to $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$ than $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ and $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ has a much higher standard error than the other two estimators in this choice of number of basis. (2) When $k = 1,024$, $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ blows up badly possibly because when $k \asymp n$, the sample covariance matrix is not even consistent to the true covariance matrix in operator norm. But $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ still performs sufficiently well, correcting $0.136/0.2 = 68\%$ of the total bias, with a relatively well-controlled standard error (0.050 vs. 0.029 for $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$). Compared with $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$, the bias correction is $0.15/0.2 = 75\%$, so $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ is quite close to $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$ in terms of bias correction and standard error.

To summarize, $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ is much closer to $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$ than $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ in terms of its statistical properties from the empirical observations made in finite-sample simulations, backing up the theoretical asymptotic analysis in this section.

3.4 Empirical shrinkage second-order influence functions – simulation studies on finite-sample performances and open problems

Even though $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{lo;quasi}}\right)^{-1}\right)$ does not suffer the drawback of $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}_{\text{emp}}^{\text{tr}}\right)^{-1}\right)$ in the simulation study performed in Section 3.3, one can expect that the use of sample covariance matrix prevents practitioners pushing empirical second-order influence function or empirical higher-order influence function in general to the "ideal" scenario when $k = n$, to decrease the approximation bias as far as possible. When $k \asymp n$, we would expect that the sample covariance matrix is a bad estimator of the true covariance matrix implied by the Marčenko-Pastur law[51]. Therefore a natural idea is to use shrinkage covariance matrix estimator[17] for a better risk control instead of favoring unbiasedness of sample covariance matrix.

Empirically, we found that the nonlinear shrinkage covariance matrix estimator developed since 2011 by Olivier Ledoit, Michael Wolf and their colleagues[37,38,40,41,39] works very well in practice. Even though we do not yet have a theoretical proof why the nonlinear shrinkage works well, heuristics tell us that one possible reason is that the nonlinear shrinkage covariance matrix estimator aims to minimizing the Frobenius norm between the estimator and the true covariance matrix. In the table below, we replace the column for $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{emp}}\right)^{-1}\right)$ in Table 3.1 by $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$, where $\hat{\Sigma}^{\text{tr}}_{\text{shrink}}$ is estimated from the training sample used to estimate the function $B(\cdot)$:

Table 3.2: Finite sample performance comparison between $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$ and $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo;quasi}}_{\text{emp}}\right)^{-1}\right)$, with oracle *Monte Carlo* standard errors

# of basis	$\hat{\Psi}_1$	nls	s.e.	quasi-lo	s.e.	oracle	s.e.
512	1.20	Blow up	Blow up	-0.102	0.024	-0.094	0.021
1024	1.20	-0.154	0.029	-0.136	0.050	-0.15	0.029
2048	1.20	-0.204	0.032	Blow up	Blow up	-0.191	0.035

It is quite remarkable that $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$ can even correct more bias than $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{lo;quasi}}_{\text{emp}}\right)^{-1}\right)$ when k is 1024 and is extremely close to $\widehat{\mathbb{IF}}_{2,2}\left(\Sigma^{-1}\right)$ when $k = 2048$, which is roughly 82% of the sample size n (2500). However, when k is relatively low (512), $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$ performs badly because to use $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$, one really relies heavily on the asymptotics of $p, n \to \infty$ and $p/n \to \gamma$ where $\gamma \in (0,1)$ and it's likely that when $k = 512$, the asymptotics have not kicked in. So this leaves us an important question to address in the future: when should we use shrinkage estimator and when should we use sample covariance matrix with different choice of k. Other open problems are: what is the rate of convergence for shrinkage-type covariance matrix estimator? and what is the rate of convergence for a

functional of shrinkage-type covariance matrix estimator and $\widehat{\mathbb{IF}}_{2,2}\left(\left(\widehat{\Sigma}^{\text{tr}}_{\text{shrink}}\right)^{-1}\right)$ is one special case.

3.5 Discussion

In this paper, we discussed several practical issues in applying empirical higher-order influence functions to de-bias first-step estimators based on nonparametric/machine-learning based estimation of nuisance parameters for follow-up valid inference (a.k.a. constructing confidence intervals with either finite-sample or asymptotic correct coverage probability). Such task might have important applications in the era of big data when functional estimation/inference is concerned in fields such as causal inference. The recently proposed empirical higher-order influence function[53] can perform badly potentially because of the difficulty of estimating a large dimensional inverse sample covariance matrix. It is well known from random matrix theory[113] that when the dimension p is close to sample size n in the sense of $p/n \to \gamma$ where $0 < \gamma < 1$, the eigenvalues/vectors of sample covariance matrix is hugely biased from the eigenvalues/vectors of the true covariance matrix. We demonstrated that certain modifications of the empirical higher-order influence function estimators originally proposed in Mukherjee et al.[53] can render better theoretical properties or finite-sample performance, using the idea of jackknifing and shrinkage. A more complete theoretical investigation on shrinkage empirical higher-order influence function is in working progress. Finally, the analysis conducted in this chapter is asymptotic ($n \to \infty$) in nature, and high-dimensional non-asymptotic analysis[108] of higher-order influence function with potentially tight constants is still an important open problem which might lead to non-asymptotic inference for statistical functionals.

A

Appendix for Chapter 2

A.1 Asymptotic equivalence between logistic binary DGP and homoscedastic conditional outcome variance in $\sqrt{n}$-perturbation regimes

In Section A.1, we will show that under Definition 2.1.1 if $Y \in \{0,1\}$, and assuming

$$Y|X \sim Bernoulli\left(p = expit\left\{b_0 + \frac{h(K_{d_Y})}{\sqrt{n}}\right\}\right).$$

where $expit(\cdot) = exp(\cdot)/(1+exp(\cdot))$, then we have asymptotically the same DGP as Definition 2.1.1 with homoscedastic conditional variance of the outcome after suitable reparamterization.

Proof.

$$\begin{aligned} Var[Y|X] =& \ exp\left\{b_0 + \frac{h(K_{d_Y})}{\sqrt{n}}\right\} \Big/ \left(1 + exp\left\{b_0 + \frac{h(K_{d_Y})}{\sqrt{n}}\right\}\right)^2 \\ \approx& \ \frac{e^{b_0} + h(K_{d_Y})/\sqrt{n}}{1 + 2\left(e^{b_0} + h(K_{d_Y})/\sqrt{n}\right) + o(1/\sqrt{n})} \\ \approx& \ \left(e^{b_0} + \frac{h_1}{\sqrt{n}}\right) \cdot \left(\frac{1}{1+2e^{b_0}} + \frac{h_2}{\sqrt{n}}\right) \\ \approx& \ \frac{e^{b_0}}{1+2e^{b_0}} + \frac{h_3}{\sqrt{n}}. \end{aligned}$$

The approximations are followed due to Taylor expansions up to order $O\left(\frac{1}{\sqrt{n}}\right)$. The representation for $\mathbb{E}[Y|X]$ follows similar strategy and we will omit the calculation here. Therefore we have a conditional outcome variance of Y not depending on any covariates X asymptotically up to suitable reparameterization.

□

A.2 Proof of results related to M-CV prediction risks and ∞-CV prediction risks

First we prove Proposition 2.4.2

Proof. According to the notations in Proposition 2.4.2, as $n \to \infty$ and $n/M \to \infty$

$$\begin{aligned} & \eta^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\ =& \frac{n}{M}\sum_{t=1}^{M} \left\{ \begin{array}{l} \mathbb{P}^t_{n/M}\left[\frac{R}{\pi^2_{K_m}}\right]\left(\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -2\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_m}}\left(Y - b_0 - \frac{\hat{\epsilon}^{\setminus t}_{K_i}}{\pi_{K_i}} - \cdots - \frac{\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_k;K_\ell}}{\pi_{K_\ell}}\right)\right] \end{array} \right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M}\sum_{t=1}^{M}\left\{\begin{array}{l}
\frac{M}{M-1}\mathbb{P}^t_{n/M}\left[\frac{R}{\pi^2_{K_m}}\right]\left(\sqrt{\frac{n(M-1)}{M}}\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\
-2\sqrt{\frac{M^2}{M-1}}\sqrt{\frac{n(M-1)}{M}}\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{\frac{n}{M}}\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_m}}(Y-b_0)\right] \\
+2n\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\left(\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_m}\pi_{K_i}}\right]\hat{\epsilon}^{\setminus t}_{K_i}+\cdots+\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_m}\pi_{K_\ell}}\right]\hat{\epsilon}^{\setminus t}_{K_i,K_j,\cdots,K_k;K_\ell}\right)
\end{array}\right\} \\
&\xrightarrow{d} \frac{1}{M}\sum_{t=1}^{M}\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\times \\
&\quad\left\{\begin{array}{l}
\frac{M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_{m-1}}\right)^2 \\
-\frac{2M}{\sqrt{M-1}}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_\ell}\right) \\
\times\left(v^{1/2}_{K_m,K_m}Z^{t}_{K_m}+\sqrt{\frac{1}{M}}(p_{K_m}+q)\right) \\
+\frac{2M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_\ell}\right) \\
\times\left(v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}+\sqrt{\frac{M-1}{M}}(p_{K_\ell}+q)\right)
\end{array}\right\} \\
&= \frac{1}{M}\sum_{t=1}^{M}\mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\times \\
&\quad\left\{\begin{array}{l}
\frac{M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right)^2-2\sqrt{M}(p_{K_m}-p_{K_\ell})v^{1/2}_{K_m,K_m}Z^{t}_{K_m} \\
-\frac{2M}{\sqrt{M-1}}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right)\left(v^{1/2}_{K_m,K_m}Z^{t}_{K_m}-\frac{1}{\sqrt{M-1}}v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right) \\
-(p_{K_m}-p_{K_\ell})^2+2\sqrt{\frac{M}{M-1}}(p_{K_m}-p_{K_\ell})v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}
\end{array}\right\} \\
&= \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1}\times \\
&\quad\left\{\begin{array}{l}
\frac{1}{(M-1)^2}\sum_{t=1}^{M}\left((2M-1)v^{1/2}_{K_m,K_m}Z^{t}_{K_m}+v^{1/2}_{K_\ell,K_\ell}Z^{t}_{K_\ell}\right)\left(v^{1/2}_{K_m,K_m}Z^{t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{t}_{K_\ell}\right) \\
-\frac{M}{(M-1)^2}\left(Mv^{1/2}_{K_m,K_m}Z_{K_m}-(M-2)v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right) \\
-2(p_{K_m}-p_{K_\ell})\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)-(p_{K_m}-p_{K_\ell})^2
\end{array}\right\}.
\end{aligned}
$$

The line of the convergence in distribution statement is due to central limit theorem and Slutsky's theorem. Other steps are purely algebra. □

Therefore one can easily derive the asymptotic distributions for the estimated risks using M-fold cross validation based on Proposition 2.4.2. Taking expectation over the Normal random variables to get the asymptotic expectation of the cross-validation risk, one gets

$$\mathbb{E}\left[\tilde{\eta}^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] = \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{\frac{M}{M-1} v_{K_m,K_m} - \frac{M}{M-1} v_{K_{m-1},K_{m-1}} - \left(p_{K_m} - p_{K_{m-1}}\right)^2\right\}, \tag{A.1}$$

compared to the asymptotic expectation of true prediction risk:

$$\mathbb{E}\left[\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] = \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]^{-1} \cdot \left\{v_{K_m,K_m} - v_{K_{m-1},K_{m-1}} - \left(p_{K_m} - p_{K_{m-1}}\right)^2\right\}. \tag{A.2}$$

Therefore as $M \to \infty$, $\mathbb{E}\left[\tilde{\eta}^{\dagger M}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right] \to \mathbb{E}\left[\tilde{\eta}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}\right]$.

Now we give the derivations for eq. (2.41) and eq. (2.42): As $n \to \infty$ and $n/M \to \infty$

$$\begin{aligned}
&\eta^{\dagger M \text{ quasi}}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&= \frac{n}{M}\sum_{t=1}^{M} \mathbb{P}^t_{n/M}\left[R\left(Y - b_0 - \hat{\varepsilon}^{\setminus t}_{K_i} - \cdots - \hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 - R\left(Y - b_0 - \hat{\varepsilon}^{\setminus t}_{K_i} - \cdots - \hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_k;K_\ell}\right)^2\right] \\
&= \frac{n}{M}\sum_{t=1}^{M} \left\{ \begin{array}{l} \mathbb{P}^t_{n/M}[R]\left(\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -2\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\mathbb{P}^t_{n/M}\left[R\left(Y - b_0 - \hat{\varepsilon}^{\setminus t}_{K_i} - \cdots - \hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_k;K_\ell}\right)\right] \end{array} \right\} \\
&= \frac{1}{M}\sum_{t=1}^{M} \left\{ \begin{array}{l} \frac{M}{M-1}\mathbb{P}^t_{n/M}[R]\left(\sqrt{\frac{n(M-1)}{M}}\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\right)^2 \\ -2\sqrt{\frac{M^2}{M-1}}\sqrt{\frac{n(M-1)}{M}}\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\sqrt{\frac{n}{M}}\mathbb{P}^t_{n/M}\left[\frac{R}{\pi_{K_0}}(Y - b_0)\right]\pi_{K_0} \\ +2\mathbb{P}^t_{n/M}[R] n\hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_\ell;K_m}\left(\hat{\varepsilon}^{\setminus t}_{K_i} + \cdots + \hat{\varepsilon}^{\setminus t}_{K_i,K_j,\cdots,K_k;K_\ell}\right) \end{array} \right\}
\end{aligned}$$

$$
\begin{aligned}
&\xrightarrow{d} \frac{1}{M}\sum_{t=1}^{M}\pi_{K_0}\times \\
&\left\{\begin{array}{l}
\frac{M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_{m-1}}\right)^2 \\
-\frac{2M}{\sqrt{M-1}}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_\ell}\right) \\
\times\left(v^{1/2}_{K_0,K_0}Z^{t}_{K_0}+\sqrt{\frac{1}{M}}(p_{K_0}+q)\right) \\
+\frac{2M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}+\sqrt{\frac{M-1}{M}}p_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}-\sqrt{\frac{M-1}{M}}p_{K_\ell}\right) \\
\times\left(v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}+\sqrt{\frac{M-1}{M}}(p_{K_\ell}+q)\right)
\end{array}\right\} \\
&=\frac{\pi_{K_0}}{M}\sum_{t=1}^{M}\left\{\begin{array}{l}
\frac{M}{M-1}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right)^2-2\sqrt{M}(p_{K_m}-p_{K_\ell})v^{1/2}_{K_0,K_0}Z^{t}_{K_0} \\
-\frac{2M}{\sqrt{M-1}}\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right)\left(v^{1/2}_{K_0,K_0}Z^{t}_{K_0}-\frac{1}{\sqrt{M-1}}v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right) \\
+(p_{K_m}-p_{K_\ell})(p_{K_m}+p_{K_\ell}-2p_{K_0})+2\sqrt{\frac{M}{M-1}}(p_{K_m}-p_{K_\ell})v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell} \\
+2\sqrt{\frac{M}{M-1}}(p_{K_m}-p_{K_0})\left(v^{1/2}_{K_m,K_m}Z^{\setminus t}_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z^{\setminus t}_{K_\ell}\right)
\end{array}\right\} \\
&=\pi_{K_0}\left\{\begin{array}{l}
\frac{M(M-2)}{(M-1)^2}v_{K_m,K_m}Z^2_{K_m}+\frac{1}{(M-1)^2}v_{K_m,K_m}\sum_{t=1}^{M}Z^{t\,2}_{K_m} \\
-\frac{M(M-2)}{(M-1)^2}v_{K_\ell,K_\ell}Z^2_{K_\ell}-\frac{1}{(M-1)^2}v_{K_\ell,K_\ell}\sum_{t=1}^{M}Z^{t\,2}_{K_\ell} \\
-\frac{2M}{M-1}v^{1/2}_{K_m,K_m}v^{1/2}_{K_0,K_0}Z_{K_m}Z_{K_0}+\frac{2}{M-1}v^{1/2}_{K_m,K_m}v^{1/2}_{K_0,K_0}\sum_{t=1}^{M}Z^{t}_{K_m}Z^{t}_{K_0} \\
+\frac{2M}{M-1}v^{1/2}_{K_\ell,K_\ell}v^{1/2}_{K_0,K_0}Z_{K_\ell}Z_{K_0}-\frac{2}{M-1}v^{1/2}_{K_\ell,K_\ell}v^{1/2}_{K_0,K_0}\sum_{t=1}^{M}Z^{t}_{K_\ell}Z^{t}_{K_0} \\
+2\left(v^{1/2}_{K_m,K_m}Z_{K_m}-v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}\right)(p_{K_m}-p_{K_0}) \\
+2\left(v^{1/2}_{K_\ell,K_\ell}Z_{K_\ell}-v^{1/2}_{K_0,K_0}Z_{K_0}\right)(p_{K_m}-p_{K_\ell})+(p_{K_m}-p_{K_\ell})(p_{K_m}+p_{K_\ell}-2p_{K_0})
\end{array}\right\} \\
&:=\eta^{\dagger M\text{ quasi}}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell}
\end{aligned}
$$

Again let's take the limit as $M\to\infty$, the quasi-∞-CV prediction risks for Joffe's estimator then have the following asymptotic distribution and again we can decompose the whole expression into one piece

corresponding to the asymptotic distribution of quasi-within-sample prediction risk and the other piece which is a constant. This is similar to what we observed in the CTMLE estimators.

$$
\begin{aligned}
&\tilde{\eta}^{\dagger \text{ quasi}}_{K_i,K_j,\cdots,K_\ell,K_m:K_i,K_j,\cdots,K_\ell} \\
&= \pi_{K_0} \cdot \left\{ \begin{array}{l} v_{K_m,K_m} Z^2_{K_m} - v_{K_\ell,K_\ell} Z^2_{K_\ell} - 2 v^{1/2}_{K_0,K_0} v^{1/2}_{K_m,K_m} Z_{K_0} Z_{K_m} + 2 v^{1/2}_{K_0,K_0} v^{1/2}_{K_\ell,K_\ell} Z_{K_0} Z_{K_\ell} \\ +2 \left(v^{1/2}_{K_m,K_m} Z_{K_m} - v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} \right) (p_{K_m} - p_{K_0}) + 2 \left(v^{1/2}_{K_\ell,K_\ell} Z_{K_\ell} - v^{1/2}_{K_0,K_0} Z_{K_0} \right) (p_{K_m} - p_{K_\ell}) \\ +(p_{K_m} - p_{K_\ell})(p_{K_m} + p_{K_\ell} - 2p_{K_0}) + 2 v_{K_0,K_m} - 2 v_{K_0,K_\ell} \end{array} \right\} \\
&= \text{R.H.S. of eq. (2.24)} + \pi_{K_0} (2 v_{K_0,K_m} - 2 v_{K_0,K_\ell}).
\end{aligned}
\tag{A.3}
$$

Similarly

$$
\begin{aligned}
&\eta^{\dagger M \text{ quasi}}_{K_m:} \\
&:= \frac{n}{M} \sum_{t=1}^{M} \mathbb{P}^t_{n/M} \left[R \left(Y - b_0 - \hat{\varepsilon}^{\setminus t}_{K_m} \right)^2 - R(Y - b_0)^2 \right] \\
&= \frac{1}{M} \sum_{t=1}^{M} \left\{ \frac{M}{M-1} \mathbb{P}^t_{n/M}[R] \left(\sqrt{\frac{n(M-1)}{M}} \hat{\varepsilon}^{\setminus t}_{K_m} \right)^2 - 2 \frac{M \pi_{K_0}}{\sqrt{M-1}} \sqrt{\frac{n}{M}} \hat{\varepsilon}^t_{K_0} \sqrt{\frac{n(M-1)}{M}} \hat{\varepsilon}^{\setminus t}_{K_m} \right\} \\
&\xrightarrow{d} \frac{\pi_{K_0}}{M-1} \sum_{t=1}^{M} \left\{ v^{1/2}_{K_m,K_m} \sqrt{\frac{M}{M-1}} \left(Z_{K_m} - \sqrt{\frac{1}{M}} Z^t_{K_m} \right) + \sqrt{\frac{M-1}{M}} (p_{K_m} + q) \right\}^2 \\
&\quad - \frac{2\pi_{K_0}}{\sqrt{M-1}} \sum_{t=1}^{M} \left(v^{1/2}_{K_m,K_m} \sqrt{\frac{M}{M-1}} \left(Z_{K_m} - \sqrt{\frac{1}{M}} Z^t_{K_m} \right) + \sqrt{\frac{M-1}{M}} (p_{K_m} + q) \right) \\
&\quad \cdot \left(v^{1/2}_{K_0,K_0} Z^t_{K_0} + \sqrt{\frac{1}{M}} (p_{K_0} + q) \right) \\
&= \pi_{K_0} \times
\end{aligned}
$$

$$
\left\{
\begin{array}{l}
\dfrac{M(M-2)}{(M-1)^2} v_{K_m,K_m} Z_{K_m}^2 - \dfrac{2M}{M-1} v_{K_0,K_0}^{1/2} v_{K_m,K_m}^{1/2} Z_{K_0} Z_{K_m} + \dfrac{1}{(M-1)^2} v_{K_m,K_m} \sum_{t=1}^{M} Z_{K_m}^{t\,2} \\
+ \dfrac{2}{M-1} v_{K_0,K_0}^{1/2} v_{K_m,K_m}^{1/2} \sum_{t=1}^{M} Z_{K_0}^t Z_{K_m}^t - 2 v_{K_m,K_m}^{1/2} Z_{K_m} (p_{K_0} + q) \\
+ 2 \left(v_{K_m,K_m}^{1/2} Z_{K_m} - v_{K_0,K_0}^{1/2} Z_{K_0} \right) (p_{K_m} + q) + (p_{K_m} + q)(p_{K_m} - 2p_{K_0} - q)
\end{array}
\right\}
$$

$$
:= \tilde{\eta}_{K_m:}^{\dagger M \text{ quasi}}.
$$

We take the limit as $M \to \infty$:

$$
\tilde{\eta}_{K_m:}^{\dagger \text{ quasi}} = \pi_{K_0} \cdot \{\text{R.H.S. of eq. (2.27)} + 2 v_{K_0,K_m}\}.
$$

A.3 Proof of large sample distribution of squared error loss using least-square estimation strategy

Proof. Suppose the estimator involves the following subsets of covariates $K_i \subset K_j \subset \cdots \subset K_\ell \subset K_m \subset K_{\text{full}}$, and the least-square based estimator $\bar{\Psi}_{K_i,K_j,\cdots,K_\ell,K_m} = \mathbb{P}_n \left[b_0 + \dfrac{\bar{\epsilon}_{K_i}}{\pi_{K_i}} + \dfrac{\bar{\epsilon}_{K_j}}{\pi_{K_j}} + \cdots + \dfrac{\bar{\epsilon}_{K_\ell}}{\pi_{K_\ell}} + \dfrac{\bar{\epsilon}_{K_m}}{\pi_{K_m}} \right]$. The asymptotic distribution for $[\sqrt{n}\bar{\epsilon}_{K_i}, \cdots, \sqrt{n}\bar{\epsilon}_{K_m}]'$ is

$$
\begin{pmatrix} \sqrt{n}\bar{\epsilon}_{K_i} \\ \sqrt{n}\bar{\epsilon}_{K_j} \\ \vdots \\ \sqrt{n}\bar{\epsilon}_{K_m} \end{pmatrix}
\xrightarrow{d}
\begin{pmatrix}
\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] \\
\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] \\
\vdots & \vdots & \ddots & \vdots \\
\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right]
\end{pmatrix}^{-1}
\cdot
\begin{pmatrix} v_{K_i,K_i}^{1/2} Z_{K_i} + p_{K_i} + q \\ v_{K_j,K_j}^{1/2} Z_{K_j} + p_{K_j} + q \\ \vdots \\ v_{K_m,K_m}^{1/2} Z_{K_m} + p_{K_m} + q \end{pmatrix}. \quad \text{(A.4)}
$$

Then an application of Woodbury identity[110,36] gives

$$\left(\mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] \quad \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] \quad \cdots \quad \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right] \right) \cdot \begin{pmatrix} \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] \\ \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[\frac{1}{\pi_{K_i}}\right] & \mathbb{E}\left[\frac{1}{\pi_{K_j}}\right] & \cdots & \mathbb{E}\left[\frac{1}{\pi_{K_m}}\right] \end{pmatrix}^{-1} = (0\ 0\ \cdots\ 0\ 1)\,.$$

Therefore $\bar{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m} \xrightarrow{d} \tilde{\zeta}_{K_i,K_j,\cdots,K_\ell,K_m} = \left(v^{1/2}_{K_m,K_m} Z_{K_m} + p_{K_m}\right)^2$. □

A.4 Asymptotic distribution of the statistics used for model selection in Vansteelandt et al.[107]

Again we assume $n/M \to \infty$ and $n \to \infty$ simultaneously.

$$\begin{aligned}
&\widehat{\mathrm{FIC}}^{M}_{K_m} \\
&= \frac{n}{M}\sum_{t=1}^{M}\left(\mathbb{P}^{\setminus t}_{n(M-1)/M}\left[b_0 + \frac{\hat{\epsilon}^{\setminus t}_{K_m}}{\pi_{K_m}}\right] - \mathbb{P}^{t}_{n/M}\left[b_0 + \frac{\hat{\epsilon}^{t}_{K_{\text{full}}}}{\pi_{K_{\text{full}}}}\right]\right)^2 \\
&= \frac{1}{M}\sum_{t=1}^{M}\left(\sqrt{\frac{n(M-1)}{M}}\sqrt{\frac{M}{M-1}}\mathbb{P}^{\setminus t}_{n(M-1)/M}\left[\frac{1}{\pi_{K_m}}\right]\hat{\epsilon}^{\setminus t}_{K_m} - \sqrt{\frac{n}{M}}\sqrt{M}\mathbb{P}^{t}_{n/M}\left[\frac{1}{\pi_{K_{\text{full}}}}\right]\hat{\epsilon}^{t}_{K_{\text{full}}}\right)^2 \\
&\xrightarrow{d} \frac{1}{M}\sum_{t=1}^{M}\left(\frac{M}{M-1}v^{1/2}_{K_m,K_m}Z_{K_m} + p_{K_m} - \frac{\sqrt{M}}{M-1}v^{1/2}_{K_m,K_m}Z^t_{K_m} - \sqrt{M}v^{1/2}_{K_{\text{full}},K_{\text{full}}}Z^t_{K_{\text{full}}}\right)^2 \\
&= \sum_{t=1}^{M}\left(\frac{1}{M-1}v^{1/2}_{K_m,K_m}Z^t_{K_m} + v^{1/2}_{K_{\text{full}},K_{\text{full}}}Z^t_{K_{\text{full}}}\right)^2 + \frac{M^2-2M}{(M-1)^2}v_{K_m,K_m}Z^2_{K_m} - \frac{2M}{M-1}v^{1/2}_{K_m,K_m}v^{1/2}_{K_{\text{full}},K_{\text{full}}}Z_{K_m}Z_{K_{\text{full}}} \\
&\quad - 2\left(v^{1/2}_{K_{\text{full}},K_{\text{full}}}Z_{K_{\text{full}}} - v^{1/2}_{K_m,K_m}Z_{K_m}\right)p_{K_m} + p^2_{K_m}\,.
\end{aligned}$$

References

[1] Altrock, P. M., Liu, L. L., & Michor, F. (2015). The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12), 730.

[2] Apostol, T. M. (1974). Mathematical Analysis.

[3] Athey, S., Imbens, G. W., & Wager, S. (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.

[4] Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

[5] Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R. M., Ramaswamy, S., & Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous ipsc generation. *Nature methods*, 11(11), 1170–1176.

[6] Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, (just-accepted), 1–36.

[7] Bickel, P. J., Klaassen, C. A., Ritov, Y., & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore.

[8] Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, (pp. 113–133).

[9] Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A., & Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6), 1209–1222.

[10] Carone, M., Díaz, I., & van der Laan, M. J. (2014). Higher-order targeted minimum loss-based estimation.

[11] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

[12] Claeskens, G. & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464), 900–916.

[13] Claeskens, G. & Hjort, N. L. (2008). *Model selection and model averaging*, volume 330. Cambridge University Press.

[14] Crawford, F. W., Minin, V. N., & Suchard, M. A. (2014). Estimation for general birth-death processes. *Journal of the American Statistical Association*, 109(506), 730–747.

[15] Di Stefano, B., Sardina, J. L., van Oevelen, C., Collombet, S., Kallin, E. M., Vicent, G. P., Lu, J., Thieffry, D., Beato, M., & Graf, T. (2014). C/ebpa poises b cells for rapid reprogramming into induced pluripotent stem cells. *Nature*, 506(7487), 235–239.

[16] Dinh, V., Rundell, A. E., & Buzzard, G. T. (2014). Experimental design for dynamics identification of cellular processes. *Bulletin of mathematical biology*, 76(3), 597–626.

[17] Donoho, D. L., Gavish, M., & Johnstone, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. *arXiv preprint arXiv:1311.0851*.

[18] Donoho, D. L. & Nussbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3), 290–323.

[19] Duffy, K. R., Wellard, C. J., Markham, J. F., Zhou, J. H., Holmberg, R., Hawkins, E. D., Hasbold, J., Dowling, M. R., & Hodgkin, P. D. (2012). Activation-induced b cell fates are selected by intracellular stochastic competition. *Science*, 335(6066), 338–341.

[20] Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

[21] Evans, L. C. (2010). *Partial Differential Equations 2nd Edition*. Graduate studies in mathematics. American Mathematical Society.

[22] Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1), 1–23.

[23] Foo, J., Liu, L. L., Leder, K., Riester, M., Iwasa, Y., Lengauer, C., & Michor, F. (2015). An evolutionary approach for identifying driver mutations in colorectal cancer. *PLoS computational biology*, 11(9), e1004350.

[24] Fryer Jr, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*, 126(4), 1755–1798.

[25] Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press New York, NY, USA.

[26] Gill, R. D., Van Der Laan, M. J., & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics* (pp. 255–294).: Springer.

[27] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25), 2340–2361.

[28] Giné, E. & Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press.

[29] Gruber, S. & van der Laan, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1).

[30] Guo, S., Zi, X., Schulz, V. P., Cheng, J., Zhong, M., Koochaki, S. H., Megyola, C. M., Pan, X., Heydari, K., Weissman, S. M., et al. (2014). Nonstochastic reprogramming from a privileged somatic cell state. *Cell*, 156(4), 649–662.

[31] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346), 383–393.

[32] Hanna, J., Saha, K., Pando, B., Van Zon, J., Lengner, C. J., Creyghton, M. P., van Oudenaarden, A., & Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273), 595–601.

[33] Hanna, J., Wernig, M., Markoulaki, S., Sun, C.-W., Meissner, A., Cassady, J. P., Beard, C., Brambrink, T., Wu, L.-C., Townes, T. M., et al. (2007). Treatment of sickle cell anemia mouse model with ips cells generated from autologous skin. *Science*, 318(5858), 1920–1923.

[34] Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464), 879–899.

[35] Ho, L. S. T., Xu, J., Crawford, F. W., Minin, V. N., & Suchard, M. A. (2017). Birth/birth-death processes and their computable transition probabilities with biological applications. *Journal of mathematical biology*, (pp. 1–34).

[36] Kollo, T. & von Rosen, D. (2006). *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media.

[37] Ledoit, O. & Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2), 233–264.

[38] Ledoit, O. & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2), 1024–1060.

[39] Ledoit, O. & Wolf, M. (2017a). Direct nonlinear shrinkage estimation of large-dimensional covariance matrices.

[40] Ledoit, O. & Wolf, M. (2017b). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *The Review of Financial Studies*, 30(12), 4349–4388.

[41] Ledoit, O. & Wolf, M. (2017c). Numerical implementation of the quest function. *Computational Statistics & Data Analysis*, 115, 199–223.

[42] Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01), 21–59.

[43] Lepski, O. V., Mammen, E., & Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, (pp. 929–947).

[44] Lepskii, O. V. (1990). A problem of adaptive estimation in gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya*, 35(3), 459–470.

[45] Lepskii, O. V. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4), 682–697.

[46] Lepskii, O. V. (1993). Asymptotically minimax adaptive estimation. II. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3), 433–448.

[47] Lepskii, O. V. & Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, (pp. 2512–2546).

[48] Li, L., Tchetgen, E. T., van der Vaart, A., & Robins, J. (2005). Robust inference with higher order influence functions: Part II. In *Joint Statistical Meetings, Minneapolis, Minnesota*.

[49] Liu, L. L., Brumbaugh, J., Bar-Nur, O., Smith, Z., Stadtfeld, M., Meissner, A., Hochedlinger, K., & Michor, F. (2016). Probabilistic modeling of reprogramming to induced pluripotent stem cells. *Cell reports*, 17(12), 3395–3406.

[50] Liu, L. L., Li, F., Pao, W., & Michor, F. (2015). Dose-dependent mutation rates determine optimum erlotinib dosing strategies for egfr mutant non-small cell lung cancer patients. *PloS one*, 10(11), e0141665.

[51] Marčenko, V. A. & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 457.

[52] Morris, R., Sancho-Martinez, I., Sharpee, T. O., & Belmonte, J. C. I. (2014). Mathematical approaches to modeling development and reprogramming. *Proceedings of the National Academy of Sciences*, 111(14), 5076–5082.

[53] Mukherjee, R., Newey, W. K., & Robins, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*.

[54] Mukherjee, R., Tchetgen, E. T., & Robins, J. (2015). Lepski's method and adaptive estimation of nonlinear integral functionals of density. *arXiv preprint arXiv:1508.00249*.

[55] Murrell, D. J., Dieckmann, U., & Law, R. (2004). On moment closures for population dynamics in continuous space. *Journal of Theoretical Biology*, 229(3), 421–432.

[56] Nåsell, I. (2003). Moment closure and the stochastic logistic model. *Theoretical Population Biology*, 63(2), 159–168.

[57] Nelsen, R. B. & Schmidt, H. (1991). Chains in power sets. *Mathematics Magazine*, 64(1), 23–31.

[58] Newey, W. K. & Robins, J. M. (2017). *Cross-fitting and fast remainder rates for semiparametric estimation*. Technical report, Working paper, MIT.

[59] Parzen, E. (1999). *Stochastic processes*. SIAM.

[60] Pasque, V., Tchieu, J., Karnik, R., Uyeda, M., Dimashkie, A. S., Case, D., Papp, B., Bonora, G., Patel, S., Ho, R., et al. (2014). X chromosome reactivation dynamics reveal stages of reprogramming to pluripotency. *Cell*, 159(7), 1681–1697.

[61] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688.

[62] Pearl, J. (2009). Remarks on the method of propensity score. *Statistics in Medicine*, 28(9), 1415–6.

[63] Polo, J. M., Anderssen, E., Walsh, R. M., Schwarz, B. A., Nefzger, C. M., Lim, S. M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J., et al. (2012). A molecular roadmap of reprogramming somatic cells into ips cells. *Cell*, 151(7), 1617–1632.

[64] Polo, J. M., Liu, S., Figueroa, M. E., Kulalert, W., Eminli, S., Tan, K. Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T., et al. (2010). Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nature biotechnology*, 28(8), 848–855.

[65] Pour, M., Pilzer, I., Rosner, R., Smith, Z. D., Meissner, A., & Nachman, I. (2015). Epigenetic predisposition to reprogramming fates in somatic cells. *EMBO reports*, (pp. e201439264).

[66] Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A. A., Caspi, I., Krupalnik, V., Zerbib, M., et al. (2013). Deterministic direct reprogramming of somatic cells to pluripotency. *Nature*, 502(7469), 65–70.

[67] Rinaldo, A., Wasserman, L., G'Sell, M., Lei, J., & Tibshirani, R. (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*.

[68] Robins, J., Li, L., Tchetgen, E., & van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman* (pp. 335–421). Institute of Mathematical Statistics.

[69] Robins, J., Tchetgen, E. T., Li, L., van der Vaart, A., et al. (2009). Semiparametric minimax rates. *Electronic Journal of Statistics*, 3, 1305–1321.

[70] Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9), 1393–1512.

[71] Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9), 923–945.

[72] Robins, J. M. & Greenland, S. (1986). The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*, 123(3), 392–402.

[73] Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, (pp. 550–560).

[74] Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. T., & van der Vaart, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5), 1951–1987.

[75] Robins, J. M. & Morgenstern, H. (1987). The foundations of confounding in epidemiology. *Computers & Mathematics with Applications*, 14(9), 869–916.

[76] Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22(4), 544–559.

[77] Robins, J. M., Zhang, P., Ayyagari, R., Logan, R., Tchetgen, E., Li, L., Lumley, T., van der Vaart, A., Committee, H. H. R., et al. (2013). New statistical approaches to semiparametric regression with application to air pollution research. *Research report (Health Effects Institute)*, (175), 3.

[78] Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

[79] Rotnitzky, A., Lei, Q., Sued, M., & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2), 439–456.

[80] Rotnitzky, A., Robins, J., & Babino, L. (2017). On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*.

[81] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

[82] Schnitzer, M. E., Lok, J. J., & Gruber, S. (2015). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The International Journal of Biostatistics*.

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

[84] Sherman, J. & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1), 124–127.

[85] Slutsky, E. (1925). Über stochastische asymptoten und grenzwerte. *Metron*, 5(3), 3–89.

[86] Smith, G. D. (1985). *Numerical solution of partial differential equations: finite difference methods*. Oxford university press.

[87] Smith, Z. D., Nachman, I., Regev, A., & Meissner, A. (2010). Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. *Nature biotechnology*, 28(5), 521–526.

[88] Soetaert, K., Petzoldt, T., Setzer, R. W., et al. (2010). Solving differential equations in R: package desolve. *Journal of Statistical Software*, 33(9), 1–25.

[89] Spokoiny, V. G. & Vial, C. (2009). Parameter tuning in pointwise adaptation using a propagation approach. *The Annals of Statistics*, (pp. 2783–2807).

[90] Stitelman, O. M., de Gruttola, V., & van der Laan, M. J. (2012). A general implementation of TMLE for longitudinal data applied to causal inference in survival analysis. *The International Journal of Biostatistics*, 8(1).

[91] Stitelman, O. M. & van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1).

[92] Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *cell*, 131(5), 861–872.

[93] Takahashi, K. & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4), 663–676.

[94] Tan, W. & Piantadosi, S. (1991). On stochastic growth processes with application to stochastic logistic growth. *Statistica Sinica*, 1, 527–540.

[95] Taylor, H. M. & Karlin, S. (2014). *An Introduction to Stochastic Modeling*. Academic Press.

[96] Tran, K. A., Jackson, S. A., Olufs, Z. P., Zaidan, N. Z., Leng, N., Kendziorski, C., Roy, S., & Sridharan, R. (2015). Collaborative rewiring of the pluripotency network by chromatin and signalling modulating pathways. *Nature communications*, 6.

[97] van der Laan, M. J. & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

[98] van der Laan, M. J. & Gruber, S. (2011). Targeted minimum loss based estimation of an intervention specific mean outcome.

[99] van der Laan, M. J. & Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1).

[100] van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

[101] Van der Laan, M. J. & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

[102] van der Laan, M. J. & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

[103] van der Laan, M. J. & Rose, S. (2017). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Forthcoming.

[104] van der Laan, M. J. & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

[105] van der Vaart, A. (2014). Higher order tangent spaces and influence functions. *Statistical Science*, (pp. 679–686).

[106] VanderWeele, T. J. & Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, 41(1), 196.

[107] Vansteelandt, S., Bekaert, M., & Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1), 7–30.

[108] Vershynin, R. (2018). *High-Dimensional Probability – An Introduction with Applications in Data Science*. Cambridge University Press.

[109] Vidal, S. E., Amlani, B., Chen, T., Tsirigos, A., & Stadtfeld, M. (2014). Combinatorial modulation of signaling pathways reveals cell-type-specific requirements for highly efficient and synchronous ipsc reprogramming. *Stem cell reports*, 3(4), 574–584.

[110] Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum Report*, 42, 106.

[111] Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. *Nature*, 460(7251), 49.

[112] Yan, J., Zheng, P., & Pan, X. (2014). Theoretical modelling discriminates the stochastic and deterministic hypothesis of cell reprogramming. *arXiv preprint arXiv:1409.2205*.

[113] Yao, J., Zheng, S., & Bai, Z.-D. (2015). *Large sample covariance matrices and high-dimensional data analysis*. Cambridge University Press.