High-Dimensional Knockoffs Inference for Time Series Data[*]

Chien-Ming Chi[1], Yingying Fan[2], Ching-Kang Ing[3] and Jinchi Lv[2]

Academia Sinica[1], University of Southern California[2]

and National Tsing Hua University[3]

November 10, 2024

Abstract

We make some initial attempt to establish the theoretical and methodological foundation for the model-X knockoffs inference for time series data. We suggest the method of time series knockoffs inference (TSKI) by exploiting the ideas of subsampling and e-values to address the difficulty caused by the serial dependence.

[*] Chien-Ming Chi is Assistant Research Fellow, Institute of Statistical Science, Academia Sinica, Taiwan (E-mail: *xbbchi@stat.sinica.edu.tw*). Yingying Fan is Centennial Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *fanyingy@marshall.usc.edu*). Ching-Kang Ing is Tsing Hua Chair Professor, Institute of Statistics, National Tsing Hua University, Hsinchu 300044, Taiwan (E-mail: *cking@stat.nthu.edu.tw*). Jinchi Lv is Kenneth King Stonier Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *jinchilv@marshall.usc.edu*). This work was supported by NSF Grants DMS-1953356, EF-2125142, DMS-2310981, and DMS-2324490, and by Grants 111-2118-M-001-012-MY2 and 109-2118-M-007-MY3 from the National Science and Technology Council, Taiwan.

We also generalize the robust knockoffs inference in [4] to the time series setting to relax the assumption of known covariate distribution required by model-X knockoffs, since such an assumption is overly stringent for time series data. We establish sufficient conditions under which TSKI achieves the asymptotic false discovery rate (FDR) control. Our technical analysis reveals the effects of serial dependence and unknown covariate distribution on the FDR control. We conduct a power analysis of TSKI using the Lasso coefficient difference knockoff statistic under the generalized linear time series models. The finite-sample performance of TSKI is illustrated with several simulation examples and an economic inflation study.

Running title: TSKI

Key words: Model-X knockoffs; Time series; High dimensionality; FDR control; Power analysis; Sparsity; Interpretable forecasting; E-values

1 Introduction

Identifying key economic factors among a large number of potential variables (e.g., numerous types of consumer price indices, unemployment rates, and housing prices) that can influence inflation is a long-standing research pursuit [25, 36, 15] that remains crucial due to inflation's significance. However, statistical inference for economic time series such as inflation is challenging due to the serial dependence, large number of potentially important covariates (including time series covariates, their lags, and non-time series covariates), regime shifts [21, 39], and possible nonlinear relationships.

Let us exemplify these challenges with Figure 1, which depicts the monthly inflation rate (hereafter referred to as inflation) time series of the U.S. economy from May 2013 to January 2023, sourced from the FRED-MD database [28] and the U.S. Bureau of Labor Statistics. In addition to the inflation series, the FRED-MD database includes 126 other time series variables that can be used to predict the inflation for the following month. Here, inflation is calculated as the first-order difference of the Consumer Price Index (CPI) [28] divided by the CPI of the previous month; see Section 5 for details. Figure 1 shows that during the period from 2020 to late 2022, inflation displayed an upward mean drift due to the U.S. economy's post-COVID-19 recovery and the impacts of the Russia–Ukraine conflict, along with some possible stationary phase-changing behaviors [21, 39]. To address concerns on potential nonstationarity over the entire time span caused by, say, the mean drift, the rolling-window method is commonly employed in time series analysis. The intuition behind this method is that the time series, including both the response and predictors, are more likely to be stationary within a small time window[1]. However, the use of rolling windows further complicates the problem in that the sample size in each window is usually small; for example, there are only 60 data points if sampled monthly over a five-year period. The presence of serial dependence, small sample size, high-dimensional covariates, and possibly nonlinear relationships makes statistical inference for time series regression highly challenging. Variable selection has been a popular solution to address such challenges, under the assumption that only a small subset of covariates contribute to the response of interest. Correctly selecting these important covariates can help simplify the model and improve interpretability and prediction accuracy. Our goal in this paper is to develop a reliable variable selection approach that accounts for these unique challenges in time series regression.

Popularly used measures for evaluating the performance of high-dimensional variable selection include the false discovery rate (FDR) [6], model selection consistency [43, 29, 24], and feature importance ranking [10, 30]. Our paper focuses on controlling the FDR, whose formal definition is in (1). Most existing works address the FDR control by building

[1] We also obtain numerical evidence supporting the stationarity of the inflation within each window through standard unit root tests; details are provided in Section A.5.

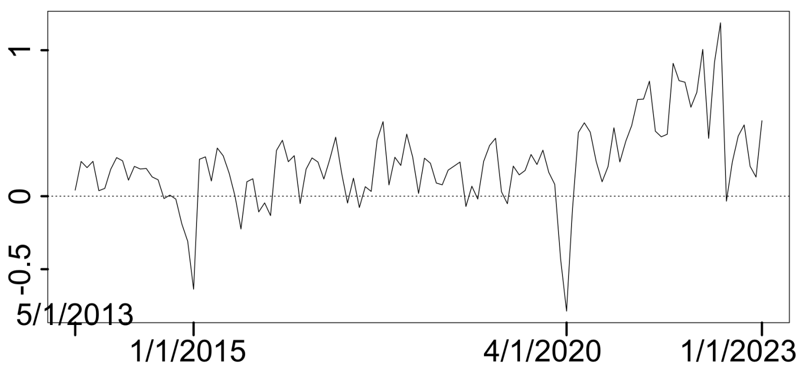


Figure 1: The U.S. inflation from May 2013 to January 2023.

procedures based on the p-values constructed for assessing the importance of individual variables; see, e.g., the seminal works of [6, 7]. Yet, the high dimensionality in covariates and the possibly complicated nonlinear model structure make many conventional ways of p-value calculations inapplicable or even completely fail [19]. To overcome such difficulties, the framework of knockoffs inference was proposed in [3, 12] to achieve the goal of exact FDR control in variable selection in finite samples, completely bypassing the use of the conventional p-values in high-dimensional regression models. It allows for an arbitrary dependence structure of the response on covariates and an arbitrary dimensionality of covariates at the cost of assuming the known joint covariate distribution. See Section 2.2 for a brief review of the model-X knockoffs framework.

Two critical assumptions in model-X knockoffs [12] are: 1) the observations across time are independent and identically distributed (i.i.d.) and 2) the joint distribution of covariate vector is known for generating the knockoff variables. Both assumptions are unreasonably strong for time series data. Time series data exhibit serial dependence, and in some applications where covariates are lagged response variables with a stationarity assumption, assuming a known covariate distribution directly reveals the set of important variables, thus invalidating the problem of variable selection. We will address these challenges by relaxing the two aforementioned assumptions.

To relax the first assumption, we adopt the popular idea of subsampling. To relax the second assumption, we generalize the robust knockoffs inference in [4] to the time series setting, where the robust knockoffs inference [4] is a recent innovation that allows for approximate covariate distribution (as opposed to known covariate distribution) developed for i.i.d. data. We name the knockoff variables generated using approximate covariate distribution as *approximate knockoffs* to ease the presentation. Our analyses reveal that with approximate knockoffs generated in a rowwise fashion ignoring the serial dependence, the FDR inflation has an upper bound depending on the Kullback–Leibler (KL) divergence between the distributions of data matrices corresponding to the approximate and exact model-X knockoff variables. Our theoretical results are comparable to Theorem 1 of [4], with the difference that we do not need the i.i.d. observations assumption. Our theory shows that there is generally no guarantee that such KL divergences asymptotically vanish in the existence of serial dependence, suggesting that the corresponding FDR could be uncontrolled. We also show that subsampling [42] can successfully address such difficulty and warrant asymptotically vanishing KL divergence if the subsampling rate is appropriately chosen, provided that the time series are β-mixing. We then apply the robust knockoffs inference to each of the subsampled data, resulting in multiple sets of selected variables. We aggregate these sets via the e-value method [40, 34]. The complete framework is presented in Section 2 and named as the *time series knockoffs inference* (TSKI). We provide a rigorous characterization of how the serial dependence and the accuracy of the approximate knockoffs affect the FDR control and prove that the TSKI procedure can achieve asymptotic FDR control when the subsampling is done appropriately and the approximate knockoffs are accurate enough.

It is well-known that FDR and power are two sides of the same coin. We then turn to the power analysis of TSKI in Section 3. Assuming the generalized linear time series models and some regularity conditions, we show that for TSKI with subsampling and e-value aggregation, the set of selected variables is either empty or enjoys the sure screening property with asymptotic probability one. These results are formally summarized in Theorem 2 and discussed after it.

We test the empirical performance of TSKI on both simulated and real data. Our simulation results in Section 4 demonstrate that the TSKI with subsampling controls the FDR when the data are generated from some popular β-mixing processes. The selection power is generally satisfactory with a sufficient sample size. Furthermore, in Section 5 we apply the TSKI to study the temporal relations between inflation and other macroeconomic time series from the U.S. economy over the past ten years.

1.1 Related work

Established methods such as the BH and BY [6, 7] are commonly used in biology applications with non-time series data. Some new developments such as [33, 8] either assume independent test statistics, require the availability of p-values, or rely on specific model structures, making them unsuitable for time series applications due to incompatible assumptions.

Among these existing methods, BY achieves the FDR control in variable selection based on a set of valid p-values with no requirements on the dependence structure among p-values. Similarly, e-BH [40] uses e-values without requirements on their dependence structure. While these methods offer the potential for valid FDR control for time series data, obtaining valid p-values or e-values remains an unresolved challenge in many applications.

Regularized regression [29] and information criterion-based model selection [24] are also popularly used for selecting important variables in time series regressions. Yet, they often assume some specific model structure (e.g., the linear model) to prove variable/model selection consistency, posing uncertainty about their performance when deployed on real data with nonlinear dependency. In addition, they are not specifically designed for controlling the variable selection error rate, and hence may not be suitable for certain applications when the error rate control is a concern.

1.2 Notation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{R}$ be the underlying probability space and the Borel σ-algebra on the real line $\mathbb{R}$, respectively. We use the boldface for random vectors and matrices, the tilde for the knockoff variables, and the vector notation $\vec{x}$ to denote vectors in the Euclidean space. For random vector $\boldsymbol{x}$, define $\boldsymbol{x}_{-j}$ as the subvector by removing the jth coordinate. We use parentheses for matrix concatenation. For any real sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ means $\limsup_{n\to\infty} \left|\frac{a_n}{b_n}\right| < \infty$, and $a_n = o(b_n)$ means $\limsup_{n\to\infty} \left|\frac{a_n}{b_n}\right| = 0$. We use $\#S$ to denote the cardinality of a given set S. Moreover, a transition kernel is defined as a map $p : (\mathbb{R}^{k_1}, \mathcal{R}^{k_2}) \longrightarrow [0,1]$ for some positive integers k_1 and k_2 satisfying that (i) for each $\mathcal{D} \in \mathcal{R}^{k_2}$, $p(\cdot, \mathcal{D})$ is a measurable function and (ii) for each $x \in \mathbb{R}^{k_1}$, $p(x, \cdot)$ is a probability measure.

2 Robust time series knockoffs inference with TSKI

Given an observed stationary time series $\{Y_t, \boldsymbol{x}_t\}_{t=1}^n$ with $Y_t \in \mathbb{R}$ a scalar response and $\boldsymbol{x}_t \in \mathbb{R}^p$ a high-dimensional covariate vector, we are interested in accurately selecting relevant covariates (i.e., non-null features) in $\boldsymbol{x}_t$, where the definition of the null feature is given below.

Definition 1. *(Null feature) Consider the response Y and the covariate vector $\boldsymbol{x} = (X_1, \cdots, X_p)^T$. Covariate X_j with $j \in \{1, \cdots, p\}$ is said to be null with respect to response Y if and only if $X_j \perp\!\!\!\perp Y | (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p)^T$.*

We denote the set of null features according to Definition 1 above as $\mathcal{H}_0 \subset \{1, \cdots, p\}$

for each t, where $\mathcal{H}_0$ is independent of t because $(Y_t, \boldsymbol{x}_t)$'s have the same distribution due to the stationarity assumption. Our goal is to estimate the important variable set $\mathcal{H}_1 = \{1, \cdots, p\} \backslash \mathcal{H}_0$ using data $\{Y_t, \boldsymbol{x}_t\}_{t=1}^n$. For an estimated set $\widehat{S} \subset \{1, \cdots, p\}$ returned by some algorithm, we measure the accuracy by evaluating the False Discovery Rate (FDR) defined as

$$\mathrm{FDR} := \mathbb{E}\left(\frac{\#(\widehat{S} \cap \mathcal{H}_0)}{(\#\widehat{S}) \vee 1}\right). \tag{1}$$

In this paper, we focus on time series data with serial dependency and high dimensionality where covariate dimensionality p can be much larger than sample size n. We *do not* assume any specific dependence structure of Y_t on $\boldsymbol{x}_t$ other than the one in Definition 1 and that Y_t is $\boldsymbol{x}_{t+1}$-measurable (see Corollary 2) for our FDR analysis. As a result, our proposed method can accommodate an unknown relationship between Y_t and $\boldsymbol{x}_t$ including both linear and nonlinear ones. Our method builds on the recent work of model-X knockoffs [12] and its robust extension [4] proposed for the non-time-series data, where we briefly review the former in the next section to set the stage.

2.1 A brief review of the model-X knockoffs framework

Let $\boldsymbol{y} = (Y_1, \cdots, Y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$ be the response vector and design matrix collecting the n observations. The model-X knockoffs [12] was proposed for the setting where the rows of the augmented matrix $(\boldsymbol{y}, \boldsymbol{X}) \in \mathbb{R}^{n \times (p+1)}$ are i.i.d. random vectors with known distribution for $\boldsymbol{x}_1$. The knockoffs inference aims at estimating $\mathcal{H}_1$ while keeping the FDR (1) under control. To this end, it constructs an $n \times p$ matrix $\widetilde{\boldsymbol{X}}$ in a rowwise fashion independently using the known joint distribution of $\boldsymbol{x}_1$ such that

$$\widetilde{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{y} | \boldsymbol{X} \quad \text{and} \quad (\boldsymbol{X}, \widetilde{\boldsymbol{X}})_{\mathrm{swap}(S)} \overset{d}{=} (\boldsymbol{X}, \widetilde{\boldsymbol{X}}) \tag{2}$$

for each $S \subset \{1, \cdots, p\}$, where swap$(S)$ denotes the swapping operation meaning that for each $j \in S$, columns j and $j + p$ are swapped, and $\overset{d}{=}$ stands for equal in distribution. At a high level, the model-X knockoffs create a "fake" covariate matrix $\widetilde{\boldsymbol{X}}$ which perfectly mimics the "behavior" of the original covariate matrix. By using these fake covariates as controls, the importance of original covariates can be inferred.

As an example, when the covariate distribution for $\boldsymbol{x}_t$ is known to be $N(\mathbf{0}, \Sigma)$ with Σ the $p \times p$ covariance matrix, a valid way to construct the corresponding ideal knockoff vector $\widetilde{\boldsymbol{x}}_t$ is to sample from the conditional multivariate Gaussian distribution

$$\widetilde{\boldsymbol{x}}_t | \boldsymbol{x}_t \overset{d}{\sim} N(\boldsymbol{x}_t - \text{diag}(\vec{s})\Sigma^{-1}\boldsymbol{x}_t, 2\text{diag}(\vec{s}) - \text{diag}(\vec{s})\Sigma^{-1}\text{diag}(\vec{s})), \tag{3}$$

where diag$(\vec{s})$ is a diagonal matrix of tuning parameters with positive diagonal entries. Larger components of $\vec{s}$ imply that the resulting knockoff variables are more independent of the original variables, thereby providing higher power in distinguishing them. For the general covariate distribution, a conditional distribution for generating knockoff variables can also be constructed following the same high-level idea above. Further details about knockoff variable sampling procedure for general distributions can be found in [12, 18].

With the knockoff variable matrix $\widetilde{\boldsymbol{X}}$, the knockoff statistics W_j's, measuring the importance of original covariates, are constructed such that the sign-flip property [3, 12] is satisfied: for each $S \subset \{1, \cdots, p\}$ and each $1 \leq j \leq p$,

$$W_j(\boldsymbol{y}, [\boldsymbol{X}, \widetilde{\boldsymbol{X}}]_{\text{swap}(S)}) = \begin{cases} -W_j(\boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) & \text{if } j \in S, \\ W_j(\boldsymbol{y}, \boldsymbol{X}, \widetilde{\boldsymbol{X}}) & \text{otherwise.} \end{cases} \tag{4}$$

As variable importance measures, high-quality knockoff statistics W_j's should have the desired properties that 1) W_j's have large positive values for $j \in \mathcal{H}_1$ and 2) for null features $j \in \mathcal{H}_0$, W_j's have small magnitude and are symmetric around zero. See [12] for the formal characterization. Examples 1–2 in the next subsection are two important instances of the knockoff statistics that satisfy the sign-flip property.

Model-X knockoffs [12] estimates $\mathcal{H}_1$ as $\widehat{S} = \{1 \leq j \leq p : W_j \geq T\}$ with the so-called knockoff threshold $T = \min\left\{t > 0 : \frac{1+\#\{j:W_j\leq -t\}}{\#\{j:W_j\geq t\}\vee 1} \leq \tau\right\}$, where $\tau \in (0,1)$ is some pre-specified target level for FDR control. It has been shown in [12] that the model-X knockoffs framework achieves FDR control in finite samples with arbitrary dimensionality of $\boldsymbol{x}_t$ and arbitrary (unknown) dependence structure of Y_t on $\boldsymbol{x}_t$.

As discussed in the Introduction, the i.i.d. row assumption for $(\boldsymbol{y}, \boldsymbol{X})$ and the known distribution assumption for $\boldsymbol{x}_t$ are too stringent for time series data. We will develop a new framework for time series knockoffs inference, which relaxes these assumptions. It is important to note that the remaining part of the paper does *not* assume i.i.d. rows in data matrix $(\boldsymbol{y}, \boldsymbol{X})$.

2.2 Outline of the TSKI framework

In this subsection, we provide an outline of the TSKI framework. Some technical details will be presented in the next subsection. TSKI has three key ingredients: subsmapling, robust knockoffs inference on each subsample, and e-value aggregation of the selected sets from different subsamples. Algorithm 1 provides a detailed implementation of TSKI. Our presentation may not strictly adhere to the order of the three ingredients stated above.

To relax the known covariate distribution assumption, we adopt the robust knockoffs framework in [4] and introduce the following Definition 2. In what follows, the knockoff generator, conditional distribution, and transition kernel are examples of regular conditional probability (r.c.p.) in probability theory.

Definition 2 (Knockoff generator). *$\kappa : \mathbb{R}^p \times \mathcal{R}^p \longmapsto \mathbb{R}^p$ is said to be a knockoff generator if 1) $\kappa(\vec{z}, \cdot)$ is a probability measure for each $\vec{z} \in \mathbb{R}^p$ and 2) $\kappa(\cdot, \mathcal{A})$ is a measurable function* *for each $\mathcal{A} \in \mathcal{R}^p$, where $\mathcal{R}$ is the smallest σ-algebra of $\mathbb{R}$ that contains all open sets.*

For each observed covariate vector $\boldsymbol{x}_t$ with $t \in [n] := \{1, \cdots, n\}$, we generate its knockoff vector $\widetilde{\boldsymbol{x}}_t$ from $\kappa(\boldsymbol{x}_t, \cdot)$. When $\boldsymbol{x}_t$ has known distribution $N(\mathbf{0}, \Sigma)$ with known parameters Σ, the conditional Gaussian distribution in (3) is an instance of the knockoff generator which yields *ideal* knockoff variables satisfying (2). If Σ in (3) is unknown and replaced with an estimated version (see (8) for details), the resulting knockoff generator produces only *approximate* knockoff variables that violate the exchangeability condition in (2). Additional examples of the knockoff generator can be found in [12, 18]. To achieve asymptotic FDR control using knockoff variables generated from $\kappa(\boldsymbol{x}, \cdot)$, we need some additional conditions on κ which will be presented in Condition 3 in the next section. Note that because of the rowwise generation of knockoff variables, although $\{\boldsymbol{x}_t\}_{t=1}^n$ have serial dependence across t, $\{\widetilde{\boldsymbol{x}}_t\}_{t=1}^n$ are independent across t conditional on $\{\boldsymbol{x}_t\}_{t=1}^n$; this violates the second property in (2) and thus the FDR control result in [12] or [4] cannot be applied directly. From now on, we work with the inference sample $\{Y_t, \boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t\}_{t=1}^n$.

To overcome such difficulty, we consider subsamples each with index set $H_k = \{k + s(q+1) : s = 0, 1, \cdots, \lfloor \frac{n-k}{q+1} \rfloor\}$ for $k \in \{1, \cdots, q+1\}$ with some integer $q \geq 0$. To simplify the technical presentation, let $\{V_t, \boldsymbol{u}_t, \widetilde{\boldsymbol{u}}_t\}_{t=1}^N$ be a generic subsample that can be any of the $q+1$ subsamples. Denote by $\boldsymbol{v} := (V_1, \cdots, V_N)^T$, $\boldsymbol{U} := (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_N)^T$, and $\widetilde{\boldsymbol{U}} := (\widetilde{\boldsymbol{u}}_1, \cdots, \widetilde{\boldsymbol{u}}_N)^T$. As in the robust knockoffs inference [4], we construct knockoff statistics $W_j(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}})$'s based on $(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}})$ and select the set of important variables using these knockoff statistics following the identical procedure as reviewed in the last subsection. Thus, we end up with $q+1$ sets of selected variables $\{j : W_j^k \geq T^k\}$ with each corresponding to a subsample k. Here, W_j^k's and T^k are the correspondingly constructed knockoff statistics and the knockoff threshold as specified in (5), respectively.

Below are two examples of the knockoff statistics. The random forests model in Example 2 can be replaced with other learning models such as the deep learning model.

Example 1 (Lasso coefficient difference (LCD))**.** *For a given sample* $(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}})$ *and a tuning parameter* $\lambda \geq 0$*, we define* $W_j = W_j(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}|$*, where* $(\widehat{\beta}_1, \cdots, \widehat{\beta}_{2p})^T$ *is given by either the Lasso estimate* $\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{2p}} \left\{ n^{-1} \sum_{t=1}^n (V_t - (\boldsymbol{u}_t^T, \widetilde{\boldsymbol{u}}_t^T)\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{2p} |\beta_j| \right\}$ *with* $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{2p})^T$*, or the generalized linear model (GLM) Lasso estimate defined in* (12)*.*

Example 2 (Random forests mean decrease accuracy (MDA))**.** *For a given sample* $(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}})$*, we define* $W_j = W_j(\boldsymbol{v}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) = N^{-1} \sum_{t=1}^N \left\{ [V_t - \widehat{m}(\boldsymbol{u}_t^{(j)}, \widetilde{\boldsymbol{u}}_t)]^2 - [V_t - \widehat{m}(\boldsymbol{u}_t, \widetilde{\boldsymbol{u}}_t^{(j)})]^2 \right\}$ *for each* $j \in \{1, \cdots, p\}$*, where* $(\boldsymbol{u}_t^{(j)}, \widetilde{\boldsymbol{u}}_t^{(j)}) = [\boldsymbol{u}_t, \widetilde{\boldsymbol{u}}_t]_{\mathsf{swap}(\{j\})}$ *and* $\widehat{m} : \mathbb{R}^{2p} \longmapsto \mathbb{R}$ *is the random forests regression function trained by regressing* V_t*'s on* $(\boldsymbol{u}_t, \widetilde{\boldsymbol{u}}_t)$*'s.*

The LCD uses the linear model as a working model, while the MDA does not assume any explicit model structure. When the underlying true model is nonlinear, the LCD is based on the misspecified model, whereas the MDA is free of such issues. Yet, MDA demands a large sample size, which is a common drawback for all nonparametric regression models. Our simulation study will provide additional insights into the performance of LCD and MDA in various model settings.

When $q > 0$, subsampling yields more than one set of selected variables. Naively taking the intersection or union over these sets would not guarantee FDR control. The TSKI uses the e-BH procedure [40] to overcome such a difficulty; see Step 3 in Algorithm 1 for the calculation of e-values and how it forms the final set of selected variables $\widehat{S}$. The e-BH procedure works similarly to the BH procedure [6] with the difference that e-values [35] are used in place of the p-values. Given a null hypothesis, we call a non-negative random variable E an "e-value" if $\mathbb{E}[E] \leq 1$ under the null. To test a hypothesis at significance level α, we can reject the null hypothesis when $E \geq 1/\alpha$, noting that $\mathbb{P}(E \geq 1/\alpha) \leq \alpha \mathbb{E}[E] \leq \alpha$ under the null. One appealing property is that the average of multiple e-values is still a valid e-value regardless of their dependence structure. This motivates us to use it to aggregate results from different subsamples. We acknowledge that the e-value idea has been used in the literature for aggregating knockoffs inference results [40, 34]. We note that tuning parameter τ_1 used for individual subsample in Step 3 should be set smaller than the overall target level τ^* in Step 4 in order to achieve high selection power; this is formally stated in Theorem 2 in Section 3.

Algorithm 1: Robust time series knockoffs inference (TSKI) via e-values

```
1 Let $0 < \tau_1 < 1$ be a constant and $0 < \tau^* < 1$ the target FDR level.
2 For each $k \in \{1, \cdots, q+1\}$, calculate the knockoff statistics $W_1^k, \cdots, W_p^k$
    satisfying (4) using sample $\{\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i, Y_i\}_{i \in H_k}$.
3 Calculate the e-value statistics $e_j = (q+1)^{-1} \sum_{k=1}^{q+1} e_j^k$, where[2]
    $$e_j^k = \frac{p \times \mathbf{1}_{\{W_j^k \geq T^k\}}}{1 + \sum_{s=1}^{p} \mathbf{1}_{\{W_s^k \leq -T^k\}}}, \quad T^k = \min\left\{ t \in \mathcal{W}_+^k : \frac{1 + \#\{j : W_j^k \leq -t\}}{\#\{j : W_j^k \geq t\} \vee 1} \leq \tau_1 \right\}, \quad (5)$$
    and $\mathcal{W}_+^k = \{|W_s^k| : |W_s^k| > 0\}$ for each $k \in \{1, \cdots, q+1\}$. Here, $\mathbf{1}_{\{\cdot\}}$ is the
    indicator function.
4 Let $\widehat{S} = \{j : e_j \geq p(\tau^* \times \widehat{k})^{-1}\}$ with $\widehat{k} = \max\{k : e_{(k)} \geq p(\tau^* \times k)^{-1}\}$, where $e_{(j)}$'s
    are the ordered statistics of $e_j$'s such that $e_{(1)} \geq \cdots \geq e_{(p)}$.
```

2.3 FDR control by TSKI

We need some technical conditions to bound the FDR of TSKI.

Condition 1. *The density function of $(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{Y})$ exists and $(Y_t, \boldsymbol{x}_t)$'s are identically distributed across t. In addition, the supports of $(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{Y})$ and $[\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{Y}]_{\mathrm{swap}(\{j\})}$ are the same for each $j \in \{1, \cdots, p\}$.*

Condition 2. *The knockoff generator $\kappa(\cdot, \cdot)$ is constructed independently of observed time series $\{\boldsymbol{x}_t, Y_t\}_{t=1}^n$.*

Condition 1 is a basic regularity condition. Condition 2 may be relaxed if we use sample splitting to obtain an asymptotically independent training subsample for estimating the unknown covariate distribution and constructing the knockoff generator.

To ease the presentation, let $(Y, \boldsymbol{x}, \widetilde{\boldsymbol{x}})$ be an independent copy of $(Y_1, \boldsymbol{x}_1, \widetilde{\boldsymbol{x}}_1)$. As outlined in the last subsection, we allow certain deviation of $\kappa(\boldsymbol{x}, \cdot)$ from the one derived from the true covariate distribution of $\boldsymbol{x}$. Consequently, $\widetilde{\boldsymbol{x}}$ generated from $\kappa(\boldsymbol{x}, \cdot)$ is only an *approximate* knockoff vector, which may violate the exchangeability condition in (2). The use of approximate knockoffs instead of ideal model-X knockoffs may incur a potential FDR inflation. Our Theorem 1, which imposes mild conditions on $\kappa(\cdot, \cdot)$, demonstrates that such FDR inflation can be controlled in time series applications. These results generalize the robust knockoffs inference results for i.i.d. observations in [4] to the time series setting.

Let $\{\boldsymbol{x}_t^\pi, \widetilde{\boldsymbol{x}}_t^\pi, Y_t^\pi\}_{t=1}^n$ be a sequence of i.i.d. random vectors such that $(\boldsymbol{x}_1^\pi, \widetilde{\boldsymbol{x}}_1^\pi, Y_1^\pi)$ and $(\boldsymbol{x}_1, \widetilde{\boldsymbol{x}}_1, Y_1)$ have the same distribution. Denote by $\mathcal{X}_k = \{\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i, Y_i\}_{i \in H_k}$ and $\mathcal{X}_k^\pi = \{\boldsymbol{x}_i^\pi, \widetilde{\boldsymbol{x}}_i^\pi, Y_i^\pi\}_{i \in H_k}$ for each $k \in \{1, \cdots, q+1\}$. Let $f_{\boldsymbol{z}}(\cdot)$ be the density function of random vector $\boldsymbol{z}$.

Theorem 1. *Let $\widehat{S}$ be the set of variables selected by TSKI with Algorithm 1. Then under Conditions 1–2 and the assumption of positive T^k's in (5), we have*

$$\mathrm{FDR} \le \inf_{\varepsilon > 0} \Big[\tau^* e^{\varepsilon} + \sum_{k=1}^{q+1} \mathbb{P}(\max_{1 \le j \le p} \widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon)\Big] + \sum_{k=1}^{q+1} \sup_{\mathcal{D} \in \mathcal{R}^{\# H_k \times (2p+1)}} |\mathbb{P}(\mathcal{X}_k \in \mathcal{D}) - \mathbb{P}(\mathcal{X}_k^\pi \in \mathcal{D})|, \quad (6)$$

where $0 < \tau^* < 1$ is the target FDR level and for each $1 \le k \le q+1$ and $1 \le j \le p$,

$$\widehat{\mathrm{KL}}_j^{k\pi} = \sum_{i \in H_k} \log \left(\frac{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(X_{ij}^\pi, \widetilde{X}_{ij}^\pi, \boldsymbol{x}_{-ij}^\pi, \widetilde{\boldsymbol{x}}_{-ij}^\pi)}{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(\widetilde{X}_{ij}^\pi, X_{ij}^\pi, \boldsymbol{x}_{-ij}^\pi, \widetilde{\boldsymbol{x}}_{-ij}^\pi)} \right). \quad (7)$$

An example is provided in Section A.1 of the Supplementary Material, where the KL divergence term on the right-hand side of (6) asymptotically vanishes. The KL divergence term can be further simplified provided that Condition 3 below, adapted from Definition 1 in [4], is satisfied. This condition concerns the knockoff generator $\kappa(\cdot, \cdot)$ and p additional coordinate-wise knockoff generators $\kappa_j : \mathbb{R}^{p-1} \times \mathcal{R} \mapsto \mathbb{R}$ for $j \in \{1, \cdots, p\}$, where each

$\kappa_j(\boldsymbol{x}_{-j}, \cdot)$ approximates the conditional distribution of X_j given $\boldsymbol{x}_{-j}$.

Condition 3 (Definition 1 in [4]). *For each* $1 \leq j \leq p$, *1) if* $\widetilde{\boldsymbol{z}} = (\widetilde{Z}_1, \cdots, \widetilde{Z}_p)^T$ *is sampled from the conditional distribution* $\kappa((X_1, \cdots, X_{j-1}, \widetilde{X}_j^\dagger, X_{j+1}, \cdots, X_p), \cdot)$, *then* $(\widetilde{X}_j^\dagger, \widetilde{Z}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ *and* $(\widetilde{Z}_j, \widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ *have the same distribution, where* $\widetilde{X}_j^\dagger$ *is sampled from* $\kappa_j(\boldsymbol{x}_{-j}, \cdot)$; *and 2) the density function of the distribution of* $(\widetilde{X}_j^\dagger, \widetilde{Z}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ *exists.*

Corollary 1. *Assume that all the conditions of Theorem 1 hold. If further Condition 3 is satisfied, then (6) holds with*

$$\widehat{\mathrm{KL}}_j^{k\pi} = \sum_{i \in H_k} \log \left(\frac{f_{X_j|\boldsymbol{x}_{-j}}(X_{ij}^\pi | \boldsymbol{x}_{-ij}^\pi) f_{\widetilde{X}_j^\dagger|\boldsymbol{x}_{-j}}(\widetilde{X}_{ij}^\pi | \boldsymbol{x}_{-ij}^\pi)}{f_{X_j|\boldsymbol{x}_{-j}}(\widetilde{X}_{ij}^\pi | \boldsymbol{x}_{-ij}^\pi) f_{\widetilde{X}_j^\dagger|\boldsymbol{x}_{-j}}(X_{ij}^\pi | \boldsymbol{x}_{-ij}^\pi)} \right),$$

where $\widetilde{X}_j^\dagger$*'s are given in Condition 3 and* $f_{\boldsymbol{z}_1|\boldsymbol{z}_2}(\boldsymbol{z}_1|\boldsymbol{z}_2) = f_{\boldsymbol{z}_1,\boldsymbol{z}_2}(\boldsymbol{z}_1, \boldsymbol{z}_2)[f_{\boldsymbol{z}_2}(\boldsymbol{z}_2)]^{-1}$ *is the conditional probability density function of* $\boldsymbol{z}_1$ *given* $\boldsymbol{z}_2$.

It is seen that when Condition 3 is met, the FDR inflation can be measured by the Kullback–Leibler (KL) divergence between the conditional distributions of $\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}$ and $X_j | \boldsymbol{x}_{-j}$, where the former is described by the coordinate-wise knockoff generator $\kappa_j(\boldsymbol{x}_{-j}, \cdot)$. When these two conditional distributions are identical, the knockoff generator $\kappa(\cdot, \cdot)$ satisfying Condition 3 reduces to the ideal knockoff generator in model-X knockoffs [12]. In this sense, Condition 3 relaxes the requirement of knowing the exact covariate distribution.

We borrow the Gaussian example discussed in [4] to help understand Condition 3. Consider the Gaussian knockoff generator described around (8). Lemma 4 of [4] demonstrates that $\kappa(\cdot, \cdot)$ and $\kappa_j(\cdot, \cdot)$ satisfy Condition 3 if the former is chosen as

$$\kappa(\boldsymbol{x}_t, \cdot) \overset{d}{\sim} N(\boldsymbol{x}_t - \mathrm{diag}(\vec{s})\widehat{\Theta}\boldsymbol{x}_t, 2\mathrm{diag}(\vec{s}) - \mathrm{diag}(\vec{s})\widehat{\Theta}\mathrm{diag}(\vec{s})), \tag{8}$$

and each $\kappa_j(\boldsymbol{x}_{-j}, \cdot)$ follows the distribution $N(\boldsymbol{x}_{-j}^T \widehat{\Theta}_{-j,j} \widehat{\Theta}_{jj}^{-1}, \widehat{\Theta}_{jj}^{-1})$, where $\widehat{\Theta}$ is a postive definite estimate of $[\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)]^{-1}$ constructed independently from the observed data $(\boldsymbol{y}, \boldsymbol{X})$,

$\widehat{\Theta}_{-j,j}$ represents the jth column of $\widehat{\Theta}$ with the jth component removed, and $\widehat{\Theta}_{jj}$ is the jth diagonal entry of $\widehat{\Theta}$. Here, for simplicity, we assume that $\boldsymbol{x}_t$ has mean zero. Corollary 1 indicates that the asymptotic FDR control depends on the estimation accuracy of $\widehat{\Theta}$. See Section A.4 for how we estimate the precision matrix in practice. Meanwhile, see [4] for other examples of knockoffs generators satisfying Condition 3.

Under Condition 4 below, a simple upper bound for the second term on the right-hand side of (6) can be derived; this result is formally summarized in Corollary 2 below. An example satisfying Condition 4 is provided in Section 2.4.

Condition 4 (h-step β-mixing with exponential decay)**.** *The covariate process $\{\boldsymbol{x}_t\}$ is a p-dimensional stationary Markov chain with a transition kernel $p : \mathbb{R}^p \times \mathcal{R}^p \longmapsto \mathbb{R}$ and a stationary distribution π. There exist a positive integer h, a measurable function $V : \mathbb{R}^p \longrightarrow [0, \infty)$, and some constants $0 \leq \rho < 1$ and $0 < C_0 < \infty$ such that for each $\vec{x} \in \mathbb{R}^p$, $\left\| p^h(\vec{x}, \cdot) - \pi(\cdot) \right\|_{TV} \leq C\rho^h V(\vec{x})$, where $C > 0$ is some constant with $C_0 \geq C \int_{\mathbb{R}^p} V(\vec{x}) d\pi(\vec{x})$ and $\|\cdot\|_{TV}$ denotes the total variation (TV) norm associated with measures. Moreover, for each $\vec{x} \in \mathbb{R}^p$, $p(\vec{x}, \cdot)$ is absolutely continuous with respect to the Lebesgue measure.*

Corollary 2. *Assume that all the conditions of Theorem 1 hold. If further $\{\boldsymbol{x}_t\}_{i \geq 1}$ satisfies Condition 4 with q-step and constants $C_0 > 0$ and $0 \leq \rho < 1$, and Y_i is $\boldsymbol{x}_{t+1}$-measurable, then (6) holds with*

$$\sum_{k=1}^{q+1} \sup_{\mathcal{D} \in \mathcal{R}^{\# H_k \times (2p+1)}} \left| \mathbb{P}(\mathcal{X}_k \in \mathcal{D}) - \mathbb{P}(\mathcal{X}_k^{\pi} \in \mathcal{D}) \right| \leq C_0 \times \rho^q \times n. \tag{9}$$

Moreover, when $(Y_t, \boldsymbol{x}_t)$'s are i.i.d., (9) holds with $\rho = 0$.

As shown in Corollary 2, in the i.i.d. data setting where $\rho = 0$, the FDR upper bound in (6) replicates the result in [4] when $q = 0$ (i.e., no subsampling). With serial dependence and general q, from (6) and (9), we observe an interesting tradeoff between the KL divergence and the TV upper bound as q changes. First, in view of the expression in Corollary 1, $\mathbb{P}(\max_{1\le j\le p}\widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon)$ depends on q in a complicated way via the subsample size $\lfloor n/(q+1)\rfloor$. In addition, since the probability is upper bounded by 1, the term $\sum_{k=1}^{q+1}\mathbb{P}(\max_{1\le j\le p}\widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon)$ increases at most linearly with q. On the other hand, Corollary 2 suggests that the TV upper bound decreases exponentially with q. Despite this tradeoff, it is unreasonable to determine the optimal choice of q by directly analyzing these upper bounds since they can be conservative. We leave the theoretical investigation of the optimal q for future research. Our empirical study suggests that $q = 1$ or 2 often yields satisfactory finite-sample control of the FDR, as long as each subsample is reasonably large.

To provide a concrete time series example where our theory in this section applies, we analyze the Gaussian ARX model (Example 3 in the next subsection) with $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)$ assumed to be known, and $q = \lceil(\log n)^{1+\delta}\rceil$ and $p = O(n^{K_0})$ for some constants $\delta > 0$ and $K_0 > 0$. We show in Section A.1 that the KL divergence is zero and prove in Section 2.4 that Condition 4 is satisfied. Thus, the FDR upper bound in Theorem 1 becomes $\tau^* + C_0 n\rho^q$. When Σ is unknown and needs to be estimated, it is possible to obtain an upper bound using an estimated covariance matrix formed from a large independent learning sample using the proof idea for Lemma 5 of [4].

When moving beyond the Gaussian time series covariates, for example, Model 1 in Section 4, we assume the high-level condition $\sum_{k=1}^{q+1}\mathbb{P}(\max_{1\le j\le p}\widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon) = o(1)$ and the β-mixing condition as in Corollary 2. Under these assumptions, our theory applies. Despite these technical assumptions, our simulation results suggest that the FDR can be controlled in broad model settings with in-sample estimated covariate distribution. Theoretical justification of such empirical results for time series data is highly challenging and is left for future investigation.

2.4 Stationary processes satisfying Condition 4

We provide a linear time series example that satisfies Condition 4. Additional examples, including nonlinear ones, are provided in Section A.2 of the Supplementary Material.

Example 3 (Autoregressive models with exogenous variables (ARX))**.** *For each t, define*

$$Y_t = \sum_{j=1}^{k_1} \alpha_j Y_{t-j} + \sum_{l=1}^{k_2} \sum_{j=1}^{k_{3l}} \beta_j^{(l)} H_{t-j+1}^{(l)} + \varepsilon_t$$

with $H_t^{(l)} = \epsilon_t^{(l)} + \sum_{j=1}^{k_{3l}} b_j^{(l)} H_{t-j}^{(l)}$, where for some positive constants $L_0, L_1,$ and L_2, it satisfies that $1 - \sum_{j=1}^{k_{3l}} b_j^{(l)} z^j \neq 0$ and $1 - \sum_{j=1}^{k_1} \alpha_j z^j \neq 0$ for each $|z| \leq 1 + L_0$ and each $l \in \{1, \cdots, k_2\}$. Additionally, $\sum_{l=1}^{k_2} \sum_{j=1}^{k_{3l}} |\beta_j^{(l)}| < L_1$ and $k_1, k_{31}, \cdots, k_{3k_2} < L_2$. Here, k_1, k_2, and $k_{31}, \cdots, k_{3k_2}$ are all positive integers. Moreover, $(\epsilon_t^{(1)}, \cdots, \epsilon_t^{(k_2)}, \varepsilon_t)$'s are $(k_2 + 1)$-dimensional i.i.d. Gaussian random vectors with zero mean and positive definite covariance matrix Σ_0 that satisfy $\mathbb{E}(\varepsilon_t \epsilon_t^{(l)}) = 0$ for each l.

Example 3 is a benchmark model for describing the behavior of macroeconomic variables. It is seen that the covariate vector with respect to response Y_t is $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-k_1}, \boldsymbol{h}_t^T)^T$ with $\boldsymbol{h}_t = (H_t^{(1)}, \cdots, H_{t-k_{31}+1}^{(1)}, \cdots, H_t^{(k_2)}, \cdots, H_{t-k_{3k_2}+1}^{(k_2)})^T$. It has been shown (e.g., [2]) that the stationary process $\{\boldsymbol{x}_t\}$ in Example 3 with fixed dimensionality satisfies Condition 4 with h-step and some constants ρ, C_0 for each positive integer h. When the dimensionality of $\boldsymbol{x}_t$ increases, certain growth conditions such as (10) below on the value of h and the dimensionality of $\boldsymbol{x}_t$ are needed for Condition 4 to hold. For this reason, we make the dependence of the stationary processes on h explicit: hereafter $\{\boldsymbol{x}_t^{(h)}\}$ denotes a p_h-dimensional stationary process satisfying Condition 4 for $h \geq 1$, as guaranteed by Proposition 1 below.

Proposition 1. *Let $\{\boldsymbol{x}_t^{(h)}\}$ be a sequence of p_h-dimensional linear process in Example 3 with constant L_i's and uniformly positive definite $\Sigma_0^{(h)}$'s. Assume that for some constant*

$C_2 > 0$ and sufficiently small $s_2 > 0$,

$$\sup_{h>0}\{p_h \exp(-s_2 h)\} \le C_2. \tag{10}$$

Then for some constants $0 \le \rho < 1$ and $C_0 > 0$ and all large h, $\{\boldsymbol{x}_t^{(h)}\}$ satisfies Condition 4.

In view of Theorem 1 and Proposition 1, if we assume Example 3 with $p = p_{q_n} = O(n^{K_0})$ and subsample with $q = q_n = \lceil(\log n)^{1+\delta}\rceil$, where $K_0 > 0$ and $\delta > 0$ are some constants, then the β-mixing convergence rate required by Theorem 1 is satisfied by the subsampled data for all large n.

3 Power analysis under generalized linear time series models

Since the selection power of any procedure depends on the signal strength, we showcase the power analysis using the GLM time series model where the signal strength can be measured conveniently by the regression coefficients. Correspondingly, we consider the LCD knockoff statistic in Example 1 because Lasso is popularly used in high-dimensional GLM regression.

The canonical GLM has the conditional mean function

$$\mathbb{E}(Y_t|\boldsymbol{x}_t) = g(\boldsymbol{x}_t^T\vec{\beta}^o), \tag{11}$$

where $\vec{\beta}^o$ is a vector of unknown coefficients and $g(\cdot)$ is the derivative of a differentiable function $r(\cdot)$. The inverse function of $r'(\cdot)$, denoted as $g^{-1}(\cdot)$, is referred to as the canonical link function. Commonly used canonical link functions include: (1) the identity link $g^{-1}(\mu) = \mu$ for the linear model, (2) the logit link $g^{-1}(\mu) = \log(\mu/(1-\mu))$ with $0 < \mu < 1$ for the logistic model, and (3) the log link $g^{-1}(\mu) = \log\mu$ with $\mu > 0$ for the Poisson model.

To form the LCD knockoff statistics, we first fit the following GLM Lasso regression

$$(\widehat{\beta}_1, \cdots, \widehat{\beta}_{2p})^T = \arg \min_{\vec{\beta} \in \mathbb{R}^{2p}} \left\{ n^{-1} \sum_{t=1}^{n} 2 \left(-Y_t \times (\boldsymbol{x}_t^T, \widetilde{\boldsymbol{x}}_t^T) \vec{\beta} + r \left((\boldsymbol{x}_t^T, \widetilde{\boldsymbol{x}}_t^T) \vec{\beta} \right) \right) + \lambda_n \sum_{j=1}^{2p} |\beta_j| \right\}, \quad (12)$$

where $\lambda_n \geq 0$ is the regularization parameter. Define $\vec{\beta}^* := (\beta_1^*, \cdots, \beta_{2p}^*)^T$ with $(\beta_1^*, \ldots, \beta_p^*)^T = \vec{\beta}^o$ and $\beta_{p+1}^* = \cdots = \beta_{2p}^* = 0$. By the conditional independence $Y_t \perp\!\!\!\perp \widetilde{\boldsymbol{x}}_t | \boldsymbol{x}_t$, we have $\mathbb{E}(Y_t | \boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t) = \mathbb{E}(Y_t | \boldsymbol{x}_t) = g(\boldsymbol{x}_t^T \vec{\beta}^o) = g((\boldsymbol{x}_t^T, \widetilde{\boldsymbol{x}}_t^T) \vec{\beta}^*)$. Thus, (12) estimates the population parameter vector $\vec{\beta}^*$. When the link function is the identity, the GLM Lasso estimate becomes the linear Lasso estimate.

Extensive research has been conducted to study the performance of Lasso with time series data. See, for example, [1, 5, 22, 26, 29, 41]. Specifically, error bounds for the linear Lasso estimates in ARX models with conditionally heteroskedastic errors have been considered in [1, 29, 41]. Moreover, [22] established error bounds and support recovery guarantees of the GLM Lasso when both the response and covariate vector are stationary time series with dependence measures (as defined therein) satisfying certain summability conditions. Our power analysis needs the following technical conditions.

Condition 5. *For some constant $c_0 > 0$ and sequence $k_{3n} > 0$ with $\lim_{n \to \infty} k_{3n} = 0$, it holds that $\mathbb{P}\left(\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0 (\# S^*) \lambda_n \right) \geq 1 - k_{3n}$, where $S^* = \{ j : |\beta_j^*| > 0 \}$.*

Condition 6. *There exists some sequence $k_{1n} > 0$ with $k_{1n} q^{-1} \to \infty$ as $n \to \infty$ such that $\min_{j \in S^*} |\beta_j^*| > k_{1n} \lambda_n$.*

Condition 7. *For some constant $c_1 \in (0, 1)$ and sequence $\{k_{2n}\}$ with $\lim_{n \to \infty} k_{2n} = 0$, it holds that $2(\tau_1 \# S^*)^{-1} < c_1$ and $\mathbb{P}(\#\{j : W_j^k \geq T^k\} \geq c_1(\# S^*)) \geq 1 - k_{2n}$ for $k \in \{1, \cdots, q+1\}$.*

We note that Condition 5 and (13) below are the L_1- and L_2-estimation error bounds for the linear Lasso, which have been established in previous studies [1, 5, 26, 29, 41] in time series settings under various different model assumptions with model-specific choice of λ_n. In these studies, k_{1n}, k_{2n}, and k_{3n} typically decrease at polynomial rates as n increases. Additionally, the GLM Lasso estimation error bounds have been established in [22]. As a concrete example, for the model in Example 3 with dimensionality p increasing at most exponentially with sample size n, we can prove that Condition 5 and (13) hold with $\lambda_n = O([\log p]^{3/2} n^{-1/2+\alpha})$ for some $\alpha \in (0, 1/2)$ following the same proof as that for Theorem 3.1 in [22] if, say, the Gaussian knockoff generator (3) is used. Since the proof is a straightforward extension, we opt to omit it to save space.

Condition 6 shares similarities with the often-imposed beta-min assumption $\min_{j \in S^*} |\beta_j^*| > \sqrt{\#S^*} \times \lambda_n$ for ensuring Lasso's model selection consistency. If q is chosen as $O((\log n)^{1+\delta})$ as suggested by Proposition 1 and $\sqrt{\#S^*} \gg (\log n)^{1+\delta}$, then Condition 6 becomes less restrictive than the beta-min condition. Condition 7 is a technical assumption that can be proved using the same derivations as in [18] under Condition 5 and the following L_2-estimation error bound condition

$$\mathbb{P}\Big(\Big[\sum_{j=1}^{2p} (\widehat{\beta}_j - \beta_j^*)^2 \Big]^{\frac{1}{2}} \le c_0 \sqrt{\#S^*} \lambda_n \Big) \ge 1 - k_{2n}. \tag{13}$$

For technical simplicity in power analysis, we assume that there are no ties in the magnitude of nonzero knockoff statistics and Lasso solutions. Furthermore, we assume that T^k's in (5) satisfy $\max_{1 \le k \le q+1} \{T^k\} < \infty$ almost surely.

Theorem 2. *Assume $S^* \subset \{1, \cdots, p\}$ with $\#S^* > 0$. Let $\widehat{S}$ be returned by Algorithm 1 with $\tau_1, \tau_* \in (0, 1)$ and the LCD knockoff statistics (Cf. Example 1). Assume that Conditions 6–7 are satisfied and Condition 5 holds for the Lasso estimates applied to each subsample H_k in Algorithm 1. Then it holds that for all large n, small*

$$\mathbb{P}\left(\{\widehat{S} = \emptyset\} \cup \left\{ \frac{\#(S^* \cap \widehat{S})}{\#S^*} \ge 1 - 4c_0(1+q)k_{1n}^{-1} \right\} \right) \ge 1 - (q+1) \times (k_{2n} + k_{3n}). \tag{14}$$

If we further assume that $\tau_1 = \tau^* \times K^{-1} \times (1 - 4(q+1)c_0 k_{1n}^{-1})$ *with some* $K > 1$*, then for all large* n*,*

$$\mathbb{E}\left(\frac{\#(S^* \cap \widehat{S})}{\#S^*}\right) \geq \left(1 - \frac{(q+1)(\tau_1 + \theta_\varepsilon)K}{K-1} - (q+1) \times (k_{2n} + k_{3n})\right) \times k_{4n}, \tag{15}$$

where $\{k_{4n}\}$ *is some increasing sequence with* $\lim_{n\to\infty} k_{4n} = 1$ *and*

$$\theta_\varepsilon = \inf\left\{\theta \geq 0 : \max_{1\leq k\leq q+1} \mathbb{E}\left(\frac{\#(\{j : W_j^k \geq T^k\} \cap (S^*)^c)}{\#\{j : W_j^k \geq T^k\} \vee 1}\right) \leq \tau_1 + \theta\right\}. \tag{16}$$

As previously discussed, k_{2n} and k_{3n} generally converge to zero at polynomial rates as n increases. Hence, setting $q = O((\log n)^{1+\delta})$ makes the right-hand side of (14) asymptotically approaching one. Thus, with asymptotic probability one, Algorithm 1 either makes no discovery or has the percentage of correct discovery $\frac{\#(S^*\cap\widehat{S})}{\#S^*}$ approaching one. The event of no discovery occurs when most individual knockoff filters from different subsamples select an abundantly large number of false positives so that the resulting e-values for all variables become too small to be selected. To further exclude the event $\{\widehat{S} = \emptyset\}$, it is essential for $(\tau_1 + \theta_\varepsilon)q$ to be asymptotically vanishing (see (15)). To understand this, note that when $(\tau_1 + \theta_\varepsilon)q$ is sufficiently small, it effectively controls the false discovery proportion for each knockoff filter, consequently providing an upper bound on $\sum_{s=1}^p \mathbf{1}_{\{W_s^k \leq -T^k\}}$ for each $k = 1, \cdots, q+1$. This further entails that the jth e-value statistic given in Algorithm 1 is sufficiently large for each $j \in S^*$ so that it can be selected by the e-value procedure and hence $\widehat{S} \neq \emptyset$. In practical implementation, in light of the definition of τ_1, we can make $\tau_1 q$ asymptotically negligible by choosing $K \gg q = O((\log n)^{1+\delta})$.

4 Simulation studies

We now investigate the finite-sample performance of the TSKI procedure with Algorithm 1. Two selection methods considered in this section are TSKI with LCD in Example 1 and TSKI with random forests MDA in Example 2; we abbreviate them as TSKI-LCD and TSKI-MDA, respectively. We generate the approximate knockoff variables using the idea of the second-order approximation [12, 18] via the knockoff generator (8); See Section A.4 for the estimation of $\widehat{\Theta}$ and the choice of $\vec{s}$.

We compare with two benchmark methods. The first method selects features using the Benjamini–Yekutieli (BY) method [7]. We choose not to use the Benjamini–Hochberg (BH) method [6] because it assumes independence among the underlying p-values, which is not expected to hold true for time series applications. Due to the limited results on obtaining p-values for high-dimensional time series regression models, for the BY method, we calculated p-values by utilizing the ordinary least squares method in our simulations when $n > p$. When $n < p$, the BY method is not applicable due to the lack of an effective p-value calculation method. The second benchmark method is the adaptive Lasso [44, 29], which was designed for feature selection without the aim of FDR control. These two approaches are abbreviated as LS-BY and adaLasso, respectively.

4.1 Simulation settings

We consider the self-exciting threshold autoregressive model with exogenous variables (SETARX [39]), which is one of the most commonly used models for regime changes in time series data. Regime changes are frequently observed in economic time series (e.g., time series in Figure 1). It is well-known that SETARX with fixed dimensionality p satisfies the β-mixing Condition 4 and is stationary; see, e.g., [2]. We present additional simulation experiments in Section A.4 of the Supplementary Material, where some other popular time series models such as the linear autoregressive model with exogenous variables and the autoregressive conditional heteroskedasticity model with exogenous variables [17] are considered.

Model 1 (SETARX model)**.** *For each integer t and $\iota \in \{0, 5\}$, we define*

$$Y_t = \begin{cases} \sum_{j=1}^{2}(-0.5)^{j-1}\beta Y_{t-j} + 0.6(\sum_{j=1}^{\iota} H_{t,j} + \sum_{j=\iota+1}^{15} H_{t,j}) + \varepsilon_t, & \text{if } Y_{t-d} > r, \\ \sum_{j=1}^{2} -(-0.5)^{j-1}\beta Y_{t-j} + 0.6(-\sum_{j=1}^{\iota} H_{t,j} + \sum_{j=\iota+1}^{15} H_{t,j}) + \varepsilon_t, & \text{otherwise,} \end{cases}$$

where we choose $r = 0.7$ and $d = 1$ as the threshold value and the threshold lag, respectively.

We set the autoregressive coefficient $\beta = 0.7$, and choose the model errors ε_t's as i.i.d. $N(0, 1)$. The time series covariates $H_{t,j}$'s are generated as $H_{t,j} = \eta \times H_{t-1,j} + \epsilon_{t,j}$ with $j \in \{1, \cdots, 50\}$ and $\eta = 0.2$, where $(\epsilon_{t,1}, \cdots, \epsilon_{t,50})$'s are i.i.d. Gaussian random vectors across t with zero mean and $\mathbb{E}(\epsilon_{t,k}\epsilon_{t,l}) = (0.2)^{|k-l|}$ for all k, l. We formulate the covariate vector with respect to response Y_t as $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-20}, \boldsymbol{h}_t, \boldsymbol{h}_{t-1}, \boldsymbol{h}_{t-2}, \boldsymbol{h}_{t-3}, \boldsymbol{h}_{t-4})$ with $\boldsymbol{h}_t = (H_{t,1}, \cdots, H_{t,50})$, giving rise to $p = 270$. Throughout the simulation, we keep $p = 270$ while varying the sample size n across experiments with $n \in \{200, 300, 500, 1000\}$. Due to the space constraint, we move the results for $n = 300$ and 1000 to Section A.4.

It is seen that the mean function is comprised of two components: a piecewise linear function of the lags and the first ι time series covariates, along with a linear function involving $(H_{t,\iota+1}, \cdots, H_{t,15})$ for $\iota \in \{0, 5\}$. The setup with $\iota = 5$ mimics a realistic and straightforward scenario where certain variables $H_{t,j}$, but not all, together with lagged variables Y_{t-j}, undergo changes in regime. The mean regression function of Model 1 is therefore nonlinear, as illustrated graphically in Figure 2. The relevant index set according to Definition 1 is $S_0 = \{1, 2, 21, \cdots, 35\}$, while the null set $\mathcal{H}_0 = \{3, \cdots, 20, 36, \cdots, 270\}$.

For implementation, the target FDR level is set as $\tau^* = 0.2$, and the R packages `glmnet` and `randomForest` are used for calculating the Lasso estimates and the random forests

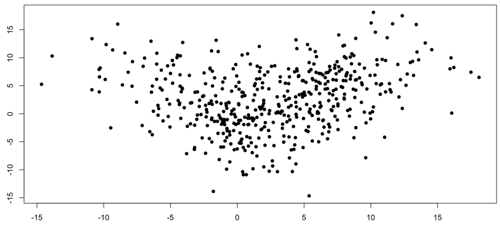


Figure 2: A graphical representation depicting Y_{t-1} on the x-axis and Y_t on the y-axis is provided under Model 1 with $(\eta, \iota) = (0.2, 5)$.

Method	$n/p/\eta/\iota$	q	FDR	Power	$n/p/\eta/\iota$	q	FDR	Power
TSKI-LCD	200/270/0.2/0	0	0.157	0.698	500/270/0.2/0	0	0.176	0.939
TSKI-LCD		1	0.026	0.051		1	0.099	0.872
TSKI-MDA		0	0.173	0.456		0	0.181	0.922
TSKI-MDA		1	0.026	0.028		1	0.092	0.550
TSKI-LCD	200/270/0.2/5	0	0.139	0.287	500/270/0.2/5	0	0.141	0.634
TSKI-LCD		1	0.023	0.019		1	0.086	0.267
TSKI-MDA		0	0.138	0.215		0	0.166	0.679
TSKI-MDA		1	0.012	0.011		1	0.084	0.216

Table 1: The simulation results on the empirical FDR and power for the TSKI with $\tau_1 = \tau^*/(q+1)$ and $\tau^* = 0.2$ under Model 1 in Section 4.1. The results for $n \in \{300, 1000\}$ and $q = 2$ are given in Section A.4

.

MDA, respectively. The TSKI Algorithm 1 with parameters $q \in \{0, 1, 2\}$, $\tau^* = 0.2$, and $\tau_1 = \tau^*/(q+1)$ is used in our simulation. The empirical versions of the FDR and power are calculated as the sample averages of the false discovery proportion and true discovery proportion across 100 repetitions, respectively. The values of (p, q, n) are included in both Tables 1–2, and the R codes are available in the Supplementary Material.

4.2 Empirical performance of TSKI

Table 1 presents the results for TSKI-LCD and TSKI-MDA with $q = 0$ and $q = 1$. The complete results, which also include $q = 2$ and a larger sample size $n = 1000$, are provided in Section A.4. It is seen from Table 1 that both methods control the FDR in finite samples below the target level of $\tau^* = 0.2$, while larger q gives lower selection power compared to that of $q = 0$ (i.e., no subsampling). The lower power for larger q is reasonable and caused by the small sample size in each subsample when fitting the time series regression model.

Adaptive Lasso			LS + BY		
$n/p/\eta/\iota$	FDR	Power	$n/p/\eta/\iota$	FDR	Power
200/270/0.2/0	0.520	0.964	200/270/0.2/0	–	–
300/270/0.2/0	0.468	0.997	300/270/0.2/0	0.000	0.001
500/270/0.2/0	0.657	1.000	500/270/0.2/0	0.027	0.763
200/270/0.2/5	0.604	0.705	200/270/0.2/5	–	–
300/270/0.2/5	0.563	0.786	300/270/0.2/5	0.018	0.006
500/270/0.2/5	0.677	0.891	500/270/0.2/5	0.026	0.276

Table 2: Left panel: the simulation results on the empirical FDR and power for the adaptive Lasso [29, 44]; there is no target FDR level for the adaptive Lasso. Right panel: the simulation results on the empirical FDR and power for the ordinary least squares + Benjamini–Yekutieli (BY [7]) with the target FDR level at 0.2; this approach does not apply to high-dimensional scenarios when $n < p$.

The knockoffs method is empirically known to be conservative when the sample size is overly small. This conservative nature offsets the FDR inflation due to serial dependence in the small sample size scenario, which explains why FDR is still controlled even without subsampling. The extended Model 1 simulation presented in Section A.4 shows that for a larger sample size $n = 1000$, we start observing FDR inflation when $q = 0$. In addition, in the additional simulation examples presented in Section A.4, we observe severe FDR inflation when $q = 0$ in various scenarios (see Table 3).

From Table 1, we see that for Model 1 with $\iota = 0$ (i.e., lower nonlinearity), the LCD-based method demonstrates superior performance over the MDA-based method in terms of power. Meanwhile, when $\iota = 5$ (i.e., higher nonlinearity), we observe from Table 1 that MDA outperforms LCD in power when the sample size is large (i.e., $n = 500$ and $q = 0$), an intuitive observation considering the nonparametric nature of the MDA measure. Indeed, the empirical performance of MDA decreases drastically in all settings as q increases, because of the smaller sample size when calculating the MDA measures.

The results from Table 1 highlight that TSKI-MDA may not be suitable for some real data applications, such as our study where only 60 observations are accessible for each inference window. They underline the need to explore ways to enhance selection power for nonlinear time series, especially with limited samples. On the other hand, Table 2 highlights that adaLasso demonstrates one of the highest selection powers among all four methods. However, adaLasso fails to control error rates under Model 1. This is expected because adaLasso is designed for the linear model and focuses on consistent model selection, lacking an error rate tuning parameter, such as the target FDR level. LS-BY is inapplicable to high-dimensional time series applications when $n < p$, such as our real data example in Section 5, due to the lack of a reliable p-value calculation method. It is important to emphasize that obtaining valid p-values under high-dimensional linear or nonlinear time series models presents a highly challenging and currently unresolved issue; such challenge is also reflected in the poor performance of LS-BY when p becomes comparable to n (i.e., $(n, p) = (300, 270)$).

It is worth pointing out that our additional simulation experiments in Section A.4 of the Supplementary Material reveal that: 1) TSKI-LCD and TSKI-MDA with $q = 1$ consistently control FDR below the target level, whereas setting $q = 0$ yields results occasionally exceeding the target level; and 2) the LS-BY method can have FDR exceeding the target level when data is simulated from autoregressive conditional heteroskedasticity models with exogenous variables and higher value of η (as in Model 1). See the Supplementary Material for full details.

Overall, our simulation experiments show that the time series FDR is controlled stably by the TSKI procedure with subsampling parameter $q = 1$ in finite samples, whereas the selection power depends on the sample size, the subsampling parameter $q \geq 0$, and the choice of the feature importance measure. Importantly, our theoretical results (Theorem 1 and Corollary 2) and simulation results show that TSKI is among the first approaches with theoretically justified FDR control for dependent data under flexible β-mixing condition in Condition 4. For implementation, we suggest that practitioners working on time series inference with limited sample sizes initiate their diagnosis by using TSKI-LCD with q set to 1, and generate knockoff variables using knockoff generator (8).

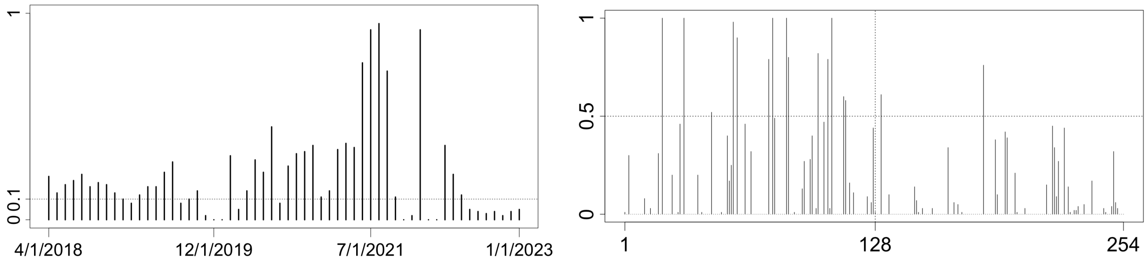


Figure 3: The left panel displays the averages of "having any selections" indicators over 100 repetitions, where the indicator at each rolling window is one if and only if any covariates are selected, and the x-axis indicates the ending time of each rolling window. The right panel shows the analogous results, but the indicator is one at each covariate index if that covariate is selected at any rolling window. The first 127 covariates are current time covariates, and the 128th to 254th covariates are one-month lag covariates in the AR(2) model. Covariates measuring similar economic values are clustered closer (see [28] for detailed definitions of these covariates). The selection method here is the TSKI-LCD without subsampling ($q = 0$).

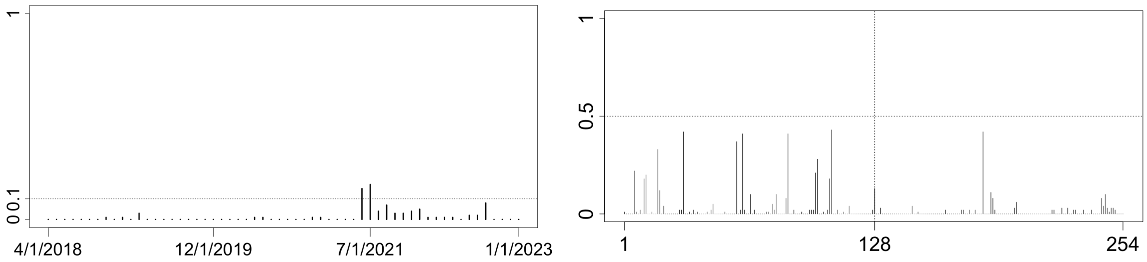
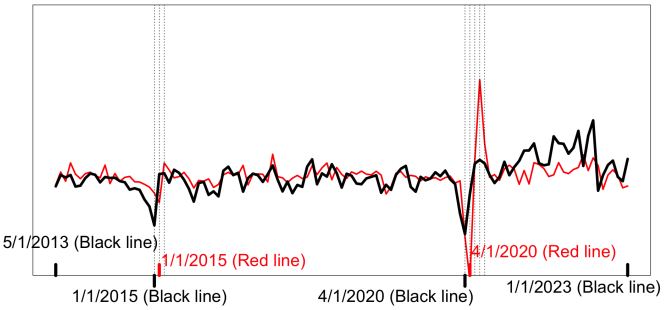


Figure 4: These two panels are analogous to those in Figure 3 but with $q = 1$ for the TSKI-LCD procedure.

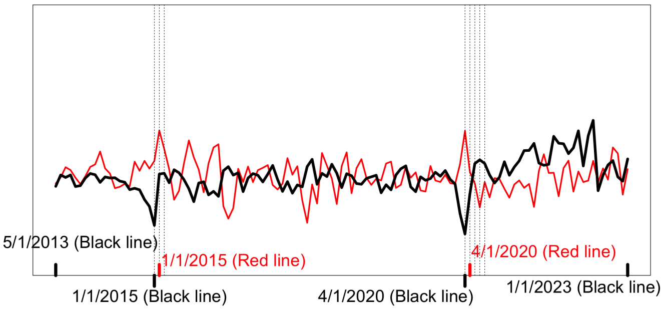
5 Real data application

We analyze the U.S. inflation series data described in the Introduction. The monthly economic time series, including numerous types of consumer price indices, unemployment rates, and housing prices, can be obtained from the FRED-MD database [28][3] and the U.S. Bureau of Labor Statistics. These time series have been pre-processed following the instructions of the FRED-MD database to make them more stationary [28]. To address the concern on potential nonstationarity over the entire time span, we break it into 58 rolling windows, each covering a five-year period. Motivated by our simulation study, we apply TSKI-LCD with $q \in \{0, 1\}$ described in Section 4.1 (with knockoff generator

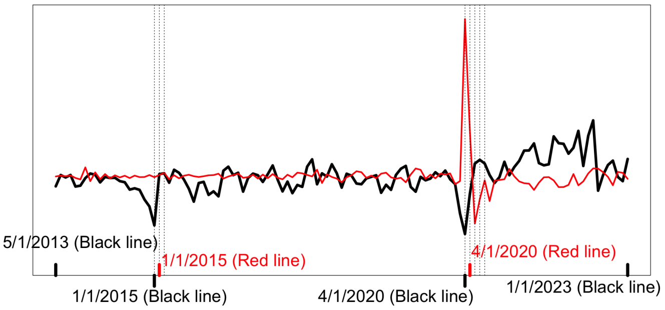
[3] The website URL: https://research.stlouisfed.org/econ/mccracken/fred-databases/.



(a) The red curve is ACOGNO



(b) The red curve is EXCAUSx



(c) The red curve is CLAIMSx

Figure 5: The black curves in the three panels are the inflation series at time $t+1$. The red curve in panel (a) is the number of new orders for consumer goods at time t, the red curve in panel (b) indicates the U.S./Canada exchange rate at time t, and the red curve in panel (c) is the U.S. initial claims for unemployment benefits at time t. All curves here are standardized and adjusted for visual comparison, and hence the values of these time series are not reported on the y-axis.

(8)) to each five-year rolling window to investigate the temporal relations between the inflation and other time series variables. The inflation at time t is defined as the adjusted consumer price index for all goods: $\text{Inflation}_t := \left(\frac{\text{CPI}_t - \text{CPI}_{t-1}}{\text{CPI}_{t-1}} \times 100\right)\%$, where CPI_t is the consumer price index for all goods at time (month) t. Each time series has a FRED-MD code. For example, CPIAUCSL is the FRED-MD code of the inflation series. To reduce randomness resulting from the use of the knockoffs construction, we repeat the inference procedure 100 times, and report the average results in Figure 3 with $q = 0$, where the left panel displays whether any significant covariates have been found at each rolling window period, and the right panel shows the selection frequency over 58 rolling window periods of the 127 time series covariates and their one month lags. That is, we model the one month ahead inflation series Inflation_{t+1} (response) using an AR(2) model with 127 time series covariates at current time t and 127 one month lags of these time series covariates. Consequently, the total covariate dimensionality is $p = 254$.

As can be seen in Figure 3, the TSKI-LCD with $q = 0$ identifies some active windows around the year 2021 (COVID) and 2022 (Russia-Ukraine conflict), with the selection frequency concentrating on a sparse set of covariates. The majority of the selected variables are covariates at the current time t. The top 10 most frequently selected covariates by the TSKI-LCD with $q = 0$ are employment-related series (HWI, CLAIMSx), consumption-related indices (ACOGNO, CPIAUCSL), housing-related series (PERMITS), U.S. bond yields (GS5, GS10), stock market indices (S.P.div.yield, S.P.500), and exchange rates (EXCAUSx) at their current time t, with their FRED-MD codes given in the parentheses.

The simulation results in Section 4 suggest that the choice of $q = 1$ has better FDR control especially when the sample size is limited. Motivated by our simulation results, we next apply the TSKI-LCD with subsampling $q = 1$ in Figure 4 with the expectation of better FDR control. The results of Figure 4 are more conservative in comparison to those in Figure 3. Despite being conservative, these new results also suggest some active windows around the same periods as $q = 0$, and the selected variables are also mostly covariates at the current time t. In addition, most covariates selected by the TSKI-LCD with $q = 1$ belong to the set selected by the TSKI-LCD with $q = 0$. In particular, the top 10 selected covariates are CPIAUCSL, CLAIMSx, PERMITS, AMDMUOx, GS10, ACOGNO, EXCAUSx, EXUSUKx, INDPRO, and IPFUELS, where only PERMITS is one-month lag covariate at $t-1$ in the AR(2) model. Among them, EXUSUKx, AMDMUOx, INDPRO, and IPFUELS are new in comparison to the list of the selected set when $q = 0$, where the first two are the U.S./U.K. exchange rate and another type of consumption-related index (the number of unfilled orders for durable goods), respectively, and the last two are industrial production indices that are related to consumption price indices. The difference in the selected sets of covariates is attributed to the fact that some economic covariates are designed to track similar economic factors and tend to be highly correlated.

The recent literature [37] suggests that inflation foresting is a difficult task in the sense that AR models with additional time series covariates rarely outperform simple AR models with only inflation lags. In other words, conditional on the lagged inflation series, additional covariates do not carry strong signals in inflation forecasting. The fact that the TSKI-LCD with $q = 1$ selects only a few time series covariates indeed supports such an argument. It is also interesting to notice that the stock market indices are not considered as important covariates by the TSKI-LCD with subsampling (i.e., $q = 1$), but are selected as important covariates when $q = 0$. In particular, the selection frequencies of S&P dividend yields and S&P 500 (both at lag $t-1$) are 100% and 79%, respectively, in Figure 3, while only 5% and 1%, respectively, in Figure 4, suggesting that stock market indices could be spurious findings.

The selection results by the TSKI-LCD motivate us to further investigate the dependency of inflation on a few time series covariates. In Figure 5, we plot three selected series, namely ACOGNO, EXCAUSx, and CLAIMSx, which are among the top 10 lists both when $q = 0$ and $q = 1$. These three selected covariates are all at their current time t in the AR(2) model. ACOGNO is the number of new orders for consumer goods, which is an important consumption index; EXCAUSx is the exchange rate from the U.S. dollar to the Canadian dollar, and CLAIMSx is the initial claim for unemployment benefits. We also include the inflation series at time $t+1$ in the same period (the black curve in each panel). For better visual comparison, all curves in Figure 5 have been standardized and adjusted.

A visual inspection of Figure 5 shows that the impacts of the COVID-19 pandemic in April 2020 on the U.S. economy are stronger than those of the gasoline price shock in January 2015. This potentially explains why our empirical findings of significant covariates concentrate mostly on this period. From Figure 5, we see that the gasoline price shock in January 2015 affects the consumption index series ACOGNO more mildly compared to the impact of COVID-19 in April 2020. Also, although there is some variation in the exchange rate EXCAUSx after January 2015, it is unclear whether such variation was caused by the gasoline price shock. In contrast, most of the U.S. economy's time series clearly responded to the pandemic to an unignorable degree. Particularly, the exchange rate and the number of initial claims dropped in March 2020, suggesting that these covariates were leading indicators of the inflation drop in April 2020. In summary, we have applied the newly suggested tool of TSKI to study the U.S. economy. Our empirical results illustrate the potential of the TSKI to obtain more instructive findings in real data applications.

References

[1] Adamek, R., S. Smeekes, and I. Wilms (2023). Lasso inference for high-dimensional time series. *Journal of Econometrics 235*, 1114–1143.

[2] An, H. and F. Huang (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica 6*, 943–956.

[3] Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*, 2055–2085.

[4] Barber, R. F., E. J. Candès, and R. J. Samworth (2020). Robust inference with knockoffs. *The Annals of Statistics 48*, 1409–1431.

[5] Basu, S. and G. Michailidis (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics 43*, 1535–1567.

[6] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B 57*, 289–300.

[7] Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of statistics 29*, 1165–1188.

[8] Bogdan, M., E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). SLOPE–adaptive variable selection via convex optimization. *The Annals of Applied Statistics 9*, 1103.

[9] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

[10] Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

[11] Brockwell, P. J., R. A. Davis, and S. E. Fienberg (1991). *Time Series: Theory and Methods.* Springer Science & Business Media.

[12] Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B 80*, 551–577.

[13] Chen, R. and R. S. Tsay (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association 88*(421), 298–308.

[14] Cline, D. B. H. and H.-M. Pu (2004). Stability and the Lyapounov exponent of threshold AR-ARCH models. *The Annals of Applied Probability 14*(4), 1920–1949.

[15] Crump, R. K., S. Eusepi, M. Giannoni, and A. Şahin (2022). The unemployment-inflation trade-off revisited: the Phillips curve in COVID times. Technical report, National Bureau of Economic Research.

[16] Durrett, R. (2019). *Probability: Theory and Examples*, Volume 49. Cambridge University Press.

[17] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*, 987–1007.

[18] Fan, Y., E. Demirkaya, G. Li, and J. Lv (2020). RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association 115*, 362–379.

[19] Fan, Y., E. Demirkaya, and J. Lv (2019). Nonuniformity of p-values can occur early in diverging dimensions. *Journal of Machine Learning Research 20*(77), 1–33.

[20] Fan, Y. and J. Lv (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics 44*(5), 2098–2126.

[21] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica 57*, 357–384.

[22] Han, Y., R. S. Tsay, and W. B. Wu (2023). High dimensional generalized linear models for temporal dependent data. *Bernoulli 29*, 105–131.

[23] Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and Its Interface 4*, 123–127.

[24] Ing, C.-K. (2020). Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics 48*, 1959–1980.

[25] King, R. G., J. H. Stock, and M. W. Watson (1995). Temporal instability of the unemployment-inflation relationship. *Economic Perspectives 19*(3), 2–13.

[26] Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics 186*(2), 325–344.

[27] Lanne, M. and P. Saikkonen (2005). Non-linear GARCH models for highly persistent volatility. *The Econometrics Journal 8*(2), 251–276.

[28] McCracken, M. W. and S. Ng (2016). FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*, 574–589.

[29] Medeiros, M. C. and E. F. Mendes (2016). ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics 191*, 255–271.

[30] Medeiros, M. C., G. F. Vasconcelos, Á. Veiga, and E. Zilberman (2021). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics 39*(1), 98–119.

[31] Meitz, M. and P. Saikkonen (2010). A note on the geometric ergodicity of a nonlinear AR-ARCH model. *Statistics & probability letters 80*(7-8), 631–638.

[32] Ozaki, T. (1985). 2 non-linear time series models and dynamical systems. *Handbook of Statistics 5*, 25–83.

[33] Ramdas, A. K., R. F. Barber, M. J. Wainwright, and M. I. Jordan (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics 47*, 2790–2821.

[34] Ren, Z. and R. F. Barber (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

[35] Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A 184*(2), 407–431.

[36] Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293–335.

[37] Stock, J. H. and M. W. Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking 39*, 3–33.

[38] Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Advances in Applied Probability 22*, 587–611.

[39] Tong, H. and K. S. Lim (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society Series B 42*, 245–292.

[40] Wang, R. and A. Ramdas (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 822–852.

[41] Wong, K. C., Z. Li, and A. Tewari (2020). Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics 48*, 1124–1142.

[42] Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability 22*, 94–116.

[43] Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory 57*, 4689–4708.

[44] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Supplementary Material to "High-Dimensional Knockoffs Inference for Time Series Data"

Chien-Ming Chi, Yingying Fan, Ching-Kang Ing and Jinchi Lv

This Supplementary Material contains the appendix of Section 2, additional simulation examples, the proofs of all main results and technical lemmas, and some additional technical details. All the notation is the same as defined in the main body of the paper. Additionally, we introduce some technical notation below. For $\vec{x} := (x_1, \cdots, x_n)^T \in \mathbb{R}^n$, we define $\|\vec{x}\|_k := (\sum_{i=1}^n |x_i|^k)^{1/k}$, $\|\vec{x}\|_\infty := \max_{1\le i\le n} |x_i|$, and $\|\vec{x}\|_0 := \sum_{i=1}^n \mathbf{1}_{\{x_i \neq 0\}}$ with $\mathbf{1}_{\{\cdot\}}$ being the indicator function. The distribution of a random mapping $\boldsymbol{X}$ is denoted as $\mu_{\boldsymbol{X}}$. For a matrix $\boldsymbol{X}$ and an index subset S, $\boldsymbol{X}(S)$ represents a submatrix of $\boldsymbol{X}$ containing only columns with indices in S. The total variation (TV) norm for any measures μ_1 and μ_2 on $(\Omega, \mathcal{F})$ is defined as $\|\mu_1 - \mu_2\|_{TV} := 2\sup_{\mathcal{D}\in\mathcal{F}} |\mu_1(\mathcal{D}) - \mu_2(\mathcal{D})|$.

A Appendix of Section 2

A.1 Example for Theorem 1

We begin with providing an example with asymptotically vanishing KL divergence. Assume that $\{\boldsymbol{x}_t\}$ follows a stationary linear Gaussian process as in Example 3 in Section 2.4 with zero mean and precision matrix (i.e., the inverse of the covariance matrix) $\Theta = [\mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t^T)]^{-1}$, and the knockoff generator is such that $\kappa(\vec{z}, \cdot)$ follows a Gaussian distribution with mean $(\boldsymbol{I}_p - D\widehat{\Theta})\vec{z}$ and variance $2D - D\widehat{\Theta}D$ for each $\vec{z} \in \mathbb{R}^p$, where $\boldsymbol{I}_p$ is the p-dimensional identity matrix, $\widehat{\Theta}$ is the estimated covariance matrix constructed from an independent learning sample, and D is a diagonal matrix with nonnegative entries such that $2D - D\widehat{\Theta}D$ is positive semidefinite. It has been shown in Lemma 5 of [4] that when the data consists of i.i.d. observations without serial dependency, it holds that for each

$\varepsilon > 0$,

$$\lim_{n\to\infty} \sum_{k=1}^{q+1} \mathbb{P}(\max_{1\le j\le p} \widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon) = 0 \tag{A.1}$$

as long as Condition 3 is satisfied, $p \gg q$, and

$$\max_{1\le j\le p} \Theta_{jj}^{-\frac{1}{2}} \left\| \Theta^{-\frac{1}{2}} (\widehat{\Theta}_j - \Theta_j) \right\|_2 = o_p \left\{ \frac{1}{\sqrt{n \log p}} \right\}, \tag{A.2}$$

where Θ_{jj} denotes the jth diagonal entry of Θ and Θ_j represents the jth column of Θ. We omit the dependence of the parameters on sample size n here for simplicity. More examples on asymptotically vanishing KL divergence for non-time series data can be found in the same paper above. Similar results can be proved for our applications of time series data using the proof techniques in [4] by replacing the concentration inequalities for i.i.d. observations with those for β-mixing data; since the extension is straightforward, we omit the details for simplicity.

We provide two remarks here. First, when Θ is known, it is obvious that the KL divergence is zero. Second, by Lemma 5 in [4], an independent learning sample of size $\widetilde{n} \gg n \log p$ is needed for (A.2) to hold. However, our simulation results indicate that using the full sample for both Θ estimation and TSKI inference can still control the FDR empirically, which suggests that the theoretical requirement may be unnecessarily strong. How to relax such an assumption in a time series setting is left for future investigation.

A.2 Additional stationary processes satisfying Condition 4

A.2.1 Various nonlinear AR-type processes

Many time-homogeneous Markov chains satisfy Condition 4. To name a few, [38, 2] showed that with some additional mild regularity conditions, $\{(Y_{t-1}, \cdots, Y_{t-k_1})\}$ in Example 4 below satisfies Condition 4 for all $h > 0$ with some constants C_0 and $0 \le \rho < 1$.

Example 4 (Nonlinear AR [38, 2])**.** *Let a measurable function $G : \mathbb{R}^{k_1} \longrightarrow \mathbb{R}$ for some constant integer $k_1 > 0$ be given such that $\sup_{\vec{z} \in \mathbb{R}^{k_1}} |G(\vec{z})| < \infty$, and $\{\varepsilon_t\}$ a sequence of i.i.d. model errors. For each t, we define*

$$Y_t = G(Y_{t-1}, \cdots, Y_{t-k_1}) + \varepsilon_t.$$

The self-exciting threshold autoregressive models (SETAR) [39, 23] also satisfies Condition 4 for all $h > 0$ according to [2]. For more examples such as the exponential AR models, see [32, 2] and the references therein.

A.2.2 ARCH-type process

Example 5 (AR(k_1)-X-ARCH(k_3) [14, 31])**.** *Let ε_t's be i.i.d. random variables with zero mean and $\mathbb{E}\varepsilon_1^2 = 1$, and $\{\boldsymbol{h}_t := (H_{t,1}, \cdots, H_{t,k_2})\}$ be a sequence of k_2-dimensional time series covariates that is independent of ε_t's. Let measurable functions $G_1 : \mathbb{R}^{k_3} \rightarrow (0, \infty)$, $G_2 : \mathbb{R}^{k_2} \rightarrow \mathbb{R}$, and $\gamma_j : \mathbb{R}^{k_1} \rightarrow \mathbb{R}$ with $1 \leq j \leq k_1$ be given such that $\sum_{j=1}^{k_1} \sup_{\vec{z} \in \mathbb{R}^{k_1}} |\gamma_j(\vec{z})| < 1$. The functional-coefficient ARX-ARCH model [13] is given by*

$$Y_t = \sum_{j=1}^{k_1} \gamma_j(Y_{t-1}, \cdots, Y_{t-k_1}) Y_{t-j} + G_2(\boldsymbol{h}_t) + \sigma_t \varepsilon_t$$

with $\sigma_t = G_1(\sigma_{t-1}\varepsilon_{t-1}, \cdots, \sigma_{t-k_3}\varepsilon_{t-k_3})$.

Example 5 above is an ARCH model with the mean function consisting of an AR component and exogenous covariates. The covariate vector is $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-k_1-k_3}, \boldsymbol{h}_t, \cdots, \boldsymbol{h}_{t-k_3})^T$ for response Y_t. In particular, Model 3 in Section 4 is an example of the ARX-ARCH model above when γ_j's are constants and the model error follows a standard ARCH process [17].

The AR component in Example 5 above can be a general functional-coefficient autoregressive model [13], and the ARCH component can take the form of a smooth transition

ARCH model [27]. With $G_2 = 0$ and additional mild regularity conditions, the Markov chain $\{Y_t, \cdots, Y_{t-k_1-k_3+1}\}_t$ satisfies Condition 4 for each $h > 0$ with constants $C_0 > 0$ and $0 \leq \rho < 1$ according to [14, 31]. In the presence of exogenous covariates $\boldsymbol{h}_t$, additional regularity conditions on $\boldsymbol{h}_t$'s are needed for Condition 4 to hold for the Markov chain; such study is beyond the scope of the current paper and is left for future investigation.

One challenge in the variable selection problem with time series data is that there does not always exist an obvious definition of the covariate vector. Taking Example 5 for instance, the existence of the ARCH component requires us to take into account both the mean function and variance function when selecting the set of non-null variables in the broad sense according to Definition 1. To better understand this, let us consider an ARX-ARCH(1) model with a standard ARCH component such that $\sigma_t = \sqrt{0.1 + 0.9(\sigma_{t-1}\varepsilon_{t-1})^2}$, and write

$$\sigma_{t-1}\varepsilon_{t-1} = Y_{t-1} - \sum_{j=1}^{k_1} \gamma_j(Y_{t-2}, \cdots, Y_{t-1-k_1})Y_{t-1-j} - G_2(\boldsymbol{h}_{t-1}). \tag{A.3}$$

In this example, in addition to variables $(Y_{t-1}, \cdots, Y_{t-k_1}, \boldsymbol{h}_t)$ which affect the mean regression function, we should also take into account the lagged covariates in the ARCH component $(Y_{t-1}, \cdots, Y_{t-k_1-1}, \boldsymbol{h}_{t-1})$ when conducting variable selection in the broad sense according to Definition 1. That is, one sensible choice of the covariate vector is $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-k_1-1}, \boldsymbol{h}_t, \boldsymbol{h}_{t-1})^T$. Omitting variables in the variance function (i.e., the ARCH component) and defining the covariate vector as $(Y_{t-1}, \cdots, Y_{t-k_1}, \boldsymbol{h}_t)$ may give us a nonsparse set of non-null variables according to Definition 1. Nevertheless, the actual variable selection performance of the TSKI depends on the specific choice of the knockoff statistics, as shown in our simulation section. If the knockoff statistics are constructed based on the mean regression function alone (e.g., the LCD and MDA discussed earlier), then the corresponding TSKI cannot be expected to have power in selecting variables that affect only the variance function. In this sense, our results in such a scenario should be interpreted as selecting the important variables contributing to the mean regression function alone.

Remark 1. *We have left out the GARCH-type process [9] in our discussion because it can be challenging to formulate meaningful covariates for variable selection purposes in such a setting. Note that a GARCH-type process can be represented as an ARCH-type process with infinite order. Thus, if the covariates vector is not well-formulated such that some active covariates are not included, the resulting regression model may no longer be sparse, rendering the FDR control problem invalid. We shall leave the variable selection problem for the GARCH-type process in future work.*

A.3 Robust TSKI without subsampling

In this section, we consider a special case of Algorithm 1 when $q = 0$, that is, no subsampling. For ease of reference, we provide a full description of the corresponding algorithm in Algorithm 2 below. Our theoretical study here has two major contributions: 1) extending the theory of robust knockoffs inference [4] to its e-value analog, where the non-robust version was first introduced and studied by [34] for i.i.d. data, and 2) further extending the results to time series applications. By similar analysis as in Theorem 3 below, we can show that the robust knockoffs inference [4] (without using the e-values) can also be extended to time series applications, but the details are omitted here for simplicity. We emphasize that our results (A.5)–(A.6) in Theorem 3 below assume *neither* i.i.d. observations *nor* the pairwise exchangibility Condition 3.

Let $\boldsymbol{X}_{-j}$ be the submatrix of $\boldsymbol{X}$ with the jth column removed, and $\boldsymbol{X}_j$ and $\widetilde{\boldsymbol{X}}_j$ the jth columns of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$, respectively. Recall that $(Y, \boldsymbol{x}, \widetilde{\boldsymbol{x}})$ is an independent copy of $(Y_1, \boldsymbol{x}_1, \widetilde{\boldsymbol{x}}_1)$ and $\widetilde{X}_j^{\dagger}$ is given in Condition 3 by the jth coordinatewise knockoff generator.

Algorithm 2: Time series knockoffs inference (TSKI) via e-values without subsampling

```
Let $0 < \tau_1 < 1$ be a constant and $0 < \tau^* < 1$ the target FDR level.
Calculate the knockoff statistics $W_1, \cdots, W_p$ satisfying (4) with the full sample
    $\{Y_t, \boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t\}_{t=1}^n$.
Let $\mathcal{W}_+ = \{|W_s| : |W_s| > 0\}$. Calculate the e-value statistics $e_j$'s such that
    $e_j = \frac{p \times \mathbf{1}_{\{W_j \geq T\}}}{1 + \sum_{s=1}^p \mathbf{1}_{\{W_s \leq -T\}}}, \quad T = \min\left\{t \in \mathcal{W}_+ : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \tau_1\right\}. \qquad (A.4)$
Let $\widehat{S} = \{j : e_j \geq p(\tau^* \times \widehat{k})^{-1}\}$ with $\widehat{k} = \max\{k : e_{(k)} \geq p(\tau^* \times k)^{-1}\}$, where $e_{(j)}$'s
    are the ordered statistics of $e_j$'s such that $e_{(1)} \geq \cdots \geq e_{(p)}$.
```

Theorem 3. *Let $\widehat{S}$ be the set of variables selected by the TSKI Algorithm 2 and $0 < \tau^* < 1$ the target FDR level. Assume that Condition 1 holds and T in (A.4) is positive. Then we have*

$$\mathrm{FDR} \leq \inf_{\varepsilon > 0} \Big[\tau^* \times e^{\varepsilon} + \mathbb{P}(\max_{1 \leq j \leq p} \widehat{\mathrm{KL}}_j > \varepsilon)\Big], \tag{A.5}$$

where for each $1 \leq j \leq p$,

$$\widehat{\mathrm{KL}}_j = \log\left(\frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}\right). \tag{A.6}$$

If we further assume that Condition 2 is satisfied and $\boldsymbol{X}_j$ is independent of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}_{-j}$ for each $j \in \mathcal{H}_0$, then we have

$$\widehat{\mathrm{KL}}_j = \log\left(\frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}}(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j})}{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}}(\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j})}\right). \tag{A.7}$$

Moreover, if $(\boldsymbol{x}, \widetilde{\boldsymbol{x}}), (\boldsymbol{x}_1, \widetilde{\boldsymbol{x}}_1), \cdots, (\boldsymbol{x}_n, \widetilde{\boldsymbol{x}}_n)$ are further assumed to be i.i.d., then we have

$$\widehat{\mathrm{KL}}_j = \sum_{t=1}^n \log\left(\frac{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(X_{tj}, \widetilde{X}_{tj}, \boldsymbol{x}_{-tj}, \widetilde{\boldsymbol{x}}_{-tj})}{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(\widetilde{X}_{tj}, X_{tj}, \boldsymbol{x}_{-tj}, \widetilde{\boldsymbol{x}}_{-tj})}\right). \tag{A.8}$$

If further Condition 3 is satisfied, then we have

$$\begin{aligned}\widehat{\mathrm{KL}}_j &= \sum_{t=1}^{n} \log \left(\frac{f_{X_j, \boldsymbol{x}_{-j}}(X_{tj}, \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(\widetilde{X}_{tj}, \boldsymbol{x}_{-tj})}{f_{X_j, \boldsymbol{x}_{-j}}(\widetilde{X}_{tj}, \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(X_{tj}, \boldsymbol{x}_{-tj})} \right) \\ &= \sum_{t=1}^{n} \log \left(\frac{f_{X_j | \boldsymbol{x}_{-j}}(X_{ij} | \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(\widetilde{X}_{tj} | \boldsymbol{x}_{-tj})}{f_{X_j | \boldsymbol{x}_{-j}}(\widetilde{X}_{tj} | \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(X_{tj} | \boldsymbol{x}_{-tj})} \right),\end{aligned} \tag{A.9}$$

where $f_{\boldsymbol{z}_1|\boldsymbol{z}_2}(\boldsymbol{z}_1|\boldsymbol{z}_2)$ denotes the conditional probability density function of $\boldsymbol{z}_1$ given $\boldsymbol{z}_2$.

The proof of Theorem 3 follows mainly those in [4, 34, 40] and is presented in Section B.3 later. Comparing (A.9) to (A.8), we see that the KL divergences become invariant to $\widetilde{\boldsymbol{x}}_{-j}$ thanks to the additional assumption Condition 3. The simplified form in (A.9) is important in proving the asymptotic FDR control as in (A.1). In addition, Condition 3 allows for deviation of the conditional distribution of $\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}$ from the true underlying conditional distribution of $X_j | \boldsymbol{x}_{-j}$, making the procedure more practically applicable, as verified in examples given in [4].

A.4 Appendix of Section 4

A.4.1 ARX and ARXARCH models

In this section, we present additional simulation examples: the autoregressive model with exogenous variables and the autoregressive conditional heteroskedasticity model with exogenous variables, as detailed in [17]. All the symbols and notation are consistent with those in Section 4.

Model 2 (ARX model)**.** *For each integer t, we define*

$$Y_t = \sum_{j=1}^{2} (-0.5)^{j-1} \beta Y_{t-j} + \sum_{j=1}^{15} 0.6 \times H_{t,j} + \varepsilon_t.$$

Model 3 (ARX-ARCH model)**.** *For each t, we define*

$$Y_t = \sum_{j=1}^{2}(-0.5)^{j-1}\beta Y_{t-j} + \sum_{j=1}^{15} 0.6 \times H_{t,j} + \sigma_t \varepsilon_t$$

with $\sigma_t^2 = 0.1 + 0.9(\sigma_{t-1}\varepsilon_{t-1})^2$.

The model error $\{\varepsilon_t\}$ is a sequence of i.i.d. standard Gaussian random variables and $\beta = 0.7$. The time series covariates are given by $H_{t,j} = \eta \times H_{t-1,j} + \epsilon_{t,j}$ with $j \in \{1, \cdots, 50\}$ and some $\eta \in \{0.2, 0.95\}$, where $(\epsilon_{t,1}, \cdots, \epsilon_{t,50})$'s are i.i.d. Gaussian random vectors with zero mean and $\mathbb{E}(\epsilon_{t,k}\epsilon_{t,l}) = (0.2)^{|k-l|}$ for all k, l. We formulate the coveriate vector with respect to response Y_t as $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-20}, \boldsymbol{h}_t, \boldsymbol{h}_{t-1}, \boldsymbol{h}_{t-2}, \boldsymbol{h}_{t-3}, \boldsymbol{h}_{t-4})$, where $\boldsymbol{h}_t = (H_{t,1}, \cdots, H_{t,50})$, giving rise to $p = 270$. We set $p = 270$ but vary the sample size n across experiments with $n \in \{200, 300, 500\}$.

In Models 2 and 3, the mean functions both depend linearly on the covariates. It is worth mentioning that because of the ARCH component, for Model 3, the relevant and null sets according to definition 1 are $S_{\text{arch}} = \{1, 2, 3, 21, \cdots, 35, 71, \cdots 85\}$ and $\mathcal{H}_{\text{arch}} = \{4, \cdots, 20, 36, \cdots, 70, 86, \cdots, 270\}$, respectively. The sets S_0 and $\mathcal{H}_0$ defined previously are the sets of active and null covariates, respectively, in the mean regression function. Although S_0 and $\mathcal{H}_0$ differ from S_{arch} and $\mathcal{H}_{\text{arch}}$, respectively, in Model 3, we examine the empirical power and FDR of the TSKI with respect to $\mathcal{S}_0$ and $\mathcal{H}_0$ for two reasons: 1) this is an interesting problem in time series inference and 2) random forests and Lasso are both algorithms designed for fitting the mean regression and thus are not expected to detect variables that affect only the variance function.

For implementation, the target FDR level is set as $\tau^* = 0.2$, and the R packages `glmnet` and `randomForest` are used for calculating the Lasso estimates and the random forests MDA, respectively. We generate the approximate knockoff variables using the idea of the second-order approximation [12, 18]. Specifically, for the ideal scenario with a zero-mean random vector $\boldsymbol{x}$ given, we sample its knockoff vector from the multivariate Gaussian distribution (also see (8))

$$\widetilde{\boldsymbol{x}}|\boldsymbol{x} \sim N(\boldsymbol{x} - \text{diag}(\vec{s})\Sigma^{-1}\boldsymbol{x}, 2\text{diag}(\vec{s}) - \text{diag}(\vec{s})\Sigma^{-1}\text{diag}(\vec{s})), \tag{A.10}$$

where $\Sigma = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T)$, $\vec{s} \in \mathbb{R}^p$ denotes the tuning parameters, and diag($\vec{s}$) is a diagonal matrix with diagonal entries in $\vec{s}$. Larger components of $\vec{s}$ imply that the resulting knockoff variables deviate more from the original features, thereby providing higher power in distinguishing them. Further details about this knockoff variable sampling procedure can be found in [12, 18]. In practical applications, we provide an estimate of the precision matrix $\widehat{\Sigma}^{-1}$ using the full sample and the method developed in [20], and select $\vec{s} = (\widehat{s}, \cdots, \widehat{s})^T$ with $\widehat{s}$ the inverse of the maximum eigenvalue of $\widehat{\Sigma}^{-1}$. Notably, this method matches the first two moments of the original covariates and their knockoffs counterparts and is thus termed the second-order approximation method. The TSKI Algorithm 1 with subsampling parameter $q \in \{0, 1\}$, $\tau^* = 0.2$, and $\tau_1 = \tau^*/(q+1)$ is considered in our simulation. The R code for the simulation experiments is available in the online Supplementary Material.

A.4.2 Empirical performance of TSKI

For all simulation experiments reported in Tables 1 and 3, both TSKI-LCD and TSKI-MDA with $q = 1$ control the FDR in finite samples at the target value of $\tau^* = 0.2$, but at the cost of lower selection power compared to the case of $q = 0$ (i.e., no subsampling). In contrast, TSKI with $q = 0$ has FDR exceeding the target FDR level in some cases.

When analyzing Models 2–3 with linear mean regression functions, the LCD-based method demonstrates superior performance over the MDA-based method in terms of power, as evidenced in Table 3. However, for Model 1 when $\iota = 5$ (i.e., high nonlinearity), we observe from Table 1 that MDA outperforms LCD in power when the sample

$n/p/\eta$	Method	q	FDR	Power	$n/p/\eta$	Method	q	FDR	Power
200/270/0.2	TSKI-LCD	0	0.237	0.992	200/270/0.2	TSKI-LCD	0	0.203	0.985
	TSKI-LCD	1	0.108	0.529		TSKI-LCD	1	0.122	0.755
	TSKI-MDA	0	0.273	0.391		TSKI-MDA	0	0.220	0.292
	TSKI-MDA	1	0.026	0.021		TSKI-MDA	1	0.044	0.021
300/270/0.2	TSKI-LCD	0	0.189	0.999	300/270/0.2	TSKI-LCD	0	0.233	0.999
	TSKI-LCD	1	0.110	0.97		TSKI-LCD	1	0.124	0.986
	TSKI-MDA	0	0.222	0.520		TSKI-MDA	0	0.185	0.387
	TSKI-MDA	1	0.049	0.057		TSKI-MDA	1	0.017	0.025
500/270/0.2	TSKI-LCD	0	0.188	1	500/270/0.2	TSKI-LCD	0	0.181	0.999
	TSKI-LCD	1	0.142	0.999		TSKI-LCD	1	0.166	0.996
	TSKI-MDA	0	0.208	0.729		TSKI-MDA	0	0.142	0.476
	TSKI-MDA	1	0.068	0.164		TSKI-MDA	1	0.037	0.076
500/270/0.95	TSKI-LCD	0	0.299	0.972	500/270/0.95	TSKI-LCD	0	0.312	0.961
	TSKI-LCD	1	0.172	0.979		TSKI-LCD	1	0.195	0.969
	TSKI-MDA	0	0.042	0.019		TSKI-MDA	0	0.032	0.015
	TSKI-MDA	1	0.000	0.000		TSKI-MDA	1	0.000	0.000

Table 3: The simulation results on the empirical FDR and power for the TSKI with $\tau_1 = \tau^*/(q+1)$ and $\tau^* = 0.2$ under Model 2 (left panel) and Model 3 (right panel) in Section 4.1.

Model 2 (ARX)			Model 3 (ARXARCH)		
$n/p/\eta$	FDR	Power	$n/p/\eta$	FDR	Power
200/270/0.2	–	–	200/270/0.2	–	–
300/270/0.2	0.007	0.009	300/270/0.2	0.025	0.04
500/270/0.2	0.029	0.989	500/270/0.2	0.094	0.983
500/270/0.95	0.099	0.999	500/270/0.95	0.233	0.987

Table 4: The simulation results on the empirical FDR and power for the ordinary least squares + Benjamini–Yekutieli (BY [7]) with the target FDR level at 0.2. This approach does not apply to high-dimensional scenarios with more features than observations.

Model 2 (ARX)			Model 3 (ARX-ARCH)		
$n/p/\eta$	FDR	Power	$n/p/\eta$	FDR	Power
200/270/0.2	0.001	0.999	200/270/0.2	0.009	0.999
300/270/0.2	0.000	1.000	300/270/0.2	0.001	1.000
500/270/0.2	0.131	1.000	500/270/0.2	0.069	0.999
500/270/0.95	0.001	0.991	500/270/0.95	0.008	1.000

Table 5: The simulation results on the empirical FDR and power for the adaptive Lasso [29, 44]. There is no target FDR level for the adaptive Lasso.

size is large (i.e., $n = 500$ and $q = 0$), an intuitive observation considering the nonparametric nature of the MDA measure. Indeed, the empirical performance of MDA decreases drastically in all settings when q increases from 0 to 1, because of the smaller sample size when calculating the MDA measures.

In Table 4, LS-BY exhibits a slightly higher error rate than $\tau^* = 0.2$ for Model 3 with $\eta = 0.95$. In addition, it is inapplicable to high-dimensional time series applications when $n < p$, such as our real data example in Section 5, due to the lack of a reliable p-value calculation method. It is important to emphasize that obtaining valid p-values under high-dimensional linear or nonlinear time series models, such as Models 1–3, presents a highly challenging and currently unresolved issue. Such a challenge is also reflected by the deteriorating performance of LS-BY when p becomes comparable to n. On the other hand, Table 5 highlights that adaLasso demonstrates the highest overall selection powers among all four methods. However, adaLasso fails to control error rates under Model 1, which is expected because it is developed for the linear model and focuses on consistent model selection instead of FDR control.

To sum up, the additional simulation results in this section support our conclusion in Section 4. For implementation, we suggest that practitioners working on time series inference with limited sample sizes initiate their diagnosis using TSKI-LCD with parameter q set to either 0 or 1, along with our knockoffs sampling procedure (8).

Remark 2. *Regarding the subsampling parameter q, we remark that by the construction of* (5) *in Algorithm 1, TSKI may have decent asymptotic power only when the number of relevant features is no less than $\tau_1^{-1} = (q+1)/\tau^*$. To see the intuition, let us consider an ideal scenario when the number of relevant features is less than τ_1^{-1} and these relevant features' knockoff statistics are the largest (positive) among all knockoff statistics. Then, even if the other knockoff statistics are all zero, we have $T^k = \infty$ in* (5)*, and, hence the knockoff filter screens out all features. It is worth mentioning that a similar requirement* *is assumed in Condition 7 for the asymptotic power.*

A.4.3 Additional simulation results for Model 1 with $q = 2$

Method	$n/p/\eta/\iota$	q	FDR	Power	$n/p/\eta/\iota$	q	FDR	Power
TSKI-LCD	200/270/0.2/0	0	0.157	0.698	300/270/0.2/0	0	0.164	0.870
TSKI-LCD		1	0.026	0.051		1	0.075	0.413
TSKI-LCD		2	0.000	0.000		2	0.007	0.012
TSKI-MDA		0	0.173	0.456		0	0.157	0.718
TSKI-MDA		1	0.026	0.028		1	0.041	0.102
TSKI-MDA		2	0.000	0.000		2	0.005	0.005
TSKI-LCD	200/270/0.2/5	0	0.139	0.287	300/270/0.2/5	0	0.160	0.514
TSKI-LCD		1	0.023	0.019		1	0.032	0.048
TSKI-LCD		2	0.000	0.000		2	0.000	0.000
TSKI-MDA		0	0.138	0.215		0	0.196	0.506
TSKI-MDA		1	0.012	0.011		1	0.038	0.036
TSKI-MDA		2	0.000	0.000		2	0.000	0.000

Method	$n/p/\eta/\iota$	q	FDR	Power	$n/p/\eta/\iota$	q	FDR	Power
TSKI-LCD	500/270/0.2/0	0	0.176	0.939	1000/270/0.2/0	0	0.222	0.975
TSKI-LCD		1	0.099	0.872		1	0.119	0.946
TSKI-LCD		2	0.047	0.216		2	0.126	0.879
TSKI-MDA		0	0.181	0.922		0	0.178	0.971
TSKI-MDA		1	0.092	0.550		1	0.107	0.939
TSKI-MDA		2	0.027	0.053		2	0.067	0.432
TSKI-LCD	500/270/0.2/5	0	0.141	0.634	1000/270/0.2/5	0	0.186	0.755
TSKI-LCD		1	0.086	0.267		1	0.117	0.613
TSKI-LCD		2	0.004	0.006		2	0.022	0.054
TSKI-MDA		0	0.166	0.679		0	0.199	0.849
TSKI-MDA		1	0.084	0.216		1	0.115	0.649
TSKI-MDA		2	0.000	0.000		2	0.027	0.068

Table 6: The simulation results on the empirical FDR and power for the TSKI with $\tau_1 = \tau^*/(q+1)$ and $\tau^* = 0.2$ under Model 1 in Section 4.1. The results in this table are the same as those in Table 1, except for the experiments with $q = 2$ or $n \in \{300, 1000\}$.

In this section, we present additional simulation for Model 1 with $q = 2$ and $n \in \{300, 1000\}$. The results with $q = 2$ and $n \in \{300, 1000\}$ in Table 6 complements those in Table 1; the other results in these two tables are the same.

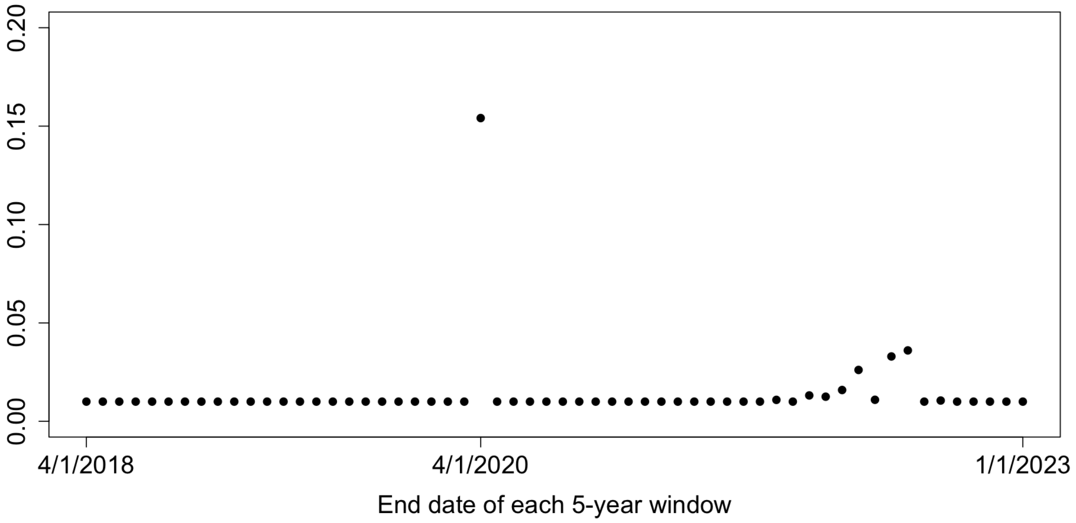
Let us begin with commenting on the results of $q = 2$ in Table 6 here. For all simulation experiments reported in Table 6, both TSKI-LCD and TSKI-MDA with $q = 2$ control the FDR in finite samples below the target value of $\tau^* = 0.2$, but at the cost of lower selection power compared to the case of $q \in \{0, 1\}$. When $q = 2$ and the sample size is small, TSKI becomes overly conservative with both low FDR and low power in most cases, with the MDA-based method suffering more severely from these issues. On the other hand, the additional results for $n \in \{300, 1000\}$ provide a clearer understanding of how the selection power increases with larger sample sizes. These results also suggest that the subsample size needs to be reasonably large for TSKI to have good power.

In summary, given the simulation results in Table 6, we recommend the use of $q = 1$ in practice as we are considering finite samples with limited sample sizes in our real data applications.

A.5 Augmented Dickey–Fuller test for unit roots

We run the augmented Dickey–Fuller (ADF) test implemented with the R package **aTSA** to test for unit roots. The null hypothesis of the ADF test is that the time series contains a unit root and is non-stationary, while the alternative hypothesis is that the time series is a stationary linear AR model. The unit root AR models considered by the ADF test may include b lags, where $b \geq 0$ is a tuning parameter. The unit root model of the ADF test also encompasses a drift term and a trend term. For more details, we refer to the R package **aTSA**. It is noteworthy that the ADF test with no lags ($i.e., b = 0$) is equivalent to the Dickey–Fuller test for the unit root.

The ADF test result for each rolling window is displayed in Figure 6. From Figure 6, it is observed that most periods do not exhibit clear numerical evidence of non-stationarity with unit roots (at p-value significance level 0.05), except for 4/1/2020, when COVID-19 occurred. In addition to the tests for rolling windows, we run the ADF test for the entire inflation series from 5/1/2013 to 1/1/2023 and rejected the null hypothesis of non-stationarity at $\alpha = 0.01$ for $b \in \{0, 1\}$.



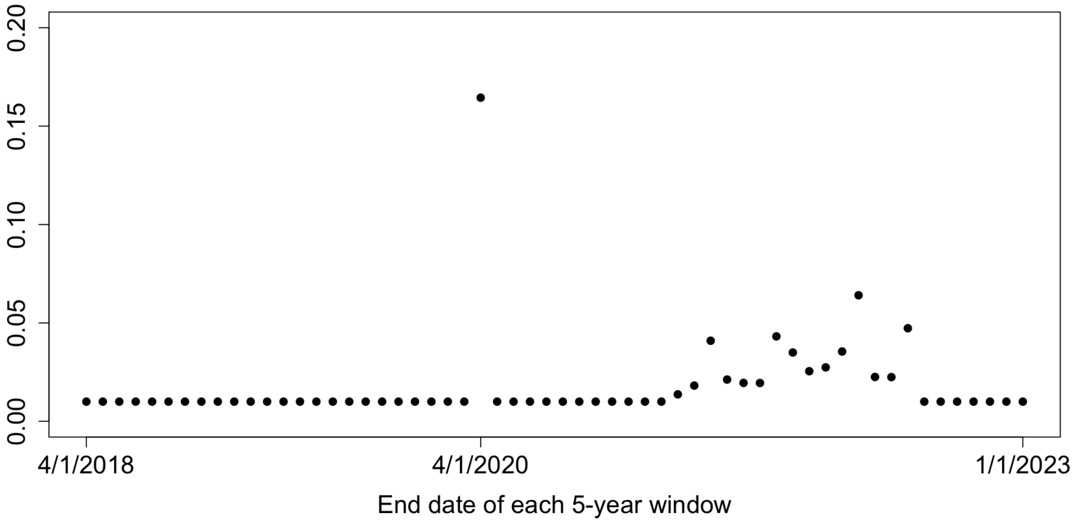


Figure 6: The results of the augmented Dickey–Fuller (ADF) test, where the unit root AR models include b lags. The y-axis displays the p-values of the tests. The top panel displays the results of the ADF test with one lag ($b = 1$) for each rolling window. The bottom panel displays the results of the ADF test with no lags ($b = 0$).

B Proofs of Theorems 1–2, Corollaries 1–2, and Proposition 1

B.1 Proof of Theorem 1

For simplicity, in this proof we use the notation $\boldsymbol{U}, \widetilde{\boldsymbol{U}}, \boldsymbol{V}$ to denote a generic subsample of $\{\boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t, Y_t\}_{t=1}^n$ or their independent and identically distributed (i.i.d.) counterparts $\{\boldsymbol{x}_t^\pi, \widetilde{\boldsymbol{x}}_t^\pi, Y_t^\pi\}_{t=1}^n$, where their exact meaning will be made explicit whenever confusion is possible. Let us define

$$\widehat{\mathrm{KL}}_j^k \equiv \log\left(\frac{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V}}(\boldsymbol{U}_j, \widetilde{\boldsymbol{U}}_j, \boldsymbol{U}_{-j}, \widetilde{\boldsymbol{U}}_{-j}, \boldsymbol{V})}{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V}}(\widetilde{\boldsymbol{U}}_j, \boldsymbol{U}_j, \boldsymbol{U}_{-j}, \widetilde{\boldsymbol{U}}_{-j}, \boldsymbol{V})}\right), \tag{A.11}$$

where $\boldsymbol{U} = (\boldsymbol{x}_i, i \in H_k)^T$, $\widetilde{\boldsymbol{U}} = (\widetilde{\boldsymbol{x}}_i, i \in H_k)^T$, and $\boldsymbol{V} = (Y_i, i \in H_k)^T$. We can use (A.37) in the proof of Theorem 3 in Section B.3 to conclude that for each $k \in \{1, \cdots, q+1\}$,

$$\mathbb{E}(\sum_{j\in\mathcal{H}_0} e_j^k \times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j^k \leq \varepsilon\}}) \leq p \times e^{\varepsilon}.$$

Then it holds that

$$\mathbb{E}(\sum_{j\in\mathcal{H}_0} e_j^{(\varepsilon)}) \leq p \times e^{\varepsilon},$$

where

$$e_j^{(\varepsilon)} = (q+1)^{-1} \sum_{k=1}^{q+1} e_j^k \times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j^k \leq \varepsilon\}}.$$

Denote by $\widehat{S}_\varepsilon$ be the set of selected features when applying the e-BH method to $e_j^{(\varepsilon)}$'s at the target FDR level τ^*. Then using similar arguments to those for (A.38) in Section B.3, we can show that

$$\mathbb{E}\left(\frac{\#(\widehat{S}_\varepsilon \cap \mathcal{H}_0)}{(\#\widehat{S}_\varepsilon) \vee 1}\right) \leq \tau^* \times e^{\varepsilon}.$$

Thus, it follows that

$$\begin{aligned}\mathbb{E}\left(\frac{\#(\widehat{S}\cap\mathcal{H}_0)}{(\#\widehat{S})\vee 1}\right) &\leq \mathbb{E}\left(\frac{\#(\widehat{S}_\varepsilon\cap\mathcal{H}_0)}{(\#\widehat{S}_\varepsilon)\vee 1}\times \mathbf{1}_{\{\widehat{S}=\widehat{S}_\varepsilon\}}+\mathbf{1}_{\{\widehat{S}\neq\widehat{S}_\varepsilon\}}\right)\\ &\leq \tau^*\times e^\varepsilon+\sum_{k=1}^{q+1}\mathbb{P}(\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j^k>\varepsilon)\end{aligned} \tag{A.12}$$

in view of $\{\widehat{S}\neq\widehat{S}_\varepsilon\}\subset\{\max_{1\leq k\leq q+1}\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j^k>\varepsilon\}$. Since (A.12) holds for each $\varepsilon>0$, we can further obtain that

$$\begin{aligned}\mathbb{E}\left(\frac{\#(\widehat{S}\cap\mathcal{H}_0)}{(\#\widehat{S})\vee 1}\right) \leq& \inf_{\varepsilon>0}\left\{\tau^* e^\varepsilon+\sum_{k=1}^{q+1}\mathbb{P}(\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j^{k\pi}>\varepsilon)\right\}\\ &+\sum_{k=1}^{q+1}\left(\mathbb{P}(\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j^{k}>\varepsilon)-\mathbb{P}(\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j^{k\pi}>\varepsilon)\right).\end{aligned} \tag{A.13}$$

It remains to bound the second term on the right-hand side of (A.13) above. Recall that $\widehat{\mathrm{KL}}_j^{k\pi}$ is defined analogously to (A.11) but based on i.i.d. sample $\{\boldsymbol{x}_t^\pi,\widetilde{\boldsymbol{x}}_t^\pi,Y_t^\pi\}_{t=1}^n$. With $\boldsymbol{U}=(\boldsymbol{x}_t^\pi,t\in H_k)^T$, $\widetilde{\boldsymbol{U}}=(\widetilde{\boldsymbol{x}}_t^\pi,t\in H_k)^T$, and $\boldsymbol{V}=(Y_t^\pi,i\in H_k)^T$, we can deduce that

$$\begin{aligned}\widehat{\mathrm{KL}}_j^{k\pi}&=\sum_{t\in H_k}\log\left(\frac{f_{X_j,\widetilde{X}_j,\boldsymbol{x}_{-j},\widetilde{\boldsymbol{x}}_{-j}}(X_{tj}^\pi,\widetilde{X}_{tj}^\pi,\boldsymbol{x}_{-tj}^\pi,\widetilde{\boldsymbol{x}}_{-tj}^\pi)}{f_{X_j,\widetilde{X}_j,\boldsymbol{x}_{-j},\widetilde{\boldsymbol{x}}_{-j}}(\widetilde{X}_{tj}^\pi,X_{tj}^\pi,\boldsymbol{x}_{-tj}^\pi,\widetilde{\boldsymbol{x}}_{-tj}^\pi)}\right)\\ &=\log\left(\frac{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j}}(\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j})}{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j}}(\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j})}\right)\\ &=\log\left(\frac{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V}}(\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V})}{f_{\boldsymbol{U}_j,\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V}}(\widetilde{\boldsymbol{U}}_j,\boldsymbol{U}_j,\boldsymbol{U}_{-j},\widetilde{\boldsymbol{U}}_{-j},\boldsymbol{V})}\right),\end{aligned} \tag{A.14}$$

where the third equality above follows from similar analysis to that of (A.7) in Theorem 1. The conditional column independence required by (A.7) holds for each $j\in\mathcal{H}_0$ because (A.14) involves i.i.d. samples.

We can now see that $\widehat{\mathrm{KL}}_j^k$ is well-defined thanks to Condition 1. Moreover, by the fact that the supports of $(\boldsymbol{x},\widetilde{\boldsymbol{x}})$ and $[\boldsymbol{x},\widetilde{\boldsymbol{x}}]_{\mathrm{swap}(\{j\})}$ are the same (as guaranteed by Condition 1) and the definition of $(\boldsymbol{x}_i^\pi,\widetilde{\boldsymbol{x}}_i^\pi)$'s, $\widehat{\mathrm{KL}}_j^{k\pi}$ is well-defined. Hence, in light of (A.11) and (A.14), there exists some measurable function $g : \mathbb{R}^{\# H_k \times (2p+1)} \longmapsto \mathbb{R}$ such that $\widehat{\mathrm{KL}}_j^k = g(\mathcal{X}_k)$ and $\widehat{\mathrm{KL}}_j^{k\pi} = g(\mathcal{X}_k^\pi)$ for each $\varepsilon \geq 0$, which entails that there exists some $\mathcal{D} \in \mathcal{R}^{\# H_k \times (2p+1)}$ such that

$$\begin{aligned} \{\mathcal{X}_k \in \mathcal{D}\} &= \{\max_{1 \leq j \leq p} \widehat{\mathrm{KL}}_j^k > \varepsilon\}, \\ \{\mathcal{X}_k^\pi \in \mathcal{D}\} &= \{\max_{1 \leq j \leq p} \widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon\}. \end{aligned} \tag{A.15}$$

With the aid of (A.15), it holds that

$$\begin{aligned} &|\mathbb{P}(\max_{1 \leq j \leq p} \widehat{\mathrm{KL}}_j^k > \varepsilon) - \mathbb{P}(\max_{1 \leq j \leq p} \widehat{\mathrm{KL}}_j^{k\pi} > \varepsilon)| \\ &\leq \sup_{\mathcal{D} \in \mathcal{R}^{\# H_k \times (2p+1)}} |\mathbb{P}(\mathcal{X}_k \in \mathcal{D}) - \mathbb{P}(\mathcal{X}_k^\pi \in \mathcal{D})|. \end{aligned} \tag{A.16}$$

Therefore, from (A.13)–(A.14) and (A.16) we can obtain the desired conclusion, which completes the proof of Theorem 1.

B.2 Proof of Theorem 2

Proof of (15). We aim to prove the second assertion (15) of Theorem 2, and will defer the proof of (14) to the end. Let the knockoff thresholds T^k's, knockoff statistics W_j^k's, statistics e_j^k's, and e-values e_j's be given as in Algorithm 1. Let us first outline the proof idea for (15) as follows. Using the inclusion-exclusion principle, we will show that $\cap_{k=1}^{q+1}\{j : e_j^k > 0\}$ includes most features in S^* with high probability. We then prove that each e_j with j in $\cap_{k=1}^{q+1}\{j : e_j^k > 0\}$ is sufficiently large to be selected by the e-BH procedure; that is, $\cap_{k=1}^{q+1}\{j : e_j^k > 0\} \subset \widehat{S}$. Combining all these results will complete the proof of (15). We will provide the full details of the proof next.

First, recall that $\widehat{S} = \{j : e_j \geq p(\tau^* \times \widehat{k})^{-1}\}$ with $\widehat{k} = \max\{k : e_{(k)} \geq p(\tau^* \times k)^{-1}\}$, where $e_{(j)}$'s are the ordered statistics of e_j's such that $e_{(1)} \geq \cdots \geq e_{(p)}$. Let $K > 1$ be the constant specified in Theorem 2. Let us consider two events given by

$$\cap_{k=1}^{q+1}\{\#\{j : W_j^k \geq T^k\} \leq K(\#S^*)\} \tag{A.17}$$

and

$$\cap_{k=1}^{q+1}\left\{\frac{\#(S^* \cap \{j : W_j^k \geq T^k\})}{\#S^*} \geq 1-(1+\phi)c_0k_{1n}^{-1}\right\}, \tag{A.18}$$

where c_0 is given in Condition 5 and that $\phi > 0$ is some real constant such that $\phi^2-\phi-1 = 0$. We will show that conditional on the two events in (A.17) and (A.18) above, it holds that

$$e_{(\#\mathcal{S})} = \min_{j\in\mathcal{S}} e_j \geq p(\tau^*(\#\mathcal{S}))^{-1}, \tag{A.19}$$

where

$$\mathcal{S} \coloneqq \cap_{k=1}^{q+1}\{j : W_j^k \geq T^k\}. \tag{A.20}$$

Then it follows from (A.19) and the definition of $\widehat{k}$ that $\widehat{k} \geq \#\mathcal{S}$ and $\mathcal{S} \subset \widehat{S}$. Such results along with an application of the inclusion–exclusion principle entail that conditional on the intersection of events (A.17) and (A.18), we have

$$\frac{\#(S^* \cap \widehat{S})}{\#S^*} \geq \frac{\#(S^* \cap \mathcal{S})}{\#S^*} \geq 1-(q+1)(1+\phi)c_0k_{1n}^{-1}. \tag{A.21}$$

Since it holds that

$$\begin{aligned}\mathbb{E}\Big[\frac{\#(S^* \cap \widehat{S})}{\#S^*}\Big] &\geq \Big(1-(q+1)(1+\phi)c_0k_{1n}^{-1}\Big)\\ &\quad\times \mathbb{P}\Big(\frac{\#(S^* \cap \widehat{S})}{\#S^*} \geq 1-(q+1)(1+\phi)c_0k_{1n}^{-1}\Big)\\ &\geq \Big(1-(q+1)(1+\phi)c_0k_{1n}^{-1}\Big)\mathbb{P}\Big(\text{event in (A.17)} \cap \text{event in (A.18)}\Big),\end{aligned} \tag{A.22}$$

to establish (15) we need only to prove (A.19) and construct the probability lower bounds for events in (A.17) and (A.18).

To show (A.19), note that by the definition of T^k's and the assumption that $T^k < \infty$, we have that conditional on event given in (A.17),

$$\begin{aligned} 1 + \#\{j : W_j^k \leq -T^k\} &\leq \tau_1(\#\{j : W_j^k \geq T^k\}) \\ &\leq \tau_1 K(\#S^*) \end{aligned} \tag{A.23}$$

for each $k \in \{1, \cdots, q+1\}$. In view of (A.23), it holds that for each nonzero e_j^k,

$$\begin{aligned} e_j^k &= \frac{p}{\#\{s : W_s^k \leq -T^k\} + 1} \\ &\geq \frac{p}{\tau_1 K(\#S^*)} \\ &\geq \frac{p}{\tau^* \times (1 - (1+q)(1+\phi)c_0 k_{1n}^{-1}) \times (\#S^*)}, \end{aligned} \tag{A.24}$$

where we have used the definitions of e_j^k's, ϕ, and τ_1. Then from (A.24) and the definition $e_j = (q+1)^{-1} \sum_{k=1}^{q+1} e_j^k$, we can deduce that conditional on event (A.17),

$$\min_{j \in \mathcal{S}} e_j \geq \frac{p}{\tau^* \times (1 - (1+q)(1+\phi)c_0 k_{1n}^{-1}) \times (\#S^*)},$$

which establishes (A.19).

It remains to provide the probability upper bounds for the complementary events of (A.17) and (A.18), which are given in (A.27) and (A.25), respectively, below. By the assumptions (Conditions 6–7 and that Condition 5 is satisfied for the Lasso estimates applied to each subsample in H_k in Algorithm 1), we can show that

$$\mathbb{P}(\text{complementary event of (A.18)}) \leq (q+1) \times (k_{2n} + k_{3n}) \tag{A.25}$$

for all large n. We postpone the detailed proof of (A.25) to a later part.

On the other hand, it follows from the assumption of $\#S^* > 0$ that conditional on event $\{\#\{j : W_j^k \geq T^k\} > K(\#S^*)\}$,

$$\begin{aligned}\frac{\#(\{j : W_j^k \geq T^k\} \cap (S^*)^c)}{\#\{j : W_j^k \geq T^k\} \vee 1} &\geq \frac{\#\{j : W_j^k \geq T^k\} - \#S^*}{\#\{j : W_j^k \geq T^k\} \vee 1} \\ &> \frac{\#f\{j : W_j^k \geq T^k\} - \#\{j : W_j^k \geq T^k\} \times K^{-1}}{\#\{j : W_j^k \geq T^k\} \vee 1} \\ &\geq \frac{K-1}{K}.\end{aligned} \tag{A.26}$$

Therefore, from (16) and (A.26) some simple calculations give that

$$\begin{aligned}&\mathbb{P}(\{\#\{j : W_j^k \geq T^k\} > K(\#S^*)\}) \times \frac{K-1}{K} + \mathbb{P}(\{\#\{j : W_j^k \geq T^k\} \leq K(\#S^*)\}) \times 0 \\ &\leq \mathbb{E}\left(\frac{\#(\{j : W_j^k \geq T^k\} \cap (S^*)^c)}{\#\{j : W_j^k \geq T^k\} \vee 1}\right) \\ &\leq \tau_1 + \theta_\varepsilon,\end{aligned}$$

which yields that

$$\mathbb{P}(\text{complementary event of (A.17)}) \leq (q+1) \times \frac{(\tau_1 + \theta_\varepsilon)K}{K-1}. \tag{A.27}$$

This establishes the desired conclusion in (15) of Theorem 2.

Proof of (A.25). We need to prove for each $k \in \{1, \cdots, p+1\}$ that conditional on event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\#S^*)\lambda_n\} \cap \{\#\{j : W_j^k \geq T^k\} \geq c_1(\#S^*)\}$, it holds that

$$\frac{\#(S^* \cap \{j : W_j^k \geq T^k\})}{\#S^*} \geq 1 - (1+\phi)c_0 k_{1n}^{-1}, \tag{A.28}$$

where we recall that c_0 is from Condition 5 and that $\phi > 0$ is a real constant such that $\phi^2 - \phi - 1 = 0$. Then (A.25) follows from Conditions 5–6.

We now proceed with establishing (A.25). Without loss of generality, we consider the case $k = 1$, and omit the superscripts "k" on e_j^k's, W_j^k's, and T^k's for simplicity.

Assume without loss of generality that

$$|W_1| \geq \cdots \geq |W_p|.$$

Let $j^* \in \{1, \cdots, p\}$ be given such that $j^* \in \{s : |W_s| = T\}$. Such j^* always exists because of the assumption that $T < \infty$. Then it follows that

$$-T < W_{j^*+1} \leq 0$$

by the definition of T (because otherwise T would take a smaller value than $|W_{j*}|$) and the assumption that there are no ties in $\{|W_j| : |W_j| > 0\}$. We will analyze two cases separately, where the first case considers $W_{j^*+1} = 0$ and the second case considers $-T < W_{j^*+1} < 0$.

Let us consider the first case of $W_{j^*+1} = 0$. Denote by $\widetilde{q} = \phi c_0 k_{1n}^{-1}$ with $\phi > 0$ and $\phi^2 - \phi - 1 = 0$. We will discuss the scenarios of $\#\{j : W_j < 0\} \leq \widetilde{q}(\#S^*)$ and $\#\{j : W_j < 0\} > \widetilde{q}(\#S^*)$ separately, where the former case will be examined here and the latter one will be left to a later part. For the scenario of $\#\{j : W_j < 0\} \leq \widetilde{q}(\#S^*)$, some simple calculations together with $W_{j*+1} = 0$ give that

$$\begin{aligned} \#(\{j : W_j \geq T\} \cap S^*) &= \#(\{j : |W_j| > 0\} \cap S^*) - \#(\{j : W_j < 0\} \cap S^*) \\ &\geq \#(\{j : |W_j| > 0\} \cap S^*) - \widetilde{q}(\#S^*). \end{aligned} \tag{A.29}$$

We will deal with the term $\#(\{j : |W_j| > 0\} \cap S^*)$ on the RHS of (A.29) below. On the event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\# S^*)\lambda_n\}$, we can deduce that

$$\begin{aligned} c_0 \lambda_n (\# S^*) &\geq \sum_{j \in \widehat{S}_1 \cap S^*} |\beta_j^*| \\ &\geq \#(\widehat{S}_1 \cap S^*) \times (\min_{j \in S^*} |\beta_j^*|), \end{aligned}$$

where $\widehat{S}_1 = \{j : |\widehat{\beta}_j| = 0\}$. Such result and Condition 6 entail that

$$c_0(\# S^*) k_{1n}^{-1} \geq \#(\widehat{S}_1 \cap S^*).$$

Hence, it follows from the assumption that there are no ties in $\{|\widehat{\beta}_j| : |\widehat{\beta}_j| > 0\}$ that

$$\begin{aligned} \#(\{j : |W_j| > 0\} \cap S^*) &= \#((\widehat{S}_1)^c \cap S^*) \\ &\geq (1 - c_0 k_{1n}^{-1}) \times (\# S^*). \end{aligned} \tag{A.30}$$

Then combining (A.29)–(A.30), we can obtain that conditional on event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\# S^*)\lambda_n\}$,

$$\begin{aligned} \frac{\#(\{j : W_j \geq T\} \cap S^*)}{\# S^*} &\geq 1 - c_0 k_{1n}^{-1} - \widetilde{q} \\ &= 1 - (1 + \phi) c_0 k_{1n}^{-1}, \end{aligned} \tag{A.31}$$

which establishes (A.28). Moreover, observe that the second scenario of $\#\{j : W_j < 0\} > \widetilde{q}(\# S^*)$ when $W_{j^*+1} = 0$ implies that

$$\#\{j : W_j \leq -T\} = \#\{j : W_j < 0\} > \widetilde{q}(\# S^*),$$

which reduces to the same form as in (A.32) below. Thus, the proof provided below can be applied here to conclude the proof for the first case $W_{j^*+1} = 0$.

We now consider the case of $-T < W_{j^*+1} < 0$. From the definition of T and the assumption that there are no ties in $\{|W_j| : |W_j| > 0\}$, it holds on event $\{\#\{j : W_j \geq T\} \geq c_1(\#S^*)\}$ that

$$\begin{aligned} \#\{j : W_j \leq -T\} + 2 &\geq \tau_1 \times \#\{j : W_j \geq T\} \\ &\geq \tau_1 c_1(\#S^*). \end{aligned} \tag{A.32}$$

Meanwhile, on the event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\#S^*)\lambda_n\}$, since $\beta_{j+p}^* = 0$ for all $j > 0$ and $|\widehat{\beta}_{j+p}| \geq T + |\widehat{\beta}_j|$ for all $j \in \{s : W_s \leq -T\}$, we have that

$$\begin{aligned} c_0\lambda_n(\#S^*) &\geq \sum_{j:W_j \leq -T} |\widehat{\beta}_{j+p} - \beta_{j+p}^*| \\ &= \sum_{j:W_j \leq -T} |\widehat{\beta}_{j+p}| \\ &\geq \#\{j : W_j \leq -T\} \times T. \end{aligned} \tag{A.33}$$

Then from (A.32)–(A.33) and Conditions 6–7, we can obtain that conditional on event $\{\#\{j : W_j \geq T\} \geq c_1(\#S^*)\} \cap \{\#\{j : W_j \geq T\} \geq c_1(\#S^*)\}$,

$$T \leq \frac{c_0\lambda_n(\#S^*)}{c_1\tau_1(\#S^*) - 2} \leq k_{1n}\lambda_n\phi^{-1} \tag{A.34}$$

for all large n.

Further, conditional on event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\#S^*)\lambda_n\} \cap \{\#\{j : W_j \geq T\} \geq$

$c_1(\#S^*)\}$, we can deduce that for all large n,

$$\begin{aligned}
c_0\lambda_n(\#S^*) &\geq \sum_{j\in S^*\cap(\{j:W_j\geq T\})^c} (|\widehat{\beta}_j - \beta_j^*| + |\widehat{\beta}_{j+p}|) \\
&\geq \sum_{j\in S^*\cap(\{j:W_j\geq T\})^c} (|\widehat{\beta}_j - \beta_j^*| + |\widehat{\beta}_j| - T) \\
&\geq \sum_{j\in S^*\cap(\{j:W_j\geq T\})^c} (|\beta_j^*| - T) \\
&\geq \#(S^*\cap(\{j:W_j\geq T\})^c) \times k_{1n}(1-\phi^{-1})\lambda_n,
\end{aligned} \tag{A.35}$$

where the second inequality above is from the fact that $|\widehat{\beta}_{j+p}| > |\widehat{\beta}_j| - T$ for j in $\{j : W_j \geq T\}^c$, the third inequality above is due to the triangle inequality, and the last inequality above results from Condition 6 and (A.34).

In light of (A.35), it holds on event $\{\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \leq c_0(\#S^*)\lambda_n\} \cap \{\#\{j : W_j \geq T\} \geq c_1(\#S^*)\}$ that for all large n,

$$\begin{aligned}
\frac{\#(S^*\cap\{j:W_j\geq T\})}{\#S^*} &= 1 - \frac{\#(S^*\cap(\{j:W_j\geq T\})^c)}{\#S^*} \\
&\geq 1 - \frac{c_0}{k_{1n}(1-\phi^{-1})} \\
&= 1 - (1+\phi)c_0k_{1n}^{-1},
\end{aligned} \tag{A.36}$$

where the second equality above follows from the definition of ϕ. This establishes (A.28). Thus, combining the above results concludes the proof for (A.25).

Proof of (14). Finally, we show the second assertion (14) of Theorem 2. Let us observe that by the construction of Algorithm 1 and the definition of $\mathcal{S}$ in (A.20), it holds that

$$\mathbb{P}(\{\widehat{S} = \emptyset\} \cup \{\mathcal{S} \subset \widehat{S}\}) = 1.$$

Then by (A.18)–(A.21), (A.25), and the fact that $\phi \leq 3$ (recall that ϕ is defined at the beginning of this proof), we can obtain the desired result in (14). It is worth mentioning that we do not require $\tau_1 \leq \tau^*$ here. This completes the proof of Theorem 2.

B.3 Proof of Theorem 3

Let us first make a useful claim that with $\widehat{\mathrm{KL}}_j$'s given in (A.6), it holds that for each $\varepsilon > 0$,

$$\sum_{j \in \mathcal{H}_0} \mathbb{E}(e_j \times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j \leq \varepsilon\}}) \leq p \times e^{\varepsilon}. \tag{A.37}$$

Then we consider an application of the e-BH method [40] to $e_j^{(\varepsilon)} := e_j \times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j \leq \varepsilon\}}$ with the target FDR level τ^*, yielding a set of selected features $\widehat{S}_\varepsilon \subset \{1, \cdots, p\}$ defined as

$$\widehat{S}_\varepsilon = \{j : e_j^{(\varepsilon)} \geq p(\tau^* \times \widehat{k}_\varepsilon)^{-1}\}$$

with $\widehat{k}_\varepsilon \equiv \max\{k : e_{(k)}^{(\varepsilon)} \geq p(\tau^* \times k)^{-1}\}$. Here, $e_{(j)}^{(\varepsilon)}$'s are the ordered statistics of $e_j^{(\varepsilon)}$'s such that $e_{(1)}^{(\varepsilon)} \geq \cdots \geq e_{(p)}^{(\varepsilon)}$. It is easy to see that $\#\widehat{S}_\varepsilon = \widehat{k}_\varepsilon$.

In view of the definition of $\widehat{k}_\varepsilon$, we can deduce that

$$\begin{aligned}
\mathbb{E}\left(\frac{\#(\widehat{S}_\varepsilon \cap \mathcal{H}_0)}{(\#\widehat{S}_\varepsilon) \vee 1}\right) &= \mathbb{E}\left(\frac{\sum_{j \in \mathcal{H}_0} \mathbf{1}_{\{j \in \widehat{S}_\varepsilon\}}}{\widehat{k}_\varepsilon \vee 1}\right) \\
&\leq \mathbb{E}\left(\frac{\sum_{j \in \mathcal{H}_0} \mathbf{1}_{\{j \in \widehat{S}_\varepsilon\}} \times \tau^* \times e_j^{(\varepsilon)}}{p}\right) \\
&\leq \tau^* p^{-1} \times \mathbb{E}\left(\sum_{j \in \mathcal{H}_0} e_j^{(\varepsilon)}\right) \\
&\leq \tau^* e^{\varepsilon},
\end{aligned} \tag{A.38}$$

where the last inequality above is from (A.37). Then combining (A.38) and $\{\widehat{S} \neq \widehat{S}_\varepsilon\} \subset$

$\cup_{j=1}^{p}\{\widehat{\mathrm{KL}}_j > \varepsilon\}$ leads to

$$\mathbb{E}\left(\frac{\#(\widehat{S}\cap\mathcal{H}_0)}{(\#\widehat{S})\vee 1}\right) \leq \mathbb{E}\left(\mathbf{1}_{\{\widehat{S}=\widehat{S}_\varepsilon\}}\times\frac{\#(\widehat{S}_\varepsilon\cap\mathcal{H}_0)}{(\#\widehat{S}_\varepsilon)\vee 1}+\mathbf{1}_{\{\widehat{S}\neq\widehat{S}_\varepsilon\}}\right)$$
$$\leq \tau^* e^{\varepsilon} + \mathbb{P}(\max_{1\leq j\leq p}\widehat{\mathrm{KL}}_j > \varepsilon)$$

for each $\varepsilon \geq 0$. This concludes the proof for the desired result (A.5). We will provide the proofs of (A.37) and (A.7)–(A.9), separately.

Proof of (A.37). Let us define

$$T_j \equiv \min\left\{t\in\mathcal{W}_+^\dagger : \frac{1+\#\{s: W_s^\dagger \leq -t\}}{\#\{s: W_s^\dagger \geq t\}\vee 1}\leq \tau_1\right\},$$

where $W_k^\dagger = W_k$ if $k\neq j$ and $W_j^\dagger = |W_j|$, $\mathcal{W}_+^\dagger = \{|W_s^\dagger| : |W_s^\dagger| > 0\}$, and $\min\emptyset$ is defined as infinity. We further define $\boldsymbol{X}_j^{(0)}$ and $\boldsymbol{X}_j^{(1)}$ such that $\boldsymbol{X}_j^{(0)} = \boldsymbol{X}_j$ and $\boldsymbol{X}_j^{(1)} = \widetilde{\boldsymbol{X}}_j$ if $W_j \geq 0$, and $\boldsymbol{X}_j^{(1)} = \boldsymbol{X}_j$ and $\boldsymbol{X}_j^{(0)} = \widetilde{\boldsymbol{X}}_j$ if $W_j < 0$.

For each $\varepsilon \geq 0$, we can deduce that

$$\begin{aligned}
&\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}_{\{W_j\geq T\}}\times\mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}}{1+\sum_{s=1}^{p}\mathbf{1}_{\{W_s\leq -T\}}}\right)\\
&=\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}_{\{W_j\geq T_j\}}\times\mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}}{1+\sum_{s=1}^{p}\mathbf{1}_{\{W_s\leq -T_j\}}}\right)\\
&=\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}_{\{W_j\geq T_j\}}\times\mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}}{1+\sum_{s=1,s\neq j}^{p}\mathbf{1}_{\{W_s\leq -T_j\}}}\right) \qquad\qquad (A.39)\\
&=\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}_{\{W_j>0\}}\times\mathbf{1}_{\{|W_j|\geq T_j\}}\times\mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}}{1+\sum_{s=1,s\neq j}^{p}\mathbf{1}_{\{W_s\leq -T_j\}}}\right)\\
&=\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbb{P}(W_j>0,\widehat{\mathrm{KL}}_j\leq\varepsilon\big|\ \boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})\times\mathbf{1}_{\{|W_j|\geq T_j\}}}{1+\sum_{s=1,s\neq j}^{p}\mathbf{1}_{\{W_s\leq -T_j\}}}\right),
\end{aligned}$$

where the first three equalities above hold because when $\mathbf{1}_{\{W_j\geq T\}} = 1$, we have $W_j > 0$, $T = T_j$, and $\mathbf{1}_{\{W_j\leq -T_j\}} = 0$, and the last equality above holds since $|W_j|$, T_j, and $W_1,\cdots,W_{j-1}$, $W_{j+1},\cdots,W_p$ are functions of $(\boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})$ due to the sign-flip property (4).

From the definitions of $\boldsymbol{X}_j^{(0)}$, $\boldsymbol{X}_j^{(1)}$, and $\widehat{\mathrm{KL}}_j$, we can obtain that

$$\begin{aligned}
&\mathbb{P}(W_j > 0, \widehat{\mathrm{KL}}_j \le \varepsilon \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= \mathbb{P}(W_j > 0, \widehat{\mathrm{KL}}_j^{(01)} \le \varepsilon \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= \mathbb{P}(W_j > 0 \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j^{(01)} \le \varepsilon\}},
\end{aligned} \tag{A.40}$$

where

$$\widehat{\mathrm{KL}}_j^{(01)} \equiv \log\left(\frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(1)}, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}\right).$$

Furthermore, we will show that it holds almost surely that

$$\begin{aligned}
&\mathbb{P}(W_j > 0 \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&\le \frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(1)}, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})} \\
&\qquad \times \mathbb{P}(W_j < 0 \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= e^{\widehat{\mathrm{KL}}_j^{(01)}} \times \mathbb{P}(W_j < 0 \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}),
\end{aligned} \tag{A.41}$$

where Condition 1 is assumed to avoid division by zero on the right-hand side (RHS) of the second inequality above. The proof of (A.41) is deferred to right after the proof of (A.37).

By (A.40)–(A.41), it holds that

$$
\begin{aligned}
&\text{RHS of (A.39)}\\
&\leq \sum_{j\in\mathcal{H}_0} \mathbb{E}\left(\frac{e^{\widehat{\mathrm{KL}}_j^{(01)}}\mathbb{P}(W_j<0\big|\, \boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})\,\mathbf{1}_{\{|W_j|\geq T_j\}}\,\mathbf{1}_{\{\widehat{\mathrm{KL}}_j^{(01)}\leq\varepsilon\}}}{1+\sum_{s=1,s\neq j}^p \mathbf{1}_{\{W_s\leq -T_j\}}}\right)\\
&\leq e^{\varepsilon}\sum_{j=1}^p \mathbb{E}\left(\frac{\mathbf{1}_{\{W_j\leq -T_j\}}}{1+\sum_{s=1,s\neq j}^p \mathbf{1}_{\{W_s\leq -T_j\}}}\right)\\
&\leq e^{\varepsilon}\,\mathbb{E}\left(\frac{\sum_{j=1}^p \mathbf{1}_{\{W_j\leq -T_j\}}}{1+\sum_{s=1,s\neq j}^p \mathbf{1}_{\{W_s\leq -T_s\}}}\right)\\
&\leq e^{\varepsilon}\,\mathbb{E}\left(\frac{\sum_{j=1}^p \mathbf{1}_{\{W_j\leq -T_j\}}}{1\vee\left(\sum_{s=1}^p \mathbf{1}_{\{W_s\leq -T_s\}}\right)}\right)\\
&\leq e^{\varepsilon},
\end{aligned}
\tag{A.42}
$$

where RHS is short for the right-hand side, and the third inequality above follows from Lemma 6 in [4]. Hence, by resorting to (A.39), (A.42), and the fact that

$$
\sum_{j\in\mathcal{H}_0}\mathbb{E}(e_j\times \mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}) = p\times\sum_{j\in\mathcal{H}_0}\mathbb{E}\left(\frac{\mathbf{1}_{\{W_j\geq T\}}\times\mathbf{1}_{\{\widehat{\mathrm{KL}}_j\leq\varepsilon\}}}{1+\sum_{s=1}^p\mathbf{1}_{\{W_s\leq -T\}}}\right),
$$

we can establish (A.37).

Proof of (A.41). Denote by

$$
F_{>0}(\boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y}) \ \text{ and } \ F_{<0}(\boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})
$$

the versions of

$$
\mathbb{P}(W_j>0|\boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y}) \ \text{ and } \ \mathbb{P}(W_j<0|\boldsymbol{X}_j^{(0)},\boldsymbol{X}_j^{(1)},\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y}),
$$

respectively. We will show that functions $F_{>0}:\mathbb{R}^{n(2p+1)}\longmapsto\mathbb{R}$ and $F_{<0}:\mathbb{R}^{n(2p+1)}\longmapsto\mathbb{R}$

satisfiy that

$$\begin{aligned} F_{>0}(\vec{z}) &= \frac{\mathbf{1}_{\{w_j(\vec{z})>0\}} \times f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z})}, \\ F_{<0}(\vec{z}) &= \frac{\mathbf{1}_{\{w_j(\vec{z}_{\text{swap}})<0\}} \times f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z}_{\text{swap}})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z})}, \end{aligned} \tag{A.43}$$

respectively, where $\vec{z} = (\vec{z}_1, \vec{z}_2, \vec{z}_3, \vec{z}_4, \vec{z}_5)$, $\vec{z}_{\text{swap}} = (\vec{z}_2, \vec{z}_1, \vec{z}_3, \vec{z}_4, \vec{z}_5)$ with $\vec{z}_1 \in \mathbb{R}^n$, $\vec{z}_2 \in \mathbb{R}^n$, $\vec{z}_3 \in \mathbb{R}^{n(p-1)}$, $\vec{z}_4 \in \mathbb{R}^{n(p-1)}$, $\vec{z}_5 \in \mathbb{R}^n$, and $w_j : \mathbb{R}^{n(2p+1)} \longmapsto \mathbb{R}$ denotes the knockoff statistic function of W_j. From the definitions of $\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}$, and $w_j(\cdot)$ along with the sign-flip property (4), we have that almost surely,

$$\begin{aligned} & w_j(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \geq 0, \\ & w_j(\boldsymbol{X}_j^{(1)}, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) < 0. \end{aligned} \tag{A.44}$$

Then an application of (A.43)–(A.44) and the fact that the probability density function is nonnegative yields that

$$\begin{aligned} F_{>0}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) &\leq \frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}, \\ F_{<0}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) &= \frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(1)}, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}, \end{aligned}$$

which entail that

$$
\begin{aligned}
&\mathbb{P}(W_j > 0 | \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= F_{>0}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&\leq \frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}, \\
&\mathbb{P}(W_j < 0 | \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= F_{<0}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) \\
&= \frac{f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(1)}, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}{f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y})}.
\end{aligned}
\tag{A.45}
$$

Hence, a combination of (A.45) and Condition 1 (which ensures that the denominator is nonzero) establishes the result in (A.41).

It remains to prove (A.43). To this end, observe that for any Borel sets $A_1 \in \mathcal{R}^n$, $A_2 \in \mathcal{R}^n$, $A_3 \in \mathcal{R}^{n(p-1)}$, $A_4 \in \mathcal{R}^{n(p-1)}$, and $A_5 \in \mathcal{R}^n$, it holds that

$$
\begin{aligned}
&\int_{\vec{z} \in A_1 \times \cdots \times A_5} F_{>0}(\vec{z}) f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z}) d\vec{z} \\
&= \mathbb{P}(W_j > 0, \boldsymbol{X}_j \in A_1, \widetilde{\boldsymbol{X}}_j \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5) \\
&= \mathbb{P}(W_j > 0, \boldsymbol{X}_j^{(0)} \in A_1, \boldsymbol{X}_j^{(1)} \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5) \\
&= \int_{\{\boldsymbol{X}_j^{(0)} \in A_1, \boldsymbol{X}_j^{(1)} \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5\}} \mathbb{P}(W_j > 0 \big| \ \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) d\mathbb{P},
\end{aligned}
\tag{A.46}
$$

where the second equality above holds by the definitions of $\boldsymbol{X}_j^{(0)}$ and $\boldsymbol{X}_j^{(1)}$. Similarly, for any Borel sets $A_1 \in \mathcal{R}^n$, $A_2 \in \mathcal{R}^n$, $A_3 \in \mathcal{R}^{n(p-1)}$, $A_4 \in \mathcal{R}^{n(p-1)}$, and $A_5 \in \mathcal{R}^n$, we have

$$\begin{aligned}
&\int_{\vec{z} \in A_1 \times \cdots \times A_5} F_{<0}(\vec{z}) f_{\boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z}) d\vec{z} \\
&= \mathbb{P}(W_j < 0, \widetilde{\boldsymbol{X}}_j \in A_1, \boldsymbol{X}_j \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5) \\
&= \mathbb{P}(W_j < 0, \boldsymbol{X}_j^{(0)} \in A_1, \boldsymbol{X}_j^{(1)} \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5) \\
&= \int_{\{\boldsymbol{X}_j^{(0)} \in A_1, \boldsymbol{X}_j^{(1)} \in A_2, \boldsymbol{X}_{-j} \in A_3, \widetilde{\boldsymbol{X}}_{-j} \in A_4, \boldsymbol{Y} \in A_5\}} \mathbb{P}(W_j < 0 \big| \, \boldsymbol{X}_j^{(0)}, \boldsymbol{X}_j^{(1)}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}) d\mathbb{P},
\end{aligned} \tag{A.47}$$

where the second equality above holds by the definitions of $\boldsymbol{X}_j^{(0)}$ and $\boldsymbol{X}_j^{(1)}$. With the aid of (A.46)–(A.47), we can resort to the $\pi - \lambda$ Theorem [16] and the definition of the conditional expectation to obtain (A.43). Thus, we have established (A.41).

Proof of (A.7). We now aim to show (A.7) under Condition 2 and the assumption that $\boldsymbol{X}_j$ is independent of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}_{-j}$ for each $j \in \mathcal{H}_0$. First, in light of Condition 2 and Definition 2 of the knockoff generator, we see that $\boldsymbol{Y}$ is independent of $\widetilde{\boldsymbol{X}}$ conditional on $\boldsymbol{X}$. Next, we can deduce that

$$\begin{aligned}
&f_{\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y}}(\vec{z}_1, \vec{z}_2, \vec{z}_3, \vec{z}_4, \vec{z}_5) \\
&= f_{\boldsymbol{X}}(\vec{z}_1, \vec{z}_3) \times f_{\widetilde{\boldsymbol{X}}_j, \widetilde{\boldsymbol{X}}_{-j}, \boldsymbol{Y} | \boldsymbol{X}}(\vec{z}_2, \vec{z}_4, \vec{z}_5 | \vec{z}_1, \vec{z}_3) \\
&= f_{\boldsymbol{X}}(\vec{z}_1, \vec{z}_3) \times f_{\boldsymbol{Y} | \boldsymbol{X}}(\vec{z}_5 | \vec{z}_1, \vec{z}_3) \times f_{\widetilde{\boldsymbol{X}} | \boldsymbol{X}}(\vec{z}_2, \vec{z}_4 | \vec{z}_1, \vec{z}_3) \\
&= f_{\boldsymbol{X}, \widetilde{\boldsymbol{X}}}(\vec{z}_1, \vec{z}_3, \vec{z}_2, \vec{z}_4) \times f_{\boldsymbol{Y} | \boldsymbol{X}}(\vec{z}_5 | \vec{z}_1, \vec{z}_3) \\
&= f_{\boldsymbol{X}, \widetilde{\boldsymbol{X}}}(\vec{z}_1, \vec{z}_3, \vec{z}_2, \vec{z}_4) \times f_{\boldsymbol{Y} | \boldsymbol{X}_{-j}}(\vec{z}_5 | \vec{z}_3),
\end{aligned} \tag{A.48}$$

where the second equality above follows from the conditional independence property of the knockoffs, and the last equality above holds because of the column-wise conditional independence assumption on the null features.

From (A.48), it holds that

$$\begin{aligned}&f_{\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y}}(\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})\\&=f_{\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j}}(\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j})\times f_{\boldsymbol{Y}|\boldsymbol{X}_{-j}}(\boldsymbol{Y}|\boldsymbol{X}_{-j})\end{aligned}$$

and

$$\begin{aligned}&f_{\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y}}(\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j},\boldsymbol{Y})\\&=f_{\boldsymbol{X}_j,\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j}}(\widetilde{\boldsymbol{X}}_j,\boldsymbol{X}_j,\boldsymbol{X}_{-j},\widetilde{\boldsymbol{X}}_{-j})\times f_{\boldsymbol{Y}|\boldsymbol{X}_{-j}}(\boldsymbol{Y}|\boldsymbol{X}_{-j}),\end{aligned}$$

which establish (A.7).

Proofs of (A.8) *and* (A.9). The proof of (A.8) is straightforward using the additional assumption of i.i.d. observations and hence, is omitted here for simplicity. We now focus on proving (A.9). Fixing a feature index j, let us consider a random vector $(\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}})$ such that $\widetilde{X}_j^\dagger$ is generated by the jth coordinatewise knockoff generator $\kappa_j(\boldsymbol{x}_{-j},)$ given $\boldsymbol{x}_{-j}$ and that $\widetilde{\boldsymbol{z}} = (\widetilde{Z}_j, \widetilde{\boldsymbol{z}}_{-j})$ is a knockoff vector of $(\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j})$ generated from the knockoff filter $\kappa((\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}),)$, where κ_j and κ are as given in Condition 3. In view of Condition 3, we see that $(\widetilde{X}_j^\dagger, \widetilde{Z}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ and $(\widetilde{Z}_j, \widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ have the same distribution and the corresponding density functions exist, which entail that for each $(z_1, z_2, \vec{z}_3, \vec{z}_4) \in \mathbb{R}^{2p}$,

$$f_{\widetilde{X}_j^\dagger,\widetilde{Z}_j,\boldsymbol{x}_{-j},\widetilde{\boldsymbol{z}}_{-j}}(z_1, z_2, \vec{z}_3, \vec{z}_4) = f_{\widetilde{X}_j^\dagger,\widetilde{Z}_j,\boldsymbol{x}_{-j},\widetilde{\boldsymbol{z}}_{-j}}(z_2, z_1, \vec{z}_3, \vec{z}_4). \tag{A.49}$$

Next, it follows from the definition of $(\widetilde{X}_j^\dagger, \widetilde{Z}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j})$ that

$$\begin{aligned}f_{\widetilde{X}_j^\dagger,\widetilde{Z}_j,\boldsymbol{x}_{-j},\widetilde{\boldsymbol{z}}_{-j}}(z_1, z_2, \vec{z}_3, \vec{z}_4) &= f_{\boldsymbol{x}_{-j}}(\vec{z}_3) f_{\widetilde{X}_j^\dagger|\boldsymbol{x}_{-j}}(z_1|\vec{z}_3) f_{\widetilde{Z}_j,\widetilde{\boldsymbol{z}}_{-j}|\widetilde{X}_j^\dagger,\boldsymbol{x}_{-j}}(z_2, \vec{z}_4|z_1, \vec{z}_3)\\&= f_{\boldsymbol{x}_{-j}}(\vec{z}_3) f_{\widetilde{X}_j^\dagger|\boldsymbol{x}_{-j}}(z_1|\vec{z}_3) f_{\widetilde{X}_j,\widetilde{\boldsymbol{x}}_{-j}|X_j,\boldsymbol{x}_{-j}}(z_2, \vec{z}_4|z_1, \vec{z}_3),\end{aligned} \tag{A.50}$$

where $f_{\widetilde{Z}_j,\widetilde{\boldsymbol{z}}_{-j}|\widetilde{X}_j^\dagger,\boldsymbol{x}_{-j}}(z_2, \vec{z}_4|z_1, \vec{z}_3) = f_{\widetilde{X}_j,\widetilde{\boldsymbol{x}}_{-j}|X_j,\boldsymbol{x}_{-j}}(z_2, \vec{z}_4|z_1, \vec{z}_3)$ because a knockoff generator outputs random vectors with the same distribution if the input values are the same due to Definition 2. Similarly, it holds that

$$\begin{aligned} f_{\widetilde{X}_j^\dagger, \widetilde{Z}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{z}}_{-j}}(z_2, z_1, \vec{z}_3, \vec{z}_4) &= f_{\boldsymbol{x}_{-j}}(\vec{z}_3) f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(z_2 | \vec{z}_3) f_{\widetilde{Z}_j, \widetilde{\boldsymbol{z}}_{-j} | \widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(z_1, \vec{z}_4 | z_2, \vec{z}_3) \\ &= f_{\boldsymbol{x}_{-j}}(\vec{z}_3) f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(z_2 | \vec{z}_3) f_{\widetilde{X}_j, \widetilde{\boldsymbol{x}}_{-j} | X_j, \boldsymbol{x}_{-j}}(z_1, \vec{z}_4 | z_2, \vec{z}_3). \end{aligned} \tag{A.51}$$

From (A.49)–(A.51), we can deduce that

$$f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(z_1 | \vec{z}_3) f_{\widetilde{X}_j, \widetilde{\boldsymbol{x}}_{-j} | X_j, \boldsymbol{x}_{-j}}(z_2, \vec{z}_4 | z_1, \vec{z}_3) = f_{\widetilde{X}_j^\dagger | \boldsymbol{x}_{-j}}(z_2 | \vec{z}_3) f_{\widetilde{X}_j, \widetilde{\boldsymbol{x}}_{-j} | X_j, \boldsymbol{x}_{-j}}(z_1, \vec{z}_4 | z_2, \vec{z}_3),$$

which results in

$$\frac{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(z_1, z_2, \vec{z}_3, \vec{z}_4)}{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(z_2, z_1, \vec{z}_3, \vec{z}_4)} = \frac{f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(z_2, \vec{z}_3) \times f_{X_j, \boldsymbol{x}_{-j}}(z_1, \vec{z}_3)}{f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(z_1, \vec{z}_3) \times f_{X_j, \boldsymbol{x}_{-j}}(z_2, \vec{z}_3)}. \tag{A.52}$$

Setting $(z_1, z_2, \vec{z}_3, \vec{z}_4) = (X_{ij}, \widetilde{X}_{ij}, \boldsymbol{x}_{-ij}, \widetilde{\boldsymbol{x}}_{-ij})$ in (A.52) above, we can obtain that

$$\begin{aligned} &\sum_{t=1}^{n} \log \left(\frac{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(X_{tj}, \widetilde{X}_{tj}, \boldsymbol{x}_{-tj}, \widetilde{\boldsymbol{x}}_{-ij})}{f_{X_j, \widetilde{X}_j, \boldsymbol{x}_{-j}, \widetilde{\boldsymbol{x}}_{-j}}(\widetilde{X}_{tj}, X_{tj}, \boldsymbol{x}_{-tj}, \widetilde{\boldsymbol{x}}_{-tj})} \right) \\ &= \sum_{t=1}^{n} \log \left(\frac{f_{X_j, \boldsymbol{x}_{-j}}(X_{ij}, \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(\widetilde{X}_{tj}, \boldsymbol{x}_{-tj})}{f_{X_j, \boldsymbol{x}_{-j}}(\widetilde{X}_{ij}, \boldsymbol{x}_{-tj}) f_{\widetilde{X}_j^\dagger, \boldsymbol{x}_{-j}}(X_{tj}, \boldsymbol{x}_{-tj})} \right), \end{aligned}$$

which establishes (A.9). This concludes the proof of Theorem 3.

B.4 Proof of Corollary 1

The conclusion of Corollary 1 follows from the proof of (A.9) in Theorem 3 given in Section B.3.

B.5 Proof of Corollary 2

Under Condition 2 and the assumptions that $\{\boldsymbol{x}_t\}_{t\geq 1}$ satisfies Condition 4 with h-step and constants $C_0 > 0$ and $0 \leq \rho < 1$ and Y_t is $\boldsymbol{x}_{t+1}$-measurable, we will prove in Section C.1 that for each $1 \leq k \leq q+1$,

$$\sup_{\mathcal{D}\in\mathcal{R}^{\#H_k\times(2p+1)}} |\mathbb{P}(\mathcal{X}_k \in \mathcal{D}) - \mathbb{P}(\mathcal{X}_k^{\pi} \in \mathcal{D})| \leq \#H_k \times \rho^q \times C_0, \tag{A.53}$$

where we recall that $\mathcal{X}_k = \{\boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t, Y_t\}_{t\in H_k}$ and $\mathcal{X}_k^{\pi} = \{\boldsymbol{x}_t^{\pi}, \widetilde{\boldsymbol{x}}_t^{\pi}, Y_t^{\pi}\}_{t\in H_k}$ for each $k \in \{1, \cdots, q+1\}$. Combining (A.53) and $\sum_{k=1}^{q+1} \#H_k \leq n$ leads to the desired result in (9).

Next, we deal with the second assertion of Corollary 2. When Condition 2 holds and $\{Y_t, \boldsymbol{x}_t\}_{t=1}^n$ is also an i.i.d. sample, $\{Y_t, \boldsymbol{x}_t, \widetilde{\boldsymbol{x}}_t\}_{t=1}^n$ is an i.i.d. sample. Therefore, it follows from the fact that $(Y_t^{\pi}, \boldsymbol{x}_i^{\pi}, \widetilde{\boldsymbol{x}}_t^{\pi})$'s are i.i.d. with $(Y_t^{\pi}, \boldsymbol{x}_t^{\pi}, \widetilde{\boldsymbol{x}}_t^{\pi})$ having the same distribution as $(Y_1, \boldsymbol{x}_1, \widetilde{\boldsymbol{x}}_1)$ that (A.53) holds with $\rho = 0$, which concludes the proof of Corollary 2.

B.6 Proof of Proposition 1

For the reader's convenience, we provide some basic knowledge about time-homogeneous Markov chains here. Two sufficient conditions for a process $\{\boldsymbol{Q}_t\}$ to admit a transition kernel are 1) for each Borel set A and each t,

$$\mathbb{P}(\boldsymbol{Q}_t \in A|\boldsymbol{Q}_{t-1}) = \mathbb{P}(\boldsymbol{Q}_t \in A|\boldsymbol{Q}_{t-j}, j < 1),$$

the so-called the "Markov property;" and 2) the conditional distribution of $\boldsymbol{Q}_t$ given $\boldsymbol{Q}_{t-1}$ are the same for each t. Processes satisfying these two conditions are known as time-homogeneous Markov chains. A useful sufficient condition for verifying that a process is a time-homogeneous Markov chain is to check whether the process can be written as $\boldsymbol{Q}_t = F(\boldsymbol{Q}_{t-1}, \boldsymbol{\varepsilon}_t)$ for some measurable $F(\cdot,\cdot)$ and identically distributed innovative random vectors $\{\boldsymbol{\varepsilon}_t\}$ such that $\boldsymbol{\varepsilon}_t$ is independent of $\boldsymbol{Q}_{t-j}$ with $j \geq 1$. It can be shown that $\{\boldsymbol{x}_t\}$ in Example 3 is a time-homogeneous Markov chain, and we omit the details on proving such claim for simplicity.

Next, let us consider Example 6 below, which is more general than Example 3. In particular, $\{\boldsymbol{x}_t\}$ in Example 3 is a special case of $\{\boldsymbol{z}_t\}$ in Example 6.

Example 6 (Gaussian linear processes). *Let $\{\boldsymbol{z}_t := (Y_{t1}, \cdots, Y_{tp})^T\}$ be such that for $l = 1, \cdots, p$, $Y_{tl} = \sum_{i=0}^{\infty}(\vec{w}_i(l))^T \boldsymbol{\delta}_{t-i}$, where $\vec{w}_i(l)$ is an ι-dimensional coefficient vector such that for each $h \geq 0$,*

$$\max_{1 \leq l \leq p} \sum_{i \geq h} \|\vec{w}_i(l)\|_1 \leq C_1 e^{-s_1 h} \tag{A.54}$$

with some positive C_1 and s_1, and $\boldsymbol{\delta}_t$'s are i.i.d. ι-dimensional Gaussian random vectors with zero mean and covariance matrix Σ. In addition, assume that $\lambda_{\max}(\Sigma) < L_3$ and $\lambda_{\min}(\mathbb{E}(\boldsymbol{z}_1 \boldsymbol{z}_1^T)) > l_1$ for some positive L_3 and l_1, where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of a given matrix, respectively.

We use Propositions 3.1.1–3.1.2 of Brockwell and Davis [11] to obtain the stationarity of Example 6; the details on this are omitted. Thus, $\{\boldsymbol{x}_t\}$ in Example 3 is a stationary time-homogeneous Markov chain; equivalently, the stationary distribution and transition kernel of $\{\boldsymbol{x}_t\}$ exist.

Remark 3. *Notice that time-homogeneous Markov chains are not always stationary; particularly, a random walk process can be a time-homogeneous Markov chain. Also, note that Example 6 may not be a time-homogeneous Markov chain. With regularity conditions assumed, for a stationary $Q_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}$ to be a time-homogeneous Markov chain, it is usually required that Q_t can be written as $Q_t = \sum_{j=1}^{k} \gamma_j Q_{t-j} + \varepsilon_t$ for some positive integer k. Without further assumptions, the linear processes in Example 6 may not admit such representation.*

Let $\{\boldsymbol{z}_t^{(h)}\}$ in Example 6 with dimensionality p_h and stationary distribution $\pi_h(\cdot)$ be given. Note that we do not assume a transition kernel for Example 6 and that all parameters (except for constant C_1, s_1, l_1, and L_3) in Example 6 may change for each h, but we drop the superscript or subscript h for simplicity of presentation whenever there is no confusion. Since Example 3 admits a transition kernel and it is a special case of Example 6, to prove Proposition 1 it suffices to show that for all large h, there exist some constants $0 \leq \rho < 1$, $0 < C_0 < \infty$, and measurable functions $V_h : \mathbb{R}^{p_h} \longrightarrow [0, \infty)$ such that for each integer t,

$$\sup_{\mathcal{D} \in \mathcal{R}^{p_h}} \left| \mathbb{P}(\boldsymbol{z}_{t+h}^{(h)} \in \mathcal{D}) - \mathbb{P}(\boldsymbol{z}_{t+h}^{(h)} \in \mathcal{D} \mid \boldsymbol{z}_t^{(h)}) \right| \leq V_h(\boldsymbol{z}_t^{(h)}) \rho^h C_3 \tag{A.55}$$

almost surely for some constant $C_3 > 0$, and

$$C_0 \geq \sup_{h>0} \int_{\mathbb{R}^{p_h}} V_h(x) \pi_h(dx).$$

To facilitate the technical presentation, we first introduce some necessary notations. For each h, denote by

$$\begin{aligned} U_{1t}^{(h)} &:= \left(\sum_{i=h}^{\infty} (\vec{w}_i(1))^{\scriptscriptstyle T} \boldsymbol{\delta}_{t-i}, \cdots, \sum_{i=h}^{\infty} (\vec{w}_i(p_h))^{\scriptscriptstyle T} \boldsymbol{\delta}_{t-i} \right)^{\scriptscriptstyle T}, \\ U_{2t}^{(h)} &:= \left(\sum_{i=0}^{h-1} (\vec{w}_i(1))^{\scriptscriptstyle T} \boldsymbol{\delta}_{t-i}, \cdots, \sum_{i=0}^{h-1} (\vec{w}_i(p_h))^{\scriptscriptstyle T} \boldsymbol{\delta}_{t-i} \right)^{\scriptscriptstyle T}, \end{aligned} \tag{A.56}$$

and let $V_{1t}^{(h)}$ and $V_{2t}^{(h)}$ be independent copies of $U_{1t}^{(h)}$ and $U_{2t}^{(h)}$, respectively, where the superscript or subscript h represents the truncation length. Observe that $U_{1t}^{(h)} + U_{2t}^{(h)}$ is an instance of $\boldsymbol{z}_t$ in Example 6. Due to the Gaussian innovations, the stationary distribution π_h is the distribution of $V_{11}^{(0)}$, which is the same as that of $V_{1t}^{(h)} + V_{2t}^{(h)}$ for each t and h.

Let us repeat the needed statement (A.55) with the newly defined notation. For all large h, there exist some constants $0 \leq \rho < 1$, $0 < C_0 < \infty$, and measurable functions $V_h : \mathbb{R}^{p_h} \longrightarrow [0, \infty)$ such that for each integer t,

$$\begin{aligned}\sup_{\mathcal{D} \in \mathcal{R}^{p_h}} &\left| \mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}) - \mathbb{P}(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \mid U_{1t}^{(h)} + U_{2t}^{(h)}) \right| \\ &\leq V_h(U_{1t}^{(h)} + U_{2t}^{(h)}) \rho^h C_3\end{aligned} \tag{A.57}$$

almost surely for some constant $C_3 > 0$, and

$$C_0 \geq \sup_{h>0} \int_{\mathbb{R}^{p_h}} V_h(x) \pi_h(dx). \tag{A.58}$$

If (A.57) holds for some t, it holds for each integer t because the process is stationary. Notice that the technical analysis here does not depend on the Markov property. For the remaining proof of Proposition 1, we tend to omit the term almost surely when the equality or inequality holds clearly almost surely.

Let us begin with establishing (A.57). In view of assumption (10), let $s_3 > 0$ and $0 < \delta_0 < 1$ be given such that $0 < s_3 < s_1$ and $s_2 < \delta_0 s_3$. For each positive integer h, we have

$$p_h \exp\left(-\delta_0 s_3 h\right) \leq C_2 \exp\left((s_2 - \delta_0 s_3) h\right). \tag{A.59}$$

We claim that for all large h and each t, it holds that for each $\mathcal{D} \in \mathcal{R}^{p_h}$,

$$\begin{aligned}&\left| \mathbb{P}\left(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\right) - \mathbb{P}\left(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right) \right| \\ &\leq \mathbb{P}\left(\left\| V_{11}^{(h)} \right\|_\infty \geq e^{-s_3 h}\right) + \mathbb{P}(\left\| U_{1(t+h)}^{(h)} \right\|_\infty \geq e^{(-s_3 h)} \mid U_{1t}^{(h)} + U_{2t}^{(h)}) \\ &\quad + 2\mathbb{P}\left(\left\| V_{21}^{(h)} \right\|_\infty \geq e^{(1-\delta_0)s_3 h} - 2e^{-s_3 h}\right) + \frac{8 p_h}{\underline{c}} e^{-\delta_0 s_3 h},\end{aligned} \tag{A.60}$$

where $\underline{c} > 0$ is some constant and $\|\vec{z}\|_\infty := \max_{1 \leq i \leq k} |z_i|$ for $\vec{z} = (z_1, \cdots, z_k)^T \in \mathbb{R}^k$. The proof of claim (A.60) above is presented in Section C.2.

We next construct some upper bounds for the first and third terms on the RHS of

(A.60). It follows from $\|\cdot\|_1^2 \geq \|\cdot\|_2^2$ and (A.54) that for each q and t,

$$\begin{aligned}\operatorname{Var}(\sum_{i\geq h} \vec{w}_i^T(q)\boldsymbol{\delta}_{t-i}) &\leq \sum_{i\geq h} \|\vec{w}_i(q)\|_2^2\, \lambda_{\max}(\Sigma_h) \\ &\leq \left(\sum_{i\geq h} \|\vec{w}_i(q)\|_1\right)^2 \lambda_{\max}(\Sigma_h) \\ &\leq \left(C_1 \exp\left(-s_1 h\right)\right)^2 \lambda_{\max}(\Sigma_h),\end{aligned}$$

where Σ_h denotes the covariance matrix of the underlying Gaussian random vectors $\boldsymbol{\delta}_t$'s (the superscript h is dropped) associated with $\{\boldsymbol{z}_t^{(h)}\}$ in Example 6. Combining this, the fact that $\boldsymbol{\delta}_t$'s are Gaussian random vectors, and Markov's inequality, it holds that for each $h \geq 0$,

$$\begin{aligned}&\mathbb{P}\left(\left\|V_{11}^{(h)}\right\|_\infty \geq \exp\left(-s_3 h\right)\right) \\ &\quad\leq p_h \mathbb{P}\left(C_1 \exp\left(-s_1 h\right)\sqrt{\lambda_{\max}(\Sigma_h)}|Z| \geq \exp\left(-s_3 h\right)\right) \\ &\quad\leq p_h \mathbb{E}(e^{|Z|}) \exp\left[-\left(C_1\sqrt{\lambda_{\max}(\Sigma_h)}\right)^{-1} \exp\left((s_1 - s_3)h\right)\right],\end{aligned} \tag{A.61}$$

where Z denotes a Gaussian random variable with zero mean and unit variance. Similarly, we can show that for all large h, it holds that

$$\begin{aligned}&\mathbb{P}\left(\left\|V_{21}^{(h)}\right\|_\infty \geq \exp\left((1-\delta_0)s_3 h\right) - 2\exp\left(-s_3 h\right)\right) \\ &\quad\leq p_h \mathbb{P}\left(C_1\sqrt{\lambda_{\max}(\Sigma_h)}|Z| \geq \exp\left((1-\delta_0)s_3 h\right) - 1\right) \\ &\quad\leq p_h \mathbb{E}(e^{|Z|}) \exp\left[-\left(C_1\sqrt{\lambda_{\max}(\Sigma_h)}\right)^{-1} \left(\exp\left((1-\delta_0)s_3 h\right) - 1\right)\right].\end{aligned} \tag{A.62}$$

We are now ready to construct the V_h function. Let g be a measurable function such that $g(U_{1t}^{(h)} + U_{2t}^{(h)})$ is a version of $\mathbb{P}(\left\|U_{1(t+h)}^{(h)}\right\|_\infty \geq e^{(-s_3 h)} \mid U_{1t}^{(h)} + U_{2t}^{(h)})$. It follows from the assumption that $\lambda_{\max}(\Sigma_h)$ is bounded by a constant, $\mathbb{E}(e^{|Z|}) < \infty$, (A.59), (A.61), and (A.62) that there exist some constants $C_3 > 0$ and $0 \leq \rho < 1$ such that for each positive

h, $C_3\rho^h$ is larger than the summation of the first, third, and fourth terms on the RHS of (A.60). For each h, let us define function V_h as

$$V_h(x) := \begin{cases} 2 \quad \text{if } g(x) \leq C_3\rho^h, \\ \\ 2\rho^{-h}C_3^{-1} \quad \text{otherwise.} \end{cases}$$

Then combining (A.60) and the definitions of ρ, C_3, and V_h leads to (A.57).

Finally, we deal with (A.58). By Markov's inequality, the definition of g, and $\mathbb{P}(\left\|V_{11}^{(h)}\right\|_\infty \geq \exp(-s_3h)) = \mathbb{P}(\left\|U_{1(t+h)}^{(h)}\right\|_\infty \geq \exp(-s_3h))$, we can deduce that

$$\begin{aligned} \int V_h(x)d\pi_h(x) &= \mathbb{E}(V_h(U_{1t}^{(h)} + U_{2t}^{(h)})) \\ &\leq 2 + 2(C_3\rho^h)^{-1}\mathbb{P}(g(U_{1t}^{(h)} + U_{2t}^{(h)}) > C_3\rho^h) \\ &\leq 2 + 2(C_3\rho^h)^{-2}\mathbb{E}(g(U_{1t}^{(h)} + U_{2t}^{(h)})) \\ &= 2 + 2(C_3\rho^h)^{-2}\mathbb{P}\left(\left\|V_{11}^{(h)}\right\|_\infty \geq \exp(-s_3h)\right). \end{aligned} \tag{A.63}$$

For the first equality, recall that $U_{1t}^{(h)} + U_{2t}^{(h)}$ has the stationary distribution. Therefore, by (A.63), (A.61), and (A.59), for all large h, it holds that $\int V_h(x)d\pi_h(x)$ is bounded by a constant, which leads to (A.58). This completes the proof of Proposition 1.

C Some key lemmas and additional technical details

In this section, we will provide additional technical details and some key lemmas. In particular, we rely on measure theory for valid arguments for the manipulation of integration when conditional distributions are involved.

C.1 Proof of Claim (A.53)

Let us first make a simple observation. For any q_1-dimensional random vectors X_1, Y_1 and q_2-dimensional random vectors X_2, Y_2 such that $X_2 = F(X_1)$ and $Y_2 = F(Y_1)$ for some measurable $F : \mathbb{R}^{q_1} \longmapsto \mathbb{R}^{q_2}$, it holds that

$$\begin{aligned}\sup_{\mathcal{D}\in\mathcal{R}^{q_2}} |\mu_{X_2}(\mathcal{D}) - \mu_{Y_2}(\mathcal{D})| &= \sup_{\mathcal{D}\in\mathcal{R}^{q_2}} |\mu_{X_1}(F^{-1}(\mathcal{D})) - \mu_{Y_1}(F^{-1}(\mathcal{D}))| \\ &\leq \sup_{\mathcal{A}\in\mathcal{R}^{q_1}} |\mu_{X_1}(\mathcal{A}) - \mu_{Y_1}(\mathcal{A})|,\end{aligned} \tag{A.64}$$

where $F^{-1}(\cdot)$ denotes the inverse mapping of $F(\cdot)$. With the aid of (A.64), we now deal with the case of $k = 1$ below. Let $\boldsymbol{M}$ be given as in (A.86) with $l = 2$, $h = q - 1$, and $\boldsymbol{z}_i = \boldsymbol{x}_i$. Similarly, let $\boldsymbol{M}^\pi$ be given as in (A.86) with $l = 2$, $h = q - 1$, and i.i.d. random vectors $(\boldsymbol{z}_1^{\pi_i}, \boldsymbol{z}_2^{\pi_i})$'s such that $(\boldsymbol{z}_1^{\pi_1}, \boldsymbol{z}_2^{\pi_1})$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ have the same distribution. Then it follows from Lemma 1 in Section C.3 that

$$\begin{aligned}&\sup_{\mathcal{D}\in\mathcal{R}^{\#H_1\times(2p)}} |\mathbb{P}((\boldsymbol{x}_{i+1}, \boldsymbol{x}_i, i \in H_1) \in \mathcal{D}) - \mathbb{P}((\boldsymbol{z}_2^{\pi_i}, \boldsymbol{z}_1^{\pi_i}, i \in H_1) \in \mathcal{D})| \\ &\leq \#H_1 \times \rho^q \times C_0.\end{aligned} \tag{A.65}$$

By the assumption that Y_i is $\boldsymbol{x}_{i+1}$-measurable, we have that $(Y_i, \boldsymbol{x}_i) = F(\boldsymbol{x}_{i+1}, \boldsymbol{x}_i)$ for some measurable $F : \mathbb{R}^{2p} \longmapsto \mathbb{R}^{1+p}$. Then it follows from the assumption that each $(Y_i^\pi, \boldsymbol{x}_i^\pi)$ and $(Y_1, \boldsymbol{x}_1)$ have the same distribution and the assumption that each $(\boldsymbol{z}_1^{\pi_i}, \boldsymbol{z}_2^{\pi_i})$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ have the same distribution that

$$\{F(\boldsymbol{z}_2^{\pi_i}, \boldsymbol{z}_1^{\pi_i})\}_{i=1}^n \quad \text{and} \quad \{(Y_i^\pi, \boldsymbol{x}_i^\pi)\}_{i=1}^n \tag{A.66}$$

have the same distribution. Hence, from (A.64)–(A.66), we can deduce that

$$\begin{aligned}
&\sup_{\mathcal{D}\in\mathcal{R}^{\#H_1\times(1+p)}} |\mathbb{P}((Y_i,\boldsymbol{x}_i,i\in H_1)\in\mathcal{D})-\mathbb{P}((Y_i^\pi,\boldsymbol{x}_i^\pi,i\in H_1)\in\mathcal{D})| \\
&=\sup_{\mathcal{D}\in\mathcal{R}^{\#H_1\times(1+p)}} |\mathbb{P}((\boldsymbol{x}_{i+1},\boldsymbol{x}_i,i\in H_1)\in F^{-1}(\mathcal{D}))-\mathbb{P}((\boldsymbol{z}_2^{\pi_i},\boldsymbol{z}_1^{\pi_i},i\in H_1)\in F^{-1}(\mathcal{D}))| \quad \text{(A.67)}\\
&\le \#H_1\times\rho^q\times C_0.
\end{aligned}$$

Finally, we can apply Lemma 7 in Section C.9 to control the distributional variation results from the inclusion of knockoffs. Using Definition 2 and Conditions 1–2, we can show that 1) $(Y_i,\boldsymbol{x}_i,\widetilde{\boldsymbol{x}}_i)$'s are identically distributed; 2) $(\widetilde{\boldsymbol{x}}_i,i\in H_1)$ are independent conditional on $(\boldsymbol{x}_i,i\in H_1)$, and that $\widetilde{\boldsymbol{x}}_i$ is independent of $(\boldsymbol{x}_q,q\in H_1\backslash\{i\})$ conditional on $\boldsymbol{x}_i$ for each $i\in H_1$; 3) $(Y_i,i\in H_1)$ is independent of $(\widetilde{\boldsymbol{x}}_i,i\in H_1)$ conditional on $(\boldsymbol{x}_i,i\in H_1)$; and 4) the above results also hold for $(Y_i^\pi,\boldsymbol{x}_i^\pi,\widetilde{\boldsymbol{x}}_i^\pi)$'s. By these results, an application of the first assertion of Lemma 7 concludes the proof of (A.53) for the case with $k=1$. The other cases with $2\le k\le q+1$ can be dealt with similarly. This concludes the proof of Claim (A.53).

C.2 Proof of Claim (A.60)

We denote by $f_h(x)$ the density function of $V_{21}^{(h)}$; that is,

$$f_h(x):=\frac{1}{\sqrt{(2\pi)^{p_h}|\Sigma_h^V|}}\exp\left(-\frac{1}{2}x^T(\Sigma_h^V)^{-1}x\right),$$

where Σ_h^V is the corresponding covariance matrix and $|A|$ stands for the determinant of a given matrix A. By the assumptions of Gaussian linear processes (this is where we need $\lambda_{\min}(\mathbb{E}(\boldsymbol{z}_1^{(h)}(\boldsymbol{z}_1^{(h)})^T))>l_1$), there exist some constant $\underline{c}>0$ and positive integer $\bar{h}$ such that

$$\min_{k\ge\bar{h}}\lambda_{\min}(\mathbb{E}(V_{21}^{(k)}V_{21}^{(k)T}))>\underline{c}>0. \quad \text{(A.68)}$$

In view of (A.68), for all $h \geq \bar{h}$ we have $\lambda_{\min}(\Sigma_h^V) > \underline{c}$.

To support the technical analysis, we will make use of the following facts.

1) For all $x < 1.79$, it holds that

$$\exp(x) \leq 1 + x + x^2. \tag{A.69}$$

To get the specific value of 1.79, we use the first and second order derivatives of $\exp(x)$ and $1 + x + x^2$ to conclude that there exists some positive number x_0 such that when $x \leq x_0$, (A.69) holds, and when $x > x_0$, it holds that $\exp(x) > 1 + x + x^2$. Then a direct calculation shows that $\exp(1.79) < 5.994 < 1 + 1.79 + 1.79^2$, which gives $x_0 \geq 1.79$.

2) By (A.68), for all large h, we have that for each $\Delta, x \in \mathbb{R}^{p_h}$ with $\|\Delta\|_2 \leq \|x\|_2$,

$$|x^T(\Sigma_h^V)^{-1}\Delta + \frac{1}{2}\Delta^T(\Sigma_h^V)^{-1}\Delta| \leq 2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1}. \tag{A.70}$$

If furthermore $2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1} < 1.79$, then it follows from (A.69)–(A.70) that

$$\begin{aligned} |f_h(x + \Delta) - f_h(x)| &\leq f_h(x)\Big|\exp\left(-x^T(\Sigma_h^V)^{-1}\Delta - \frac{1}{2}\Delta^T(\Sigma_h^V)^{-1}\Delta\right) - 1\Big| \\ &\leq f_h(x)\big(2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1} + (2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1})^2\big). \end{aligned} \tag{A.71}$$

3) We show that for each $\mathcal{D} \in \mathcal{R}^{p_h}$, $\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)})$ is a version of $\mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)})$ and in particular, for each t, $h > 0$, and $\mathcal{D} \in \mathcal{R}^{p_h}$,

$$\mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)}) = \mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)}).$$

To this end, let us define a measurable function $g(x) := \int_{\mathbb{R}^{p_h}} \mu_{V_{21}^{(h)}}(dx_2)\mathbf{1}_{x_2 \in \mathcal{D} - x}$

with $\mathcal{D} - x \coloneqq \{z - x : z \in \mathcal{D}\}$ and write $\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)}) = g(V_{1t}^{(h)})$ to see that $\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)})$ is $\sigma(V_{1t}^{(h)})$-measurable. Observe that if we can show that for each $\mathcal{A} \in \mathcal{R}^{p_h}$,

$$\int_{\mathcal{A}} \mu_{V_{1t}^{(h)}}(dx_1)\mu_{V_{21}^{(h)}}(\mathcal{D} - x_1) = \int_{\{V_{1t}^{(h)} \in \mathcal{A}\}} \mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)})d\mathbb{P}, \tag{A.72}$$

then we can apply the change of variables formula to the left-hand side of (A.72) and use the definition of conditional expectation to obtain the desired result. It remains to prove (A.72).

Since $V_{1t}^{(h)}$ and $V_{2t}^{(h)}$ are independent for each t and $h > 0$, it holds that for each $\mathcal{D}, \mathcal{A} \in \mathcal{R}^{p_h}$,

$$\begin{aligned}
&\mathbb{P}(\{V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\} \cap \{V_{1t}^{(h)} \in \mathcal{A}\}) \\
&\quad= \int_{\mathbb{R}^{2p_h}} \mu_{V_{1t}^{(h)}, V_{2t}^{(h)}}(dx_1 \times dx_2)\mathbf{1}_{x_1 + x_2 \in \mathcal{D}}\mathbf{1}_{x_1 \in \mathcal{A}} \\
&\quad= \int_{\mathbb{R}^{2p_h}} \mu_{V_{1t}^{(h)}}(dx_1)\mu_{V_{2t}^{(h)}}(dx_2)\mathbf{1}_{x_1 + x_2 \in \mathcal{D}}\mathbf{1}_{x_1 \in \mathcal{A}} \\
&\quad= \int_{\mathbb{R}^{p_h}} \mathbf{1}_{(x_1 \in \mathcal{A})}\mu_{V_{1t}^{(h)}}(dx_1) \int_{\mathbb{R}^{p_h}} \mu_{V_{2t}^{(h)}}(dx_2)\mathbf{1}_{x_2 \in \mathcal{D} - x_1} \qquad \text{(A.73)} \\
&\quad= \int_{\mathbb{R}^{p_h}} \mathbf{1}_{(x_1 \in \mathcal{A})}\mu_{V_{1t}^{(h)}}(dx_1)\mu_{V_{2t}^{(h)}}(\mathcal{D} - x_1) \\
&\quad= \int_{\mathbb{R}^{p_h}} \mathbf{1}_{(x_1 \in \mathcal{A})}\mu_{V_{1t}^{(h)}}(dx_1)\mu_{V_{21}^{(h)}}(\mathcal{D} - x_1) \\
&\quad= \int_{\mathcal{A}} \mu_{V_{1t}^{(h)}}(dx_1)\mu_{V_{21}^{(h)}}(\mathcal{D} - x_1),
\end{aligned}$$

where the second equality above is due to independence. Moreover, since $\mathbb{P}(\{V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\} \cap \{V_{1t}^{(h)} \in \mathcal{A}\} \mid V_{1t}^{(h)}) = \mathbf{1}_{V_{1t}^{(h)} \in \mathcal{A}}\mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)})$, by the law of total expectation we have that

$$\begin{aligned}
&\mathbb{P}(\{V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\} \cap \{V_{1t}^{(h)} \in \mathcal{A}\}) \\
&= \mathbb{E}(\mathbb{P}(\{V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\} \cap \{V_{1t}^{(h)} \in \mathcal{A}\} \mid V_{1t}^{(h)})) \\
&= \int_{\{V_{1t}^{(h)} \in \mathcal{A}\}} \mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)}) d\mathbb{P}.
\end{aligned} \tag{A.74}$$

Hence, combining (A.73)–(A.74) leads to (A.72).

4) Observe that $U_{2(t+h)}^{(h)}$ is independent of $U_{1t}^{(h)} + U_{2t}^{(h)}$ and $U_{1(t+h)}^{(h)}$. Thus, for each $\mathcal{D} \in \mathcal{R}^{p_h}$, we have that

$$\begin{aligned}
&\mathbb{P}(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \mid U_{1t}^{(h)} + U_{2t}^{(h)}, U_{1(t+h)}^{(h)}) \\
&= \mathbb{P}(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \mid U_{1(t+h)}^{(h)}).
\end{aligned}$$

Using such representation, similar arguments as in 3) above, and the fact that $\mu_{U_{2(t+h)}^{(h)}}$ is identical to $\mu_{V_{21}^{(h)}}$, we can show that

$$\mathbb{P}(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \mid U_{1t}^{(h)} + U_{2t}^{(h)}, U_{1(t+h)}^{(h)}) = \mu_{V_{21}^{(h)}}(\mathcal{D} - U_{1(t+h)}^{(h)}).$$

5) Denote by $Q := \left\{ \left\| V_{1t}^{(h)} \right\|_\infty \geq e^{(-s_3 h)} \right\}$ and $G := \left\{ \left\| U_{1(t+h)}^{(h)} \right\|_\infty \geq e^{(-s_3 h)} \right\}$. Then it follows from the definition of $\mu_{V_{21}^{(h)}}$ that for each $\mathcal{D} \in \mathcal{R}^{p_h}$,

$$\begin{aligned}
&\mathbf{1}_{Q^c \cap G^c} \left(\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)}) - \mu_{V_{21}^{(h)}}(\mathcal{D} - U_{1(t+h)}^{(h)}) \right) \\
&\leq \sup_{\|\Delta_i\|_\infty < e^{-s_3 h}} \Big| \int_{\mathcal{D} - \Delta_2} \big(f_h(x + \Delta_2 - \Delta_1) - f_h(x) \big) dx \Big|.
\end{aligned} \tag{A.75}$$

We are now ready to establish the desired upper bound. For each integer $h > 0$, t, and

$\mathcal{D} \in \mathcal{R}^{p_h}$, it holds that

$$\begin{aligned}
&\left|\mathbb{P}\left(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D}\right) - \mathbb{P}\left(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right)\right| \\
&= \left|\mathbb{E}\left[\mathbb{P}(V_{1t}^{(h)} + V_{2t}^{(h)} \in \mathcal{D} \mid V_{1t}^{(h)})\right]\right. \\
&\quad \left. - \mathbb{E}\left[\mathbb{P}(U_{1(t+h)}^{(h)} + U_{2(t+h)}^{(h)} \in \mathcal{D} \mid U_{1t}^{(h)} + U_{2t}^{(h)}, U_{1(t+h)}^{(h)}) \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right]\right|.
\end{aligned} \tag{A.76}$$

By 3) and 4) above, for each $\mathcal{D} \in \mathcal{R}^{p_h}$, we have

$$\text{RHS of } (A.76) = \left|\mathbb{E}\left[\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)})\right] - \mathbb{E}\left[\mu_{V_{21}^{(h)}}(\mathcal{D} - U_{1(t+h)}^{(h)}) \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right]\right|. \tag{A.77}$$

Since $V_{1t}^{(h)}$ is an independent copy, it follows that

$$\text{RHS of } (A.77) = \left|\mathbb{E}\left[\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)}) - \mu_{V_{21}^{(h)}}(\mathcal{D} - U_{1(t+h)}^{(h)}) \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right]\right|. \tag{A.78}$$

Next, we separate the expectation according to events Q and G as

$$\begin{aligned}
&\text{RHS of } (A.78) \\
&= \left|\mathbb{E}\left[(\mathbf{1}_{Q \cup G} + \mathbf{1}_{Q^c \cap G^c})\left(\mu_{V_{21}^{(h)}}(\mathcal{D} - V_{1t}^{(h)}) - \mu_{V_{21}^{(h)}}(\mathcal{D} - U_{1(t+h)}^{(h)})\right) \;\middle|\; U_{1t}^{(h)} + U_{2t}^{(h)}\right]\right|.
\end{aligned} \tag{A.79}$$

Then by (A.75) and some simple calculations, we can show that

$$\begin{aligned}
&\text{RHS of } (A.79) \\
&\leq \mathbb{P}(G \cup Q \mid U_{1t}^{(h)} + U_{2t}^{(h)}) + \sup_{\|\Delta_i\|_\infty < e^{-s_3 h}} \left|\int_{\mathcal{D}-\Delta_2} \big(f_h(x + \Delta_2 - \Delta_1) - f_h(x)\big)dx\right|.
\end{aligned} \tag{A.80}$$

For the first term on the RHS of (A.80), it follows from the definitions of Q and G and the stationarity of $V_{1t}^{(h)}$ that

$$\begin{aligned}\mathbb{P}(G \cup Q \mid U_{1t}^{(h)} + U_{2t}^{(h)}) \leq \mathbb{P}\left(\left\|V_{11}^{(h)}\right\|_\infty \geq e^{(-s_3 h)}\right) \\ + \mathbb{P}\Big(\left\|U_{1(t+h)}^{(h)}\right\|_\infty \geq e^{(-s_3 h)} \mid U_{1t}^{(h)} + U_{2t}^{(h)}\Big).\end{aligned} \tag{A.81}$$

For the second term on the RHS of (A.80), it holds that

$$\begin{aligned}&\sup_{\|\Delta_i\|_\infty < e^{-s_3 h}} \Big| \int_{\mathcal{D}-\Delta_2} \big(f_h(x+\Delta_2-\Delta_1) - f_h(x)\big) dx \Big| \\ &\leq 2\mathbb{P}\left(\left\|V_{21}^{(h)}\right\|_\infty \geq \exp\left((1-\delta_0)s_3 h\right) - 2\exp\left(-s_3 h\right)\right) \\ &\quad + \sup_{\|\Delta_i\|_\infty < e^{-s_3 h}} \Big| \int_{x \in \mathcal{D}-\Delta_2,\, \|x\|_\infty < e^{(1-\delta_0)s_3 h}} \big(f_h(x+\Delta_2-\Delta_1) - f_h(x)\big) dx \Big|.\end{aligned} \tag{A.82}$$

We proceed with dealing with the last term on the RHS of (A.82). Let x_1, x_2 denote two vectors in $\mathbb{R}^{p_h}$. Then, it follows from (A.59) that

$$\lim_{h\to\infty} \sup \sup_{\substack{\|x_2\|_2 \leq 2\sqrt{p_h} e^{-s_3 h} \\ \|x_1\|_2 \leq \sqrt{p_h} e^{(1-\delta_0)s_3 h}}} \|x_1\|_2 \|x_2\|_2 \underline{c}^{-1} = 0. \tag{A.83}$$

Let us define $\Delta \coloneqq \Delta_2 - \Delta_1$. In light of the fact that $\|z\|_2 \leq \sqrt{p_h} \|z\|_\infty$ for all $z \in \mathbb{R}^{p_h}$, (A.71), (A.83), and the fact that $\int_{x\in\mathbb{R}^{p_h}} f_h(x)dx = 1$, it holds that for all large h,

$$\begin{aligned}&\sup_{\|\Delta_i\|_\infty < e^{-s_3 h}} \Big| \int_{x \in \mathcal{D}-\Delta_2,\, \|x\|_\infty < e^{(1-\delta_0)s_3 h}} \big(f_h(x+\Delta_2-\Delta_1) - f_h(x)\big) dx \Big| \\ &\leq \sup_{\|\Delta\|_\infty < 2e^{-s_3 h}, \|\Delta_2\|_\infty < e^{-s_3 h}} \Big| \int_{x \in \mathcal{D}-\Delta_2,\, \|x\|_\infty < e^{(1-\delta_0)s_3 h}} \big(f_h(x+\Delta) - f_h(x)\big) dx \Big| \\ &\leq \sup_{\substack{\|\Delta\|_2 \leq 2\sqrt{p_h} e^{-s_3 h} \\ \|x\|_2 \leq \sqrt{p_h} e^{(1-\delta_0)s_3 h}}} (2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1} + (2\|x\|_2 \|\Delta\|_2 \underline{c}^{-1})^2) \\ &\leq \sup_{\substack{\|\Delta\|_2 \leq 2\sqrt{p_h} e^{-s_3 h} \\ \|x\|_2 \leq \sqrt{p_h} e^{(1-\delta_0)s_3 h}}} 4\|x\|_2 \|\Delta\|_2 \underline{c}^{-1} \\ &\leq 8\underline{c}^{-1} p_h \exp\left(-\delta_0 s_3 h\right).\end{aligned} \tag{A.84}$$

Therefore, combining (A.76)–(A.82), (A.84), and the stationarity of the process yields the desired conclusion. This completes the proof of Claim (A.60).

C.3 Lemma 1 and its proof

The theoretical foundation of our subsampling method is provided by Lemma 1, which concerns the asymptotic independence of the β-mixing random vectors in each subsample. Since (A.86) below for Lemma 1 involves stacking up stationary elements column-wise (there are l elements in a row of matrices in (A.86) below), it is unclear whether we can directly apply Lemma 4.1 of [42] to our setting. Thus, we provide our self-contained proof for Lemma 1. Our technical analysis of Lemma 1 seems to be the first formal proof for results on the asymptotically independent blocks due to the β-mixing and subsampling.

Consider a p-dimensional vector-valued stationary process $\{\boldsymbol{z}_t\}$. Let n, $\bar{n}$, h, and l be positive integers such that

$$\bar{n} = \sup\{s \in \mathbb{N} : s(l+h) - h \ \ \le n\} > 0. \tag{A.85}$$

We construct two $\bar{n} \times (lp)$ design matrices as

$$\begin{aligned} \boldsymbol{M} &:= \begin{pmatrix} \boldsymbol{z}^T_{(1-1)\times(l+h)+1}, \cdots, \boldsymbol{z}^T_{(2-1)\times(l+h)-h} \\ \cdots \\ \boldsymbol{z}^T_{(\bar{n}-1)\times(l+h)+1}, \cdots, \boldsymbol{z}^T_{\bar{n}\times(l+h)-h} \end{pmatrix} \\ \text{and} \quad \boldsymbol{M}^{\pi} &:= \begin{pmatrix} (\boldsymbol{z}_1^{\pi_1})^T, \cdots, (\boldsymbol{z}_l^{\pi_1})^T \\ \cdots \\ (\boldsymbol{z}_1^{\pi_{\bar{n}}})^T, \cdots, (\boldsymbol{z}_l^{\pi_{\bar{n}}})^T \end{pmatrix}, \end{aligned} \tag{A.86}$$

where $\{(\boldsymbol{z}_1^{\pi_t}, \cdots, \boldsymbol{z}_l^{\pi_t})\}_t$ is an i.i.d. sequence such that $(\boldsymbol{z}_1^{\pi_1}, \cdots, \boldsymbol{z}_l^{\pi_1})$ has the same distribution as $(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_l)$. Here, matrix $\boldsymbol{M}$ is obtained by removing h random vectors in the process after each consecutive l random vectors, and then stacking up the remaining random vectors. Lemma 1 below characterizes the distributional distance between $\boldsymbol{M}$ with dependent rows and $\boldsymbol{M}^{\pi}$ with i.i.d. rows.

Lemma 1. *Assume that the p-dimensional process $\{\boldsymbol{z}_t\}$ satisfies Condition 4 with $(h+1)$-step and constants $0 \leq \rho < 1$ and $C_0 > 0$. Then it holds that*

$$\sup_{\mathcal{D}\in\mathcal{R}^{lp\bar{n}}} \Big|\mathbb{P}(\boldsymbol{M}\in\mathcal{D}) - \mathbb{P}(\boldsymbol{M}^{\pi}\in\mathcal{D})\Big| \leq n\rho^{h+1}C_0, \tag{A.87}$$

where random matrices $\boldsymbol{M}$ and $\boldsymbol{M}^{\pi}$ are defined in (A.86).

Proof. If we view the rows of $\boldsymbol{M}$ as random mappings, a simple version of this problem is to establish an upper bound of the total variation distance between the distributions of $U_1, \cdots, U_{\bar{n}}$ and their i.i.d. counterparts denoted as $U_1^{\pi}, \cdots, U_{\bar{n}}^{\pi}$. The main technique used here is to separate the total variation distance into $\mathrm{TV}_1, \cdots, \mathrm{TV}_{\bar{n}}$ introduced below and control them separately as

$$\begin{aligned} &U_1, U_2, U_3, \cdots, U_{\bar{n}} \underbrace{\longleftrightarrow}_{\mathrm{TV}_1} U_1^{\pi}, U_2, U_3, \cdots, U_{\bar{n}} \underbrace{\longleftrightarrow}_{\mathrm{TV}_2} U_1^{\pi}, U_2^{\pi}, U_3, \cdots, U_{\bar{n}} \\ &\underbrace{\longleftrightarrow}_{\mathrm{TV}_3} \cdots \underbrace{\longleftrightarrow}_{\mathrm{TV}_{\bar{n}}} U_1^{\pi}, U_2^{\pi}, \cdots, U_{\bar{n}}^{\pi}. \end{aligned} \tag{A.88}$$

By the technique in (A.88) above, for each step, we can focus on the total variation distance between two processes with only one distinct part. For example, for the jth and $(j+1)$th processes, the distinct part is U_j and U_j^{π}. We will present the formal proof next.

To facilitate the technical presentation, let us first introduce some notations. Denote the distributions as

$$\begin{aligned}
&\mu_1 := \mu_{(\boldsymbol{z}_{(1-1)(l+h)+1\,:\,1(l+h)-h},\cdots,\boldsymbol{z}_{(i-1)(l+h)+1\,:\,i(l+h)-h},\cdots,\boldsymbol{z}_{(\bar{n}-1)(l+h)+1\,:\,\bar{n}(l+h)-h})},\\
&\mu_{\bar{n}+1} := \mu_{(\boldsymbol{z}^{\pi_1},\cdots,\boldsymbol{z}^{\pi_{\bar{n}}})}, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{A.89})\\
&\mu_j := \mu_{(\boldsymbol{z}^{\pi_1},\cdots,\boldsymbol{z}_{(j-1)(l+h)+1\,:\,j(l+h)-h},\cdots,\boldsymbol{z}_{(\bar{n}-1)(l+h)+1\,:\,\bar{n}(l+h)-h})}
\end{aligned}$$

for $j = 2, \cdots, \bar{n}$. Observe that we have $\mu_{\bar{n}} = \mu_{\bar{n}+1}$ since the process is stationary. With the notation introduced above, the desired conclusion is an upper bound for $\frac{1}{2}\left\|\mu_1 - \mu_{\bar{n}+1}\right\|_{TV}$.

For completeness, we state some important properties of the transition kernel $p : \mathbb{R}^p \times \mathcal{R}^p \longrightarrow [0,1]$ of a stationary Markov chain with stationary distribution π. 1) For each integer t and $\mathcal{D} \in \mathcal{R}^p$, $p(\boldsymbol{z}_t, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{z}_{t+1} \in \mathcal{D} | \boldsymbol{z}_t)$. 2) For each measurable function f and $\mathcal{D} \in \mathcal{R}^p$, $\int_{\mathcal{D}} p(\vec{x}, d\vec{y}) f(\vec{y})$ is a measurable function of $\vec{x}$, and hence for each $\mathcal{D}_k \in \mathcal{R}^p$,

$$\int_{\mathcal{D}_1} \pi(d\vec{x}_1) \int_{\mathcal{D}_2} p(\vec{x}_1, d\vec{x}_2) \cdots \int_{\mathcal{D}_k} p(\vec{x}_{k-1}, d\vec{x}_k) f(\vec{x}_k)$$

is a well-defined integral. 3) For each measurable function f and $\mathcal{D} \in \mathcal{R}^p$,

$$\int_{\mathbb{R}^p} \pi(d\vec{x}) \int_{\mathcal{D}} p(d\vec{x}, d\vec{y}) f(\vec{y}) = \int_{\mathcal{D}} \pi(d\vec{x}) f(\vec{x}). \tag{A.90}$$

4) We have an expression of μ_j as given in (A.91) below, where we indicate each part of the distribution of μ_j according to

$$\underbrace{\boldsymbol{z}^{\pi_1}}_{\text{1st part}}, \cdots, \underbrace{\boldsymbol{z}_{(j-1)(l+h)+1\,:\,j(l+h)-h}}_{j\text{th part}}, \cdots, \underbrace{\boldsymbol{z}_{(\bar{n}-1)(l+h)+1\,:\,\bar{n}(l+h)-h}}_{\bar{n}\text{th part}}.$$

Such representation follows from the first two properties, and the details on deriving it can be found in Section 5.2 of [16]. For each $\mathcal{D} \in \mathcal{R}^{lp\bar{n}}$, $\mu_j(\mathcal{D})$ admits the following representation

$$\begin{aligned}\mu_j(\mathcal{D}) = \int_{\mathcal{D}} & \underbrace{\pi(d\vec{x}_1) \times \cdots \times p(\vec{x}_{l-1}, d\vec{x}_l)}_{\text{1st part}} \times \cdots \\ & \times \underbrace{\pi(d\vec{x}_{(j-1)(l+h)+1}) \times \cdots \times p(\vec{x}_{j(l+h)-h-1}, d\vec{x}_{j(l+h)-h})}_{j\text{th part}} \times \cdots \\ & \times \underbrace{p^{h+1}(\vec{x}_{j(l+h)-h}, d\vec{x}_{j(l+h)+1}) \times \cdots \times p(\vec{x}_{(j+1)(l+h)-h-1}, d\vec{x}_{(j+1)(l+h)-h})}_{(j+1)\text{th part}} \times \cdots \\ & \times \underbrace{p^{h+1}(\vec{x}_{(\bar{n}-1)(l+h)-h}, d\vec{x}_{(\bar{n}-1)(l+h)+1}) \times \cdots \times p(\vec{x}_{\bar{n}(l+h)-h-1}, d\vec{x}_{\bar{n}(l+h)-h})}_{\bar{n}\text{th part}},\end{aligned} \tag{A.91}$$

where $\underbrace{\vec{x}_1, \cdots, \vec{x}_l}_{\text{1st part}}, \cdots, \underbrace{\vec{x}_{(\bar{n}-1)(l+h)+1}, \cdots, \vec{x}_{\bar{n}(l+h)-h}}_{\bar{n}\text{th part}}$ stand for the corresponding running variables with $\vec{x}_k \in \mathbb{R}^p$ for each k.

Let us make use of a critical observation that

$$\frac{1}{2}\|\mu_1 - \mu_{\bar{n}+1}\|_{TV} \leq \sum_{i=1}^{\bar{n}} \sup_{\mathcal{D}\in\mathcal{R}^{lp\bar{n}}} |\mu_j(\mathcal{D}) - \mu_{j+1}(\mathcal{D})|. \tag{A.92}$$

We will bound each term in the above summation separately. Let us fix $1 \leq j \leq \bar{n}$. We notice that μ_j and μ_{j+1} are almost identical except for the $(j+1)$th part in (A.91). By a careful comparison, we see that for μ_j, the $(j+1)$th part starts with $p^{h+1}(\vec{x}_{j(l+h)-h}, d\vec{x}_{j(l+h)+1})$, whereas the $(j+1)$th part of μ_{j+1} starts with $\pi(d\vec{x}_{j(l+h)+1})$. To see the difficulty for bounding each term on the right-hand side (RHS) of (A.92) using such observation, note that $\int_{\mathcal{D}} |\mu_1(dx) - \mu_2(dx)|$ is not a well-defined integral for a Borel set $\mathcal{D}$ and two measures μ_1 and μ_2 since there are two dx's inside the integration. To have a valid argument for this bound, we use the Radon–Nikodym theorem [16] to replace the underlying measures with measurable functions (the Radon–Nikodym derivatives). The arguments follow mainly those for the proof of Lemma 5 in Section C.7.

By Condition 4, for each $\vec{x} \in \mathbb{R}^p$ it holds that $p(\vec{x}, \cdot)$ is dominated by the Lebesgue measure. Since p is the transition kernel of the stationary Markov chain, this entails that 1) $p^{h+1}(\vec{x},\cdot)$ is dominated by the Lebesgue measure for each $\vec{x}\in\mathbb{R}^p$ and 2) $\pi(\cdot)$ is also dominated by the Lebesgue measure. By 1) and the Radon–Nikodym Theorem, there exists a nonnegative measurable function on $\mathbb{R}^{2p}$, which is denoted as r, such that for each $\vec{x}\in\mathbb{R}^p$ and $\mathcal{D}\in\mathcal{R}^p$,

$$p^{h+1}(\vec{x},\mathcal{D}) = \int_{\mathcal{D}} r(\vec{x},\vec{y})\ d\vec{y}.$$

This measurable function is simply the Radon–Nikodym derivative [16], and $r(\vec{x},\vec{y})$ is also called the probability density functions of $\boldsymbol{z}_{t+h+1}$ conditional on $\boldsymbol{z}_t$. In particular, for each $\mathcal{D}_1,\mathcal{D}_2\in\mathcal{R}^p$, we have that

$$\mathbb{P}((\boldsymbol{z}_t,\boldsymbol{z}_{t+h+1})\in\mathcal{D}_1\times\mathcal{D}_2) = \int_{\vec{x}\in\mathcal{D}_1}\pi(d\vec{x})\int_{\vec{y}\in\mathcal{D}_2} r(\vec{x},\vec{y})d\vec{y}.$$

For more details on the conditional probability density functions, see Example 4.1.6 of [16]. Furthermore, by 2) we denote by $r_\pi(\vec{x})$ the Radon–Nikodym derivative such that for each $\mathcal{D}\in\mathcal{R}^p$,

$$\pi(\mathcal{D}) = \int_{\mathcal{D}} r_\pi(\vec{x})d\vec{x}.$$

Thus, we can obtain that

$$\begin{aligned}\mu_j(\mathcal{D}) = \int_{\mathcal{D}} \cdots & \\ \times \underbrace{r(\vec{x}_{j(l+h)-h},\vec{x}_{j(l+h)+1})d\vec{x}_{j(l+h)+1}\times\cdots\times p(\vec{x}_{(j+1)(l+h)-h-1},d\vec{x}_{(j+1)(l+h)-h})}_{(j+1)\text{th part}} & \qquad (A.93)\\ \times\cdots\times \underbrace{p^{h+1}(\vec{x}_{(\bar{n}-1)(l+h)-h},d\vec{x}_{(\bar{n}-1)(l+h)+1})\times\cdots\times p(\vec{x}_{\bar{n}(l+h)-h-1},d\vec{x}_{\bar{n}(l+h)-h})}_{\bar{n}\text{th part}}. &\end{aligned}$$

A similar expression also holds for μ_{j+1} with r_π. We will bound $|\mu_j(\mathcal{D})-\mu_{j+1}(\mathcal{D})|$ next.

It follows from (A.93) and the fact of $p \leq 1$ that for each $\mathcal{D} \in \mathcal{R}^{lp\bar{n}}$,

$$\begin{aligned}
&|\mu_j(\mathcal{D}) - \mu_{j+1}(\mathcal{D})| \\
&= \left| \int_{\mathcal{D}} \underbrace{\cdots}_{1,\cdots,j} \times \left(r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right) d\vec{x}_{j(l+h)+1} \times \cdots \right| \\
&\leq \int_{\mathcal{D}} \underbrace{\cdots}_{1,\cdots,j} \times \left| r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right| d\vec{x}_{j(l+h)+1} \times \cdots \\
&\leq \int_{\mathbb{R}^{lpj+p}} \underbrace{\cdots}_{\text{1st},\cdots,j\text{th parts}} \times \left| r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right| d\vec{x}_{j(l+h)+1}.
\end{aligned} \tag{A.94}$$

To bound the RHS of (A.94), we separate the modulus into positive and negative parts and get rid of the modulus operation. Let $\mathcal{D}_+$ and $\mathcal{D}_-$ be two disjoint Borel sets such that

$$\begin{aligned}
&\int_{\mathbb{R}^{lpj+p}} \underbrace{\cdots}_{1,\cdots,j} \times \left| r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right| d\vec{x}_{j(l+h)+1} \\
&= \int_{\mathcal{D}_+} \underbrace{\cdots}_{1,\cdots,j} \times \left(r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right) d\vec{x}_{j(l+h)+1} \\
&\quad + \int_{\mathcal{D}_-} \underbrace{\cdots}_{\text{1st},\cdots,j\text{th parts}} \times \left(r_\pi(\vec{x}_{j(l+h)+1}) - r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) \right) d\vec{x}_{j(l+h)+1}.
\end{aligned} \tag{A.95}$$

To proceed, we exploit arguments involving "cross-sections." For any $\mathcal{D} \in \mathbb{R}^{k_1+k_2}$ and $\vec{x} \in \mathbb{R}^{k_1}$ with k_1 and k_2 some positive integers, let us define the cross-section at $\vec{x}$ as $\mathcal{D}_{\vec{x}} := \{\vec{y} : (\vec{x}, \vec{y}) \in \mathcal{D}\}$. See Section C.4 for more detail on cross-sections. Then it holds that

$$\begin{aligned}
&\int_{\mathcal{D}_+} \underbrace{\cdots}_{1,\cdots,j} \times \left(r(\vec{x}_{j(l+h)-h}, \vec{x}_{j(l+h)+1}) - r_\pi(\vec{x}_{j(l+h)+1}) \right) d\vec{x}_{j(l+h)+1} \\
&= \int_{\mathcal{D}_+} \underbrace{\cdots}_{1,\cdots,j} \times p^{h+1}(\vec{x}_{j(l+h)-h}, d\vec{x}_{j(l+h)+1}) - \int_{\mathcal{D}_+} \underbrace{\cdots}_{1,\cdots,j} \times \pi(d\vec{x}_{j(l+h)+1}) \\
&= \int_{\mathbb{R}^{lpj}} \underbrace{\cdots}_{1,\cdots,j} \times p^{h+1}(\vec{x}_{j(l+h)-h}, (\mathcal{D}_+)_z) - \int_{\mathbb{R}^{lpj}} \underbrace{\cdots}_{1,\cdots,j} \times \pi((\mathcal{D}_+)_z) \\
&= \int_{\mathbb{R}^{lpj}} \underbrace{\cdots}_{\text{1st},\cdots,j\text{th parts}} \times \left(p^{h+1}(\vec{x}_{j(l+h)-h}, (\mathcal{D}_+)_z) - \pi((\mathcal{D}_+)_z) \right),
\end{aligned} \tag{A.96}$$

where z represents $(\vec{x}_1^T, \cdots, \vec{x}_{j(l+h)-h}^T)^T$ in the integration. Here, we have used the definition of the Radon–Nikodym derivative to get the first equality in (A.96). The second equality in (A.96) is justified by the fact that π is a distribution and hence a transition kernel, and an application of Lemma 3 in Section C.5. To apply Lemma 3 in (A.96), we can regard $\boldsymbol{z}_{j(l+h)+1}$ as $\boldsymbol{X}_3$, $\boldsymbol{z}_{j(l+h)-h}$ as $\boldsymbol{X}_2$, and the remaining variables as $\boldsymbol{X}_1$ in Lemma 3, and notice that (A.99) is satisfied due to the Markov property. Similar arguments can be applied to $\mathcal{D}_-$ too.

In view of (A.95) and (A.96), it follows from the fact of $\mathcal{D}_+ \cap \mathcal{D}_- = \emptyset$, (A.90), and Condition 4 that

$$\begin{aligned}
\text{RHS of } (A.94) &= \int_{\mathbb{R}^{lpj}} \underbrace{\cdots}_{1,\cdots,j} \times \Big[\big(p^{h+1}(\vec{x}_{j(l+h)-h}, (\mathcal{D}_+)_z) - \pi((\mathcal{D}_+)_z)\big) \\
&\quad - \big(p^{h+1}(\vec{x}_{j(l+h)-h}, (\mathcal{D}_-)_z) - \pi((\mathcal{D}_-)_z)\big) \Big] \\
&\leq \int_{\mathbb{R}^{lpj}} \underbrace{\cdots}_{1\text{st},\cdots,j\text{th parts}} \times \left\| p^{h+1}(\vec{x}_{j(l+h)-h}, \cdot) - \pi(\cdot) \right\|_{TV} \qquad \text{(A.97)} \\
&= \int_{\mathbb{R}^p} \pi(d\vec{x}_{j(l+h)-h}) \left\| p^{h+1}(\vec{x}_{j(l+h)-h}, \cdot) - \pi(\cdot) \right\|_{TV} \\
&\leq \int_{\mathbb{R}^p} V(\vec{x})\pi(d\vec{x})\rho^{h+1} C,
\end{aligned}$$

where $C > 0$ is given in Condition 4. By Condition 4, we can further show that

$$\text{RHS of } (A.97) \leq C_0 \rho^{h+1},$$

where $C_0 > 0$ is a constant such that $\int_{\mathbb{R}^p} V(\vec{x})\pi(d\vec{x})C \leq C_0$. Therefore, combining (A.92), (A.97), and the fact of $\bar{n} \leq n$, we can see that the upper bound is given by $n\rho^{h+1}C_0$, which concludes the proof of Lemma 1.

C.4 Lemma 2 and its proof

To facilitate the technical presentation, let us first introduce some necessary notation. For any $\mathcal{D} \subset \mathbb{R}^{k_1+k_2}$ and $x \in \mathbb{R}^{k_1}$ with k_1 and k_2 some positive integers, we define the cross-section at x as $\mathcal{D}_x \coloneqq \{y : (x, y) \in \mathcal{D}\}$. A standard operation on $\mathcal{D}_x$ is described as follows. If $\mathcal{D}_1 \subset \mathcal{D}_2$, then we have that for each x, $(\mathcal{D}_1)_x \subset (\mathcal{D}_2)_x$ and $(\mathcal{D}_2 \backslash \mathcal{D}_1)_x = (\mathcal{D}_2)_x \backslash (\mathcal{D}_1)_x$. In addition, for any set $\mathcal{D} \subset \mathbb{R}^k$ and $x \in \mathbb{R}^k$, we denote by $\mathcal{D} - x$ the set $\{y - x : y \in \mathcal{D}\}$. The expectation $\int f(x)\mu(dx)$ for a measurable function f with respect to some random vector $\boldsymbol{X}$ with distribution μ is written as $\int_{\mathbb{R}^k} f d\mu$ whenever the running variables are obvious. We also use the product notation in the integration to specify the running variables such as $\int f(x_2)\mu(dx_1 \times dx_2)$.

Lemma 2. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be k_1-dimensional and k_2-dimensional random vectors, respectively. Assume that $h : \mathbb{R}^{k_1} \times \mathcal{R}^{k_2} \longrightarrow [0, 1]$ is a transition kernel such that for each $\mathcal{D} \in \mathcal{R}^{k_2}$, $h(\boldsymbol{X}, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{Y} \in \mathcal{D} \mid \boldsymbol{X})$. Then it holds that*

1) For each $\mathcal{D} \in \mathcal{R}^{k_1+k_2}$ and $x \in \mathbb{R}^{k_1}$, $\mathcal{D}_x \in \mathcal{R}^{k_2}$.

2) For each $\mathcal{D} \in \mathcal{R}^{k_1+k_2}$, $h(\cdot, \mathcal{D}.)$ is $\mathcal{R}^{k_1}$-measurable.

3) For each $\mathcal{D} \in \mathcal{R}^{k_1+k_2}$, $\mathbb{P}((\boldsymbol{X}^T, \boldsymbol{Y}^T) \in \mathcal{D}) = \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) h(x, \mathcal{D}_x)$.

Proof. We begin with showing part 1). Let L be the collection of sets in $\mathcal{R}^{k_1+k_2}$ satisfying the required conditions; that is, for each $\mathcal{D} \in L$, it holds that for each $x \in \mathbb{R}^{k_1}$, $\mathcal{D}_x \in \mathcal{R}^{k_2}$. Then it is easy to verify that L contains all the rectangles of form $\mathcal{A} \times \mathcal{B}$, where $\mathcal{A} \in \mathcal{R}^{k_1}$ and $\mathcal{B} \in \mathcal{R}^{k_2}$. By the basic set operations, it holds that for each $x \in \mathbb{R}^{k_1}$ and $E, E_i \in \mathcal{R}^{k_1+k_2}$,

a) $(E_x)^c = (\{y : (x, y) \in E\})^c = \{y : (x, y) \in E^c\} = (E^c)_x$;

b) $\cup_i (E_i)_x = \cap_i ((E_i)_x)^c = \cap_i (E_i^c)_x = (\cap_i E_i^c)_x = (\cup_i E_i)_x$.

Thus, for $E, E_i \in L$, we have that $E^c, \cup_i E_i \in L$. This shows that L is a σ-algebra. Since L contains all the rectangles, we obtain the conclusion in part 1) of Lemma 2.

We next proceed to establish part 2). Since $\mathcal{D}_x$ is a measurable set, $h(x, \mathcal{D}_x)$ is a well-defined function of x. Let L be the collection of sets such that for each $\mathcal{D} \in L$, $h(\cdot, \mathcal{D}.)$ is $\mathcal{R}^{k_1}$-measurable. Since for each $\mathcal{A} \in \mathcal{R}^{k_1}$ and $\mathcal{B} \in \mathcal{R}^{k_2}$, it holds that

$$h(x, (\mathcal{A} \times \mathcal{B})_x) = h(x, \mathcal{B}) \mathbf{1}_{\{x \in \mathcal{A}\}},$$

which is a measurable function of x, we can see that L contains all such rectangles. Moreover, if $\mathcal{D}_1, \mathcal{D}_2 \in L$ with $\mathcal{D}_1 \subset \mathcal{D}_2$, then it follows that for each x, $(\mathcal{D}_1)_x \subset (\mathcal{D}_2)_x$ and $(\mathcal{D}_2 \backslash \mathcal{D}_1)_x = (\mathcal{D}_2)_x \backslash (\mathcal{D}_1)_x$, and hence

$$h(x, (\mathcal{D}_2 \backslash \mathcal{D}_1)_x) = h(x, (\mathcal{D}_2)_x \backslash (\mathcal{D}_1)_x) = h(x, (\mathcal{D}_2)_x) - h(x, (\mathcal{D}_1)_x).$$

Observe that the RHS of the equality above is measurable since the subtraction of measurable functions is still measurable. Next, if $\mathcal{D}_i \in L$ and $\mathcal{D}_i \subset \mathcal{D}_{i+1}$, by the continuity of measure, we have that for each $x \in \mathbb{R}^{k_1}$, $\lim_{n \to \infty} h(x, (\cup_{i=1}^n \mathcal{D}_i)_x) = h(x, (\cup_{i=1}^\infty \mathcal{D}_i)_x)$. Thus, $h(x, (\cup_{i=1}^\infty \mathcal{D}_i)_x)$ is a measurable function of x, and we have $\cup_{i=1}^\infty \mathcal{D}_i \in L$. This shows that L is a λ-system containing the set of all the rectangles. Hence, by Lemma 8 in Section C.12, we see that L contains the σ-algebra generated by the set, which concludes the proof for part 2) of Lemma 2.

Finally, let us show part 3). Note that the RHS of the assertion is well-defined due to part 2) of Lemma 2. By the definition of the conditional expectation, the change of variables formula, and the fact that for each $x \in \mathbb{R}^{k_1}$, $\mathcal{A} \in \mathcal{R}^{k_1}$, $\mathcal{B} \in \mathcal{R}^{k_2}$,

$$\mathbf{1}_{\{x \in \mathcal{A}\}} h(x, \mathcal{B}) = h(x, (\mathcal{A} \times \mathcal{B})_x),$$

we can deduce that

$$\begin{aligned}\mathbb{P}((\boldsymbol{X}^{T}, \boldsymbol{Y}^{T}) \in \mathcal{A} \times \mathcal{B}) &= \int_{\Omega} \mathbf{1}_{\{\boldsymbol{X} \in \mathcal{A}\}} \ \mathbb{P}(\boldsymbol{Y} \in \mathcal{B} \mid \boldsymbol{X}) d\mathbb{P} \\ &= \int_{\Omega} \mathbf{1}_{\{\boldsymbol{X} \in \mathcal{A}\}} \ h(\boldsymbol{X}, \mathcal{B}) d\mathbb{P} \\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ \mathbf{1}_{\{x \in \mathcal{A}\}} \ h(x, \mathcal{B}) \\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ h(x, (\mathcal{A} \times \mathcal{B})_x),\end{aligned} \tag{A.98}$$

where Ω represents the underlying probability space.

Denote by L the collection of sets in $\mathcal{R}^{k_1+k_2}$ satisfying the required condition; that is, for each $\mathcal{D} \in L$, it holds that $\mathbb{P}((\boldsymbol{X}^{T}, \boldsymbol{Y}^{T}) \in \mathcal{D}) = \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ h(x, (\mathcal{D})_x)$. In view of (A.98), L contains all the rectangles in $\mathbb{R}^{k_1+k_2}$. In addition, we will make use of the following two facts.

a) If $\mathcal{D}_1, \mathcal{D}_2 \in L$ and $\mathcal{D}_1 \subset \mathcal{D}_2$, then an application of similar arguments to those in the proof for part 2) of Lemma 2 leads to

$$\begin{aligned}\mathbb{P}((\boldsymbol{X}^{T}, \boldsymbol{Y}^{T}) \in \mathcal{D}_2 \backslash \mathcal{D}_1) &= \mathbb{P}((\boldsymbol{X}^{T}, \boldsymbol{Y}^{T}) \in \mathcal{D}_2) - \mathbb{P}((\boldsymbol{X}^{T}, \boldsymbol{Y}^{T}) \in \mathcal{D}_1) \\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ \Big(h(x, (\mathcal{D}_2)_x) - h(x, (\mathcal{D}_1)_x)\Big) \\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ h(x, (\mathcal{D}_2)_x \backslash (\mathcal{D}_1)_x) \\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx) \ h(x, (\mathcal{D}_2 \backslash \mathcal{D}_1)_x),\end{aligned}$$

which shows that $\mathcal{D}_2 \backslash \mathcal{D}_1 \in L$.

b) Assume that $\mathcal{D}_i \in L$ for each n and $\mathcal{D}_i \subset \mathcal{D}_{i+1}$. Then it follows from the continuity of measure, the definition of L, and the monotone convergence theorem that

$$\begin{aligned}\mathbb{P}((\boldsymbol{X}^T, \boldsymbol{Y}^T) \in \cup_{i=1}^{\infty}\mathcal{D}_i) &= \lim_{n\to\infty} \mathbb{P}((\boldsymbol{X}^T, \boldsymbol{Y}^T) \in \mathcal{D}_n)\\ &= \lim_{n\to\infty} \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx)\ h(x, (\mathcal{D}_n)_x)\\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx)\ \lim_{n\to\infty} h(x, (\mathcal{D}_n)_x)\\ &= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{X}}(dx)\ h(x, (\cup_{i=1}^{\infty}\mathcal{D}_i)_x),\end{aligned}$$

which shows that $\cup_{i=1}^{\infty} D_i \in L$.

Therefore, using the aforementioned facts, an application of Lemma 8 leads to the conclusion in part 3) of Lemma 2. This completes the proof of Lemma 2.

C.5 Lemma 3 and its proof

Lemma 3. *Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$ be k_1-dimensional, k_2-dimensional, and k_3-dimensional random vectors, respectively, such that for each $\mathcal{D} \in \mathcal{R}^{k_3}$,*

$$\mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D} \mid \boldsymbol{X}_2) = \mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D} \mid \boldsymbol{X}_2, \boldsymbol{X}_1). \tag{A.99}$$

We define a transition kernel $h : \mathbb{R}^{k_2} \times \mathcal{R}^{k_3} \longrightarrow [0, 1]$ such that for each $\mathcal{D} \in \mathcal{R}^{k_3}$, $h(\boldsymbol{X}_2, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D} \mid \boldsymbol{X}_2)$. Then it holds that

1) *For each $\mathcal{D} \in \mathcal{R}^{k_3}$, $h(\boldsymbol{X}_2, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D} \mid \boldsymbol{X}_2, \boldsymbol{X}_1)$.*

2) *For each $\mathcal{D} \in \mathcal{R}^{k_1+k_2+k_3}$, $h(\boldsymbol{X}_2, \mathcal{D}_{(\boldsymbol{X}_1,\boldsymbol{X}_2)})$* is a version of *$\mathbb{P}((\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) \in \mathcal{D} \mid \boldsymbol{X}_2, \boldsymbol{X}_1)$.*

Proof. We first show part 1). Since $h(\boldsymbol{X}_2, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D} \mid \boldsymbol{X}_2)$, $h(\boldsymbol{X}_2, \mathcal{D})$ is $\sigma(\boldsymbol{X}_2)$-measurable, and hence $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2)$-measurable. In conjunction with (A.99) and the definition of conditional expectation, we see that $h(\boldsymbol{X}_2, \mathcal{D})$ a version of $\mathbb{P}(\boldsymbol{X}_3 \in$

$\mathcal{D} \mid \boldsymbol{X}_2, \boldsymbol{X}_1)$. This yields the conclusion in part 1) of Lemma 3. We then establish part 2). Let us first verify that $h(\boldsymbol{X}_2, \mathcal{D}_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ is $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2)$-measurable for each $\mathcal{D} \in \mathcal{R}^{k_1+k_2+k_3}$. We start with an observation that for each $\mathcal{D}_1 \in \mathcal{R}^{k_1}$, $\mathcal{D}_2 \in \mathcal{R}^{k_2}$, $\mathcal{D}_3 \in \mathcal{R}^{k_3}$, it holds that

$$h(\boldsymbol{X}_2, (\mathcal{D}_1 \times \mathcal{D}_2 \times \mathcal{D}_3)_{\boldsymbol{X}_1,\boldsymbol{X}_2}) = h(\boldsymbol{X}_2, \mathcal{D}_3)\mathbf{1}_{\{\boldsymbol{X}_1 \in \mathcal{D}_1\}}\mathbf{1}_{\{\boldsymbol{X}_2 \in \mathcal{D}_2\}},$$

which is $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2)$-measurable. This shows that for each Borel rectangle $\mathcal{D}$, $h(\boldsymbol{X}_2, \mathcal{D}_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ is $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2)$-measurable. In conjunction with similar arguments to those in the proof of Lemma 2 in Section C.4, the desired result follows.

Let $\mathcal{D} \in \mathcal{R}^{k_1+k_2+k_3}$ be given. Then we will show that for each $\mathcal{B} \in \mathcal{R}^{k_1+k_2}$,

$$\begin{aligned} &\int h(\boldsymbol{X}_2, \mathcal{D}_{\boldsymbol{X}_1,\boldsymbol{X}_2})\ \mathbf{1}_{\{(\boldsymbol{X}_1,\boldsymbol{X}_2)\in\mathcal{B}\}}\ d\mathbb{P} \\ &= \int \mathbb{P}((\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) \in \mathcal{D} \mid \boldsymbol{X}_2, \boldsymbol{X}_1)\ \mathbf{1}_{\{(\boldsymbol{X}_1,\boldsymbol{X}_2)\in\mathcal{B}\}}\ d\mathbb{P}. \end{aligned} \tag{A.100}$$

The left-hand side of (A.100) is well-defined since $h(\boldsymbol{X}_2, \mathcal{D}_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ is measurable. To show (A.100), we again apply similar arguments to those in the proof of Lemma 2. Specifically, let L be the collection of sets in $\mathcal{R}^{k_1+k_2+k_3}$ such that (A.100) holds. Since for each $\mathcal{D}_1 \in \mathcal{R}^{k_1}$, $\mathcal{D}_2 \in \mathcal{R}^{k_2}$, $\mathcal{D}_3 \in \mathcal{R}^{k_3}$, we have

$$\begin{aligned} &\mathbb{P}((\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) \in (\mathcal{D}_1 \times \mathcal{D}_2 \times \mathcal{D}_3) \mid \boldsymbol{X}_2, \boldsymbol{X}_1) \\ &= \mathbb{P}(\boldsymbol{X}_3 \in \mathcal{D}_3 \mid \boldsymbol{X}_2, \boldsymbol{X}_1)\mathbf{1}_{\{\boldsymbol{X}_1 \in \mathcal{D}_1\}}\mathbf{1}_{\{\boldsymbol{X}_2 \in \mathcal{D}_2\}} \\ &= h(\boldsymbol{X}_2, \mathcal{D}_3)\mathbf{1}_{\{\boldsymbol{X}_1 \in \mathcal{D}_1\}}\mathbf{1}_{\{\boldsymbol{X}_2 \in \mathcal{D}_2\}} \\ &= h(\boldsymbol{X}_2, (\mathcal{D}_1 \times \mathcal{D}_2 \times \mathcal{D}_3)_{\boldsymbol{X}_1,\boldsymbol{X}_2}), \end{aligned}$$

it holds that L contains all Borel rectangles $\mathcal{D} \in \mathcal{R}^{k_1+k_2+k_3}$. The remaining arguments follow those in the proof of Lemma 2. Finally, since $h(\boldsymbol{X}_2, \mathcal{D}_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ is $\sigma(\boldsymbol{X}_1, \boldsymbol{X}_2)$-measurable, by (A.100) and the definition of conditional expectation, we can obtain the conclusion in part 2) of Lemma 3. This concludes the proof of Lemma 3.

C.6 Lemma 4 and its proof

Lemma 4. *Let $\{\boldsymbol{U}_i\}$ and $\{\boldsymbol{V}_i\}$ be sequences of k_1-dimensional and k_2-dimensional random vectors, respectively. Assume that $(\boldsymbol{U}_i, \boldsymbol{V}_i)$'s are identically distributed. Then there exists a transition kernel $g : \mathbb{R}^{k_1} \times \mathcal{R}^{k_2} \longrightarrow [0,1]$ such that for each i and $\mathcal{D} \in \mathcal{R}^{k_2}$, $g(\boldsymbol{U}_i, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{V}_i \in \mathcal{D} \mid \boldsymbol{U}_i)$.*

Proof. For each $(\boldsymbol{U}_i, \boldsymbol{V}_i)$, there exists a transition kernel $g_i : \mathbb{R}^{k_1} \times \mathcal{R}^{k_2} \longrightarrow [0,1]$ such that for each $\mathcal{D} \in \mathcal{R}^{k_2}$, $g_i(\boldsymbol{U}_i, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{V}_i \in \mathcal{D} \mid \boldsymbol{U}_i)$; see, for example, Theorem 4.1.18 of [16]. By this and the fact that $\mu_{\boldsymbol{U}_i} = \mu_{\boldsymbol{U}_1}$ for each i, we have that for each $\mathcal{A} \in \mathcal{R}^{k_1}$ and $\mathcal{B} \in \mathcal{R}^{k_2}$,

$$\begin{aligned}
\mathbb{P}((\boldsymbol{U}_i, \boldsymbol{V}_i) \in \mathcal{A} \times \mathcal{B}) &= \int_\Omega \mathbf{1}_{\{\boldsymbol{U}_i \in \mathcal{A}\}} \mathbb{P}(\boldsymbol{V}_i \in \mathcal{B} \mid \boldsymbol{U}_i) d\mathbb{P} \\
&= \int_\Omega \mathbf{1}_{\{\boldsymbol{U}_i \in \mathcal{A}\}}\ g_i(\boldsymbol{U}_i, \mathcal{B}) d\mathbb{P} \\
&= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{U}_i}(dx)\ \mathbf{1}_{\{x \in \mathcal{A}\}}\ g_i(x, \mathcal{B}) \\
&= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{U}_1}(dx)\ \mathbf{1}_{\{x \in \mathcal{A}\}}\ g_1(x, \mathcal{B}) \\
&= \int_{\mathbb{R}^{k_1}} \mu_{\boldsymbol{U}_i}(dx)\ \mathbf{1}_{\{x \in \mathcal{A}\}}\ g_1(x, \mathcal{B}) \\
&= \int_\Omega \mathbf{1}_{\{\boldsymbol{U}_i \in \mathcal{A}\}}\ g_1(\boldsymbol{U}_i, \mathcal{B}) d\mathbb{P},
\end{aligned} \tag{A.101}$$

where Ω represents the underlying probability space, the third and last equalities above follow from the change of variables formula, and the fourth and fifth equalities above are due to the assumption of identical distributions. Therefore, it follows from (A.101) and the definition of conditional expectation that $g_1(\boldsymbol{U}_i, \mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{V}_i \in \mathcal{D} \mid \boldsymbol{U}_i)$ for each i and $\mathcal{D} \in \mathcal{R}^{k_2}$. This completes the proof of Lemma 4.

C.7 Lemma 5 and its proof

Intuitively, Lemma 5 below extracts the total variation distance from a difference of two integrals. The results are natural, but the arguments are somewhat delicate. We note that if the density functions f_X and f_Y exist, then it holds that

$$\begin{aligned}\sup_{\mathcal{D}\in\mathcal{R}} |\int_{\mathcal{D}} f_X(z)dz - \int_{\mathcal{D}} f_Y(z)dz| &\le \sup_{\mathcal{D}\in\mathcal{R}} \int_{\mathcal{D}} |f_X(z) - f_Y(z)|dz \\ &= \frac{1}{2} \|\mu_X - \mu_Y\|_{TV} .\end{aligned}$$

However, the same calculation does not apply directly to distributions because the integral $\int |\mu_X(dz) - \mu_Y(dz)|$ is not well-defined due to the two dzs inside the integration. Lemma 5 provides valid arguments in such situations.

Lemma 5. *1) Let μ and ν be two probability measures and $f : \mathbb{R}^K \longrightarrow \mathbb{R}$ a measurable function with $0 \le f \le 1$. Then it holds that*

$$\sup_{\mathcal{D}\in\mathcal{R}^K} \left| \int_{\mathcal{D}} f(x)\mu(dx) - \int_{\mathcal{D}} f(x)\nu(dx) \right| \le \frac{1}{2} \|\mu - \nu\|_{TV} .$$

2) Let $p(\cdot,\cdot) : \mathbb{R}^{K_1} \times \mathcal{R}^{K_2} \longmapsto [0,1]$ be a transition kernel and μ a probability measure on $\mathcal{R}^{K_2}$ such that for each $x \in \mathbb{R}^{K_1}$, $p(x,\cdot)$ is dominated by μ. Further let ν be a probability measure on $\mathcal{R}^{K_1}$ and $0 \le f \le 1$ a measurable function on $\mathbb{R}^{K_2}$. Then it holds that

$$\begin{aligned}&\sup_{\mathcal{D}\in\mathcal{R}^{K_1+K_2}} \left| \int_{\mathcal{D}} \nu(dx_1)p(x_1,dx_2)f(x_2) - \int_{\mathcal{D}} \nu(dx_1)\mu(dx_2)f(x_2) \right| \\ &\le \frac{1}{2} \int_{\mathbb{R}^{K_1}} \nu(dx_1) \|p(x_1,\cdot) - \mu(\cdot)\|_{TV} .\end{aligned} \tag{A.102}$$

Proof. We start with proving the first assertion. By the definition of the total variation distance, we can show that there exists some set $\mathcal{A} \in \mathcal{R}^K$ such that

$$\int_{\mathcal{A}} \mu(dx) - \int_{\mathcal{A}} \nu(dx) = \frac{1}{2} \|\mu - \nu\|_{TV} .$$

Further it holds that for each $\mathcal{B} \in \mathcal{R}^K$ with $\mathcal{B} \subset \mathcal{A}$,

$$\int_{\mathcal{B}} \mu(dx) - \int_{\mathcal{B}} \nu(dx) \geq 0.$$

Let us define $\mathcal{D}_j \coloneqq \mathcal{A} \cap \{x : \frac{j-1}{M} \leq f(x) < \frac{j}{M}\}$ for $j = 1, \cdots, M+1$ with M some positive integer. Denote by $\bar{f}$ a step function such that on $\mathcal{D}_j$, it holds that $\bar{f} = \frac{j}{M}$. Then we can deduce that

$$\begin{aligned}
&\left| \int_{\cup_j \mathcal{D}_j} f(x)\mu(dx) - \int_{\cup_j \mathcal{D}_j} f(x)\nu(dx) - \left(\int_{\cup_j \mathcal{D}_j} \bar{f}(x)\mu(dx) - \int_{\cup_j \mathcal{D}_j} \bar{f}(x)\nu(dx) \right) \right| \\
&\leq \int_{\cup_j \mathcal{D}_j} |f(x) - \bar{f}(x)|\mu(dx) + \int_{\cup_j \mathcal{D}_j} |f(x) - \bar{f}(x)|\nu(dx) \\
&\leq \frac{1}{M} \left(\int_{\cup_j \mathcal{D}_j} \mu(dx) + \int_{\cup_j \mathcal{D}_j} \nu(dx) \right) \\
&\leq \frac{2}{M}.
\end{aligned} \tag{A.103}$$

In light of the construction of $\mathcal{D}_j$'s and $\bar{f}$ above, we have that

$$\begin{aligned}
\int_{\cup_j \mathcal{D}_j} \bar{f}(x)\mu(dx) - \int_{\cup_j \mathcal{D}_j} \bar{f}(x)\nu(dx) &= \sum_j \int_{\mathcal{D}_j} \bar{f}(x)\mu(dx) - \int_{\mathcal{D}_j} \bar{f}(x)\nu(dx) \\
&= \sum_j \frac{j}{M} \left(\int_{\mathcal{D}_j} \mu(dx) - \int_{\mathcal{D}_j} \nu(dx) \right) \\
&\leq \sum_j \left(\int_{\mathcal{D}_j} \mu(dx) - \int_{\mathcal{D}_j} \nu(dx) \right) \\
&= \frac{1}{2} \|\mu - \nu\|_{TV}.
\end{aligned} \tag{A.104}$$

Then it follows from $\mathcal{A} = \cup_j \mathcal{D}_j$, (A.103), (A.104), and the fact that the positive integer M can be arbitrarily large that

$$\left| \int_{\mathcal{A}} f(x)\mu(dx) - \int_{\mathcal{A}} f(x)\nu(dx) \right| \leq \frac{1}{2} \|\mu - \nu\|_{TV}. \tag{A.105}$$

Using similar arguments as above, we can show that

$$\sup_{\mathcal{D}\in\mathcal{R}^K}\left|\int_{\mathcal{D}} f(x)\mu(dx) - \int_{\mathcal{D}} f(x)\nu(dx)\right| = \left|\int_{\mathcal{A}} f(x)\mu(dx) - \int_{\mathcal{A}} f(x)\nu(dx)\right|. \tag{A.106}$$

Therefore, combining (A.105) and (A.106) results in the desired conclusion in part 1) of Lemma 5.

We now proceed with showing the second assertion. Since $\|p(x_1,\cdot)-\mu(\cdot)\|_{TV}$ is a measurable function in x_1, the RHS of (A.102) is well defined. Such a claim can be established using similar arguments to those in Theorem 5.2.2 of Durrett [16]; for simplicity, we omit the details. By the assumptions, let $f_1(x_1,x_2)$ be the Radon–Nikodym derivative such that for each $x_1\in\mathbb{R}^{K_1}$ and $\mathcal{D}\in\mathcal{R}^{K_2}$,

$$\int_{\mathcal{D}} p(x_1,dx_2) = \int_{\mathcal{D}} f_1(x_1,x_2)\mu(dx_2).$$

If μ is the Lebesgue measure and for each $\mathcal{D}\in\mathcal{R}^{K_2}$, $p(\boldsymbol{X},\mathcal{D})$ is a version of $\mathbb{P}(\boldsymbol{Y}\in\mathcal{D}|\boldsymbol{X})$ for some random mappings $\boldsymbol{Y}$ and $\boldsymbol{X}$ dominated by the Lebesgue measure, such a Radon–Nikodym derivative is usually referred to as the conditional (on the density function of $\boldsymbol{X}$) probability density function; for this, see also Example 4.1.6 in [16]. Thus, for each $\mathcal{D}\in\mathcal{R}^{K_1+K_2}$, we have that

$$\begin{aligned}&\Big|\int_{\mathcal{D}}\nu(dx_1)p(x_1,dx_2)f(x_2) - \int_{\mathcal{D}}\nu(dx_1)\mu(dx_2)f(x_2)\Big|\\ &=\Big|\int_{\mathcal{D}}\nu(dx_1)(f_1(x_1,x_2)-1)\mu(dx_2)f(x_2)\Big|.\end{aligned} \tag{A.107}$$

Next let us define $\mathcal{D}^*$ as

$$\mathcal{D}^* := \arg\sup_{\mathcal{D}\in\mathcal{R}^{K_1+K_2}}\int_{\mathcal{D}}\nu(dx_1)(f_1(x_1,x_2)-1)\mu(dx_2)f(x_2) \tag{A.108}$$

such that for each $\mathcal{A} \subset \mathcal{D}^*$, the integration of (A.108) over $\mathcal{A}$ is nonnegative. Then by (A.107), (A.108), and the assumption of $0 \leq f \leq 1$, it holds that

$$\begin{aligned} &\sup_{\mathcal{D} \in \mathcal{R}^{K_1+K_2}} \Big| \int_{\mathcal{D}} \nu(dx_1) p(x_1, dx_2) f(x_2) - \int_{\mathcal{D}} \nu(dx_1) \mu(dx_2) f(x_2) \Big| \\ &\leq \int_{\mathcal{D}^*} \nu(dx_1)(f_1(x_1, x_2) - 1)\mu(dx_2). \end{aligned} \tag{A.109}$$

Thus, it follows from the definition of f_1, Lemma 2, and the definition of the total variation norm that

$$\begin{aligned} &\int_{\mathcal{D}^*} \nu(dx_1)(f_1(x_1, x_2) - 1)\mu(dx_2) \\ &= \int_{\mathbb{R}^{K_1}} \nu(dx_1)(p(x_1, (D^*)_{x_1}) - \mu((D^*)_{x_1})) \\ &\leq \int_{\mathbb{R}^{K_1}} \nu(dx_1) \frac{1}{2} \left\| p(x_1, \cdot) - \mu(\cdot) \right\|_{TV}. \end{aligned} \tag{A.110}$$

Here Lemma 2 is applicable to the integral with μ since μ can be seen as a transition kernel. Therefore, combining (A.109) and (A.110) yields the conclusion in part 2) of Lemma 5. This concludes the proof of Lemma 5.

C.8 Lemma 6 and its proof

Lemma 6. *Let $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i)\}$ and $\{(\boldsymbol{V}_i, \boldsymbol{U}_i)\}$ be two sequences of identically distributed random vectors with $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ k_1-dimensional and k_2-dimensional, respectively. Assume that there exists some positive integer K such that for each $\mathcal{D}_i \in \mathcal{R}^{k_2}$ with $i = 1, \cdots, K$,*

$$\mathbb{P}(\cap_{i=1}^K \{\boldsymbol{X}_i \in \mathcal{D}_i\} \mid \boldsymbol{Y}_j, j = 1, \cdots, K) = \Pi_{i=1}^K \mathbb{P}(\boldsymbol{X}_i \in \mathcal{D}_i \mid \boldsymbol{Y}_i),$$
$$\mathbb{P}(\cap_{i=1}^K \{\boldsymbol{U}_i \in \mathcal{D}_i\} \mid \boldsymbol{V}_j, j = 1, \cdots, K) = \Pi_{i=1}^K \mathbb{P}(\boldsymbol{U}_i \in \mathcal{D}_i \mid \boldsymbol{V}_i).$$

Let us define $\boldsymbol{Y} := (\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_K^T)^T$ and $\boldsymbol{V} := (\boldsymbol{V}_1^T, \cdots, \boldsymbol{V}_K^T)^T$, and $\boldsymbol{X}$ and $\boldsymbol{U}$ are defined similarly. Then it holds that

1) There exists a transition kernel $h : \mathbb{R}^{Kk_1} \times \mathcal{R}^{Kk_2} \longrightarrow [0,1]$ such that for each $\mathcal{B} \in \mathcal{R}^{Kk_2}$, $h(\boldsymbol{Y}, \mathcal{B})$ and $h(\boldsymbol{V}, \mathcal{B})$ are versions of $\mathbb{P}(\boldsymbol{X} \in \mathcal{B} \mid \boldsymbol{Y})$ and $\mathbb{P}(\boldsymbol{U} \in \mathcal{B} \mid \boldsymbol{V})$, respectively.

2) We have

$$\sup_{\mathcal{D} \in \mathcal{R}^{K(k_1+k_2)}} \left| \mathbb{P}\Big((\boldsymbol{Y}^{T}, \boldsymbol{X}^{T}) \in \mathcal{D} \Big) - \mathbb{P}\Big((\boldsymbol{V}^{T}, \boldsymbol{U}^{T}) \in \mathcal{D} \Big) \right| \leq \frac{1}{2} \left\| \mu_{\boldsymbol{Y}} - \mu_{\boldsymbol{V}} \right\|_{TV},$$

where $\mu_{\boldsymbol{Y}}$ and $\mu_{\boldsymbol{V}}$ denote the distributions of $\boldsymbol{Y}$ and $\boldsymbol{V}$, respectively.

Proof. By assumption, it holds that for each $\mathcal{A} \in \mathcal{R}^{Kk_1}$ and $B_i \in \mathcal{R}^{k_2}$ with $i = 1, \cdots, K$,

$$\begin{aligned} \mathbb{P}((\boldsymbol{Y}^{T}, \boldsymbol{X}^{T}) \in \mathcal{A} \times (\bigtimes_{i=1}^{K} B_i)) &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y} \in \mathcal{A}\}}\, \mathbb{P}\Big(\cap_{i=1}^{K} \{\boldsymbol{X}_i \in B_i\} \mid \boldsymbol{Y} \Big) \Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y} \in \mathcal{A}\}}\, \Pi_{i=1}^{K} \mathbb{P}\Big(\boldsymbol{X}_i \in B_i \mid \boldsymbol{Y}_i \Big) \Big], \end{aligned} \tag{A.111}$$

where $\bigtimes_{i=1}^{K} B_i := (B_1, \cdots, B_K)$. Similarly, we can show that

$$\mathbb{P}((\boldsymbol{V}^{T}, \boldsymbol{U}^{T}) \in \mathcal{A} \times (\bigtimes_{i=1}^{K} B_i)) = \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{V} \in \mathcal{A}\}}\, \Pi_{i=1}^{K} \mathbb{P}\Big(\boldsymbol{U}_i \in B_i \mid \boldsymbol{V}_i \Big) \Big]. \tag{A.112}$$

Since $(\boldsymbol{Y}_i^{T}, \boldsymbol{X}_i^{T})$ and $(\boldsymbol{V}_i^{T}, \boldsymbol{U}_i^{T})$ with $i \geq 1$ are identically distributed, by Lemma 4 in Section C.6, there exists a transition kernel h_1 such that for each i and $\mathcal{D} \in \mathcal{R}^{k_2}$, $h_1(\boldsymbol{Y}_i, \mathcal{D})$ and $h_1(\boldsymbol{V}_i, \mathcal{D})$ are versions of $\mathbb{P}(\boldsymbol{X}_i \in \mathcal{D} \mid \boldsymbol{Y}_i)$ and $\mathbb{P}(\boldsymbol{U}_i \in \mathcal{D} \mid \boldsymbol{V}_i)$, respectively.

Let us define $h_2 : \mathbb{R}^{Kk_1} \times \mathcal{R}^{Kk_2} \longrightarrow [0,1]$ such that for each $x = (x_1^{T}, \cdots, x_K^{T})^{T} \in \mathbb{R}^{Kk_1}$, $h_2(x, \cdot)$ is a probability measure such that for each $\mathcal{D}_i \in \mathcal{R}^{k_2}$ with $i = 1, \cdots, K$,

$$h_2(x, \bigtimes_{i=1}^{K} \mathcal{D}_i) = \Pi_{i=1}^{K} h_1(x, \mathcal{D}_i). \tag{A.113}$$

We make a useful claim below.

Claim 1. *h_2 is a transition kernel satisfying (A.113).*

The proof of Claim 1 is provided in Section C.10. Then it follows from (A.111), (A.112), and Claim 1 above that for each $\mathcal{B}_i \in \mathcal{R}^{k_2}$ with $i = 1, \cdots, K$,

$$\begin{aligned}\mathbb{P}((\boldsymbol{Y}^T, \boldsymbol{X}^T) \in \mathcal{A} \times (\bigtimes_{i=1}^{K} B_i)) &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y}\in\mathcal{A}\}}\ \Pi_{i=1}^{K}\mathbb{P}\Big(\boldsymbol{X}_i \in B_i \mid \boldsymbol{Y}_i\Big)\Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y}\in\mathcal{A}\}}\ \Pi_{i=1}^{K} h_1\Big(\boldsymbol{Y}_i, B_i\Big)\Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y}\in\mathcal{A}\}}\ h_2\Big(\boldsymbol{Y}, \bigtimes_{i=1}^{K} B_i\Big)\Big],\end{aligned} \tag{A.114}$$

and similarly,

$$\mathbb{P}((\boldsymbol{V}^T, \boldsymbol{U}^T) \in \mathcal{A} \times (\bigtimes_{i=1}^{K} B_i)) = \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{V}\in\mathcal{A}\}}\ h_2\Big(\boldsymbol{V}, \bigtimes_{i=1}^{K} B_i\Big)\Big]. \tag{A.115}$$

By the construction of h_2, (A.114), (A.115), and Lemma 8 in Section C.12, it holds that for each $\mathcal{A} \in \mathcal{R}^{Kk_1}$ and $\mathcal{B} \in \mathcal{R}^{Kk_2}$,

$$\begin{aligned}\mathbb{P}((\boldsymbol{Y}^T, \boldsymbol{X}^T) \in \mathcal{A} \times \mathcal{B})) &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{Y}\in\mathcal{A}\}}\ h_2\Big(\boldsymbol{Y}, \mathcal{B}\Big)\Big], \\ \mathbb{P}((\boldsymbol{V}^T, \boldsymbol{U}^T) \in \mathcal{A} \times \mathcal{B})) &= \mathbb{E}\Big[\mathbf{1}_{\{\boldsymbol{V}\in\mathcal{A}\}}\ h_2\Big(\boldsymbol{V}, \mathcal{B}\Big)\Big].\end{aligned} \tag{A.116}$$

Since h_2 is a transition kernel, we see that for each $\mathcal{B} \in \mathcal{R}^{Kk_2}$, $h_2(\boldsymbol{Y}, \mathcal{B})$ is $\sigma(\boldsymbol{Y})$-measurable. Thus, in view of (A.116) we have that for each $\mathcal{B} \in \mathcal{R}^{Kk_2}$, $h_2(\boldsymbol{Y}, \mathcal{B})$ is a version of $\mathbb{P}(\boldsymbol{X} \in \mathcal{B} \mid \boldsymbol{Y})$. A similar result for $\boldsymbol{V}$ and $\boldsymbol{U}$ can also be obtained, which leads to the first assertion.

Finally, by Lemma 2 and Lemma 5 in Sections C.4 and C.7, respectively, we can deduce that for each $\mathcal{D} \in \mathcal{R}^{K(k_1+k_2)}$,

$$\begin{aligned}&\Big|\mathbb{P}((\boldsymbol{Y}^T, \boldsymbol{X}^T) \in \mathcal{D}) - \mathbb{P}((\boldsymbol{V}^T, \boldsymbol{U}^T) \in \mathcal{D})\Big| \\ &= \Big| \int_{\mathbb{R}^{Kk_1}} h_2(x, \mathcal{D}) \mu_{\boldsymbol{Y}}(dx) - \int_{\mathbb{R}^{Kk_1}} h_2(x, \mathcal{D}) \mu_{\boldsymbol{V}}(dx) \Big| \\ &\leq \frac{1}{2} \|\mu_{\boldsymbol{Y}} - \mu_{\boldsymbol{U}}\|_{TV},\end{aligned} \tag{A.117}$$

which yields the second assertion. This completes the proof of Lemma 6.

C.9 Lemma 7 and its proof

Let $\widetilde{\boldsymbol{U}}$ and $\widetilde{\boldsymbol{V}}$ be the knockoffs counterparts of $r \times c$ design matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. The corresponding response vectors are denoted as $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Lemma 7 below ensures that the knockoffs matrix construction does not cause additional variation in the distribution in terms of the total variation distance.

Lemma 7. *Assume that* 1) *the rows of* $(\boldsymbol{U}, \widetilde{\boldsymbol{U}})$ *and* $(\boldsymbol{V}, \widetilde{\boldsymbol{V}})$ *are identically distributed,* 2) *the row vectors of* $\widetilde{\boldsymbol{U}}$ *are independent random vectors conditional on* $\boldsymbol{U}$, 3) *the ith row of* $\widetilde{\boldsymbol{U}}$ *is independent of the other rows of* $\boldsymbol{U}$ *conditional on the ith row of* $\boldsymbol{U}$, *and* 4) $\boldsymbol{u}$ *is independent of* $\widetilde{\boldsymbol{U}}$ *conditional on* $\boldsymbol{U}$. *In addition, we assume the same for* $(\boldsymbol{v}, \boldsymbol{V}, \widetilde{\boldsymbol{V}})$. *Then it holds that*

$$\sup_{\mathcal{D} \in \mathcal{R}^{r(1+2c)}} \Big|\mathbb{P}\Big((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}\Big) - \mathbb{P}\Big((\boldsymbol{v}, \boldsymbol{V}, \widetilde{\boldsymbol{V}}) \in \mathcal{D}\Big)\Big| \leq \frac{1}{2} \|\mu_{\boldsymbol{u},\boldsymbol{U}} - \mu_{\boldsymbol{v},\boldsymbol{V}}\|_{TV}$$

and

$$\sup_{\mathcal{D} \in \mathcal{R}^{2rc}} \Big|\mathbb{P}\Big((\boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}\Big) - \mathbb{P}\Big((\boldsymbol{V}, \widetilde{\boldsymbol{V}}) \in \mathcal{D}\Big)\Big| \leq \frac{1}{2} \|\mu_{\boldsymbol{U}} - \mu_{\boldsymbol{V}}\|_{TV}.$$

Proof. We start with showing the first assertion. Denote the ith rows of $\widetilde{\boldsymbol{U}}$ and $\boldsymbol{U}$ by $\widetilde{\boldsymbol{U}}_{i\bullet}$

and $\boldsymbol{U}_{i\bullet}$, respectively. Then for each $\mathcal{D}_i \in \mathcal{R}^r$, we have that

$$\begin{aligned}\mathbb{P}(\cap_{i=1}^c\{\widetilde{\boldsymbol{U}}_{i\bullet} \in \mathcal{D}_i\}|\boldsymbol{U}) &= \Pi_{i=1}^c \mathbb{P}(\widetilde{\boldsymbol{U}}_{i\bullet} \in \mathcal{D}_i|\boldsymbol{U}) \\ &= \Pi_{i=1}^c \mathbb{P}(\widetilde{\boldsymbol{U}}_{i\bullet} \in \mathcal{D}_i|\boldsymbol{U}_{i\bullet}),\end{aligned}$$

where the first equality follows from assumption 2) and the second one is due to assumption 3). By this, an application of Lemma 6 shows that there exists a transition kernel $h : \mathbb{R}^{rc} \times \mathcal{R}^{rc} \longmapsto [0,1]$ such that for each $\mathcal{D} \in \mathcal{R}^{rc}$, $h(\boldsymbol{U}, \mathcal{D})$ and $h(\boldsymbol{V}, \mathcal{D})$ are versions of $\mathbb{P}(\widetilde{\boldsymbol{U}} \in \mathcal{D} \mid \boldsymbol{U})$ and $\mathbb{P}(\widetilde{\boldsymbol{V}} \in \mathcal{D} \mid \boldsymbol{V})$, respectively.

We will make use of the claim below.

Claim 2. *For each $\mathcal{D} \in \mathcal{R}^{r(1+2c)}$, it holds that*

$$\begin{aligned}\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}) &= \int_{\mathbb{R}^{r(1+c)}} h(x_2, \mathcal{D}_{x_1,x_2})\, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2), \\ \mathbb{P}((\boldsymbol{v}, \boldsymbol{V}, \widetilde{\boldsymbol{V}}) \in \mathcal{D}) &= \int_{\mathbb{R}^{r(1+c)}} h(x_2, \mathcal{D}_{x_1,x_2})\, \mu_{\boldsymbol{v},\boldsymbol{V}}(dx_1 \times dx_2),\end{aligned}$$

where x_1 and x_2 denote r-dimensional and (rc)-dimensional vectors, respectively.

The proof of Claim 2 is presented in Section C.11. Then it follows from Claim 2 above, Lemma 5, and the fact of $0 \leq h \leq 1$ that for each $\mathcal{D} \in \mathcal{R}^{r(1+2c)}$,

$$\begin{aligned}&\left|\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}) - \mathbb{P}((\boldsymbol{v}, \boldsymbol{V}, \widetilde{\boldsymbol{V}}) \in \mathcal{D})\right| \\ &= \left|\int_{\mathbb{R}^{r(1+c)}} h(x_2, \mathcal{D}_{x_1,x_2})\, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2) - \int_{\mathbb{R}^{r(1+c)}} h(x_2, \mathcal{D}_{x_1,x_2})\, \mu_{\boldsymbol{v},\boldsymbol{V}}(dx_1 \times dx_2)\right| \\ &\leq \frac{1}{2}\left\|\mu_{\boldsymbol{u},\boldsymbol{U}} - \mu_{\boldsymbol{v},\boldsymbol{V}}\right\|_{TV},\end{aligned}$$

which leads to the conclusion in the first assertion. Using similar arguments as above, we can establish the second assertion. This concludes the proof of Lemma 7.

C.10 Proof of Claim 1

Note that given x, the existence of the probability measure $h_2(x, \cdot)$ is guaranteed by an application of Theorem 1.7.1 in [16]. Let L be the collection of sets such that if $\mathcal{D} \in L$, $h_2(x, \mathcal{D})$ is a measurable function of x. We will make three observations due to the definition of h_2. (i) L contains all Borel rectangles since the product of measurable functions is still a measurable function. (ii) For $\mathcal{D}_1, \mathcal{D}_2 \in L$ with $\mathcal{D}_1 \subset \mathcal{D}_2$, it holds that $h_2(\cdot, \mathcal{D}_2 \backslash \mathcal{D}_1)$ is measurable and hence $\mathcal{D}_2 \backslash \mathcal{D}_1 \in L$. (iii) For $\mathcal{D}_i \subset \mathcal{D}_{i+1}$, $\mathcal{D}_i \in L$, and $\mathcal{D} := \cup_{i=1}^{\infty} \mathcal{D}_i$, we have

$$h_2(\cdot, \mathcal{D}) = \sup_i h_2(\cdot, \mathcal{D}_i)$$

which is measurable, and thus $\mathcal{D} \in L$. Therefore, it follows from these facts and Lemma 8 in Section C.12 that for each $\mathcal{D} \in \mathcal{R}^{Kk_2}$, $h_2(\cdot, \mathcal{D})$ is measurable. This completes the proof of Claim 1.

C.11 Proof of Claim 2

Let us first show the first assertion. For each Borel rectangle $\mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{A}_3 \in \mathcal{R}^r \times \mathcal{R}^{rc} \times \mathcal{R}^{rc}$, it holds that

$$\begin{aligned}
&\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{A}_3) = \mathbb{E}\Big[\mathbf{1}_{\{(\boldsymbol{u}, \boldsymbol{U}) \in \mathcal{A}_1 \times \mathcal{A}_2\}}\, \mathbb{P}(\widetilde{\boldsymbol{U}} \in \mathcal{A}_3 \mid \boldsymbol{u}, \boldsymbol{U})\Big] \\
&= \mathbb{E}\Big[\mathbf{1}_{\{(\boldsymbol{u}, \boldsymbol{U}) \in \mathcal{A}_1 \times \mathcal{A}_2\}}\, \mathbb{P}(\widetilde{\boldsymbol{U}} \in \mathcal{A}_3 \mid \boldsymbol{U})\Big] \\
&= \int_{\mathbb{R}^{r(1+c)}} h(x_2, (\mathcal{A}_1 \times \mathcal{A}_2 \times \mathcal{A}_3)_{x_1, x_2})\, \mu_{\boldsymbol{u}, \boldsymbol{U}}(dx_1 \times dx_2),
\end{aligned}$$

where the second equality is due to the assumption of Lemma 7 and the last equality is because of the definition of h. Let L be a collection of sets in $\mathcal{R}^{r(1+2c)}$ such that if $\mathcal{D} \in L$,

$$\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}) = \int_{\mathbb{R}^{r(1+c)}} h(x_2, \mathcal{D}_{x_1, x_2})\, \mu_{\boldsymbol{u}, \boldsymbol{U}}(dx_1 \times dx_2).$$

Then we can see that L contains all Borel rectangles which collection is a π-system.

Let us make a few observations below.

1) The set L contains $\mathbb{R}^{r(1+2c)}$.

2) If $\mathcal{D}_1, \mathcal{D}_2 \in L$ and $\mathcal{D}_1 \subset \mathcal{D}_2$, then some basic measure and integration operations as well as the operation of $\mathcal{D}_x$ give that

$$\begin{aligned}
&\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}_2 \backslash \mathcal{D}_1) = \mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}_2) - \mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}_1) \\
&= \int_{\mathbb{R}^{r(1+c)}} \Big(h(x_2, (\mathcal{D}_2)_{x_1,x_2}) - h(x_2, (\mathcal{D}_1)_{x_1,x_2}) \Big) \, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2) \\
&= \int_{\mathbb{R}^{r(1+c)}} h(x_2, (\mathcal{D}_2 \backslash \mathcal{D}_1)_{x_1,x_2}) \, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2),
\end{aligned}$$

which leads to $\mathcal{D}_2 \backslash \mathcal{D}_1 \in L$.

3) If $\mathcal{D}_n \in L$ and $\mathcal{D}_n \subset \mathcal{D}_{n+1}$, then it follows from the continuity of measure and the monotone convergence theorem that

$$\begin{aligned}
&\mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \cup_n \mathcal{D}_n) = \lim_n \mathbb{P}((\boldsymbol{u}, \boldsymbol{U}, \widetilde{\boldsymbol{U}}) \in \mathcal{D}_n) \\
&= \lim_n \int_{\mathbb{R}^{r(1+c)}} h(x_2, (\mathcal{D}_n)_{x_1,x_2}) \, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2) \\
&= \int_{\mathbb{R}^{r(1+c)}} h(x_2, (\cup_n \mathcal{D}_n)_{x_1,x_2}) \, \mu_{\boldsymbol{u},\boldsymbol{U}}(dx_1 \times dx_2),
\end{aligned}$$

which results in $\cup_n \mathcal{D}_n \in L$.

Therefore, using the aforementioned facts, an application of Lemma 8 in Section C.12 yields the conclusion in the first assertion. The conclusion in the second assertion can be shown in a similar fashion, which concludes the proof of Claim 2.

C.12 Lemma 8

Definition (π-system and λ-system)**.** *A collection of sets P is said to be a π-system if for any $A, B \in P$, $A \cap B \in P$. A collection L of sets in Ω is said to be a λ-system if*

1) $\Omega \in L$*;*

2) If $A \subset B$ *and* $A, B \in L$*, then* $B \backslash A \in L$*;*

3) If $A_n \in L$ *and* $A_n \subset A_{n+1}$*, then* $\cup_n A_n \in L$*.*

Lemma 8 ($\pi - \lambda$ Theorem in [16])**.** *If* P *is a* π*-system and* L *is a* λ*-system that contains* P*, then the smallest* σ*-algebra containing* P *is also contained in* L*.*