High-Dimensional Knockoffs Inference (part II)

Jinchi Lv
Data Sciences and Operations Department
Marshall School of Business
University of Southern California

http://faculty.marshall.usc.edu/jinchi-lv

USC Summer School on Uncertainty Quantification (08/09/2024)

Outline of Chi, Fan, Ing and L. (2024)

Outline of Chi, Fan, Ing and L. (2024)

- An economic forecasting example
- Time series knockoffs inference
- Numerical studies
- Theoretical justifications

Inflation prediction

Inflation prediction

- Identifying key economic factors that can influence inflation is a long-standing research pursuit (King et al., 1995, Stock and Watson, 1999, Crump et al., 2022)

- Main challenges are
  - serial dependence
  - large number of potentially important covariates (time series covariates, their lags, and non-time series covariates)
  - nonlinearities attributed to the regime shifts and structural changes (Hamilton [1989], Tong and Lim [1980])

The US inflation series

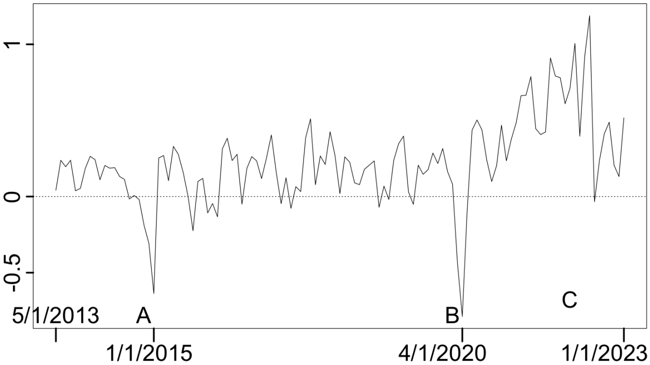


Figure 1: The U.S. inflation from May 2013 to January 2023. Number of potential time series covariates $p = 127$ (e.g., consumer price indices, unemployment rates, exchange rates, housing indices, stock market indices,...

Rolling window prediction

Pros: mitigate the effects of nonlinearity and nonstationarity

Cons: sample size in each window is usually small

- Small sample size, together with serial dependence, presence of nonlinearity, nonstationarity, and high-dimensional covariates, makes practical inference highly challenging for time series data

- Goal: develop a reliable variable selection approach specifically tailored for addressing high-dimensional time series data

Variable selection with high dimensional time series

- Problem setup:
  - Scalar time series response $\{y_t\}_{1 \le t \le n}$
  - Covariate vector $\{\boldsymbol{x}_t\}_{1 \le t \le n}$ with $\boldsymbol{x}_t \in \mathbb{R}^p$ containing time series covariates, lags, and possibly non-timeseries covariates
  - p is of high dimension
  - $(\boldsymbol{x}_t, y_t)$ is stationary across t
- Assumption: there exists $S^* \subset \{1, \cdots, p\}$ such that

$$y_t \perp\!\!\!\perp \boldsymbol{x}_{t,S^{*c}} \big| \boldsymbol{x}_{t,S^*}$$

- Stationarity guarantees that S^* is independent of t
- Goal: developing an algorithm that estimates S^* by $\widehat{S}$ such that

$$\mathrm{FDR} = \mathbb{E}\left[\frac{|(S^*)^c \cap \widehat{S}|}{|\widehat{S}|}\right] \le \tau^*$$

Related literature

- *False discovery rate* (*FDR*) has been widely used for *variable selection error rate control* (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Fan, Han and Gu, 2012; ...)

- *Valid p-values* are needed for the BH or BY framework

- Remains *largely unclear* how to construct justified p-values for many popular *nonparametric learning* tools (e.g. *random forests and deep neural networks*)

Review of Model-X Knockoffs Inference

Model-X Knockoffs Framework

- Introduced in Candès, Fan, Janson and L. (2018)
  - Bypass the use of p-values to achieve FDR control
  - Model-free: any model for the conditional dependence $Y|X_1, \cdots, X_p$
  - Dimension free: any dimension (including $p > n$)
  - Known covariate distribution: joint distribution of $\mathbf{x} = (X_1, \cdots, X_p)$ known
  - Theoretically guaranteed to have finite-sample FDR control
- Intuition:
  - Generate "fake" copies of original covariates which are irrelevant to Y but mimics the dependence structure of original covariates
  - Act as controls for assessing importance of original variables

Model-X Knockoff Variables

Definition 1 (Candès, Fan, Janson and L., 18)

Model-X knockoffs for the family of random variables $\mathbf{x} = (X_1, \cdots, X_p)^T$ are a new family of random variables $\widetilde{\mathbf{x}} = (\widetilde{X}_1, \cdots, \widetilde{X}_p)^T$ constructed such that

- for any subset $S \subset \{1, \cdots, p\}$,

$$(\mathbf{x}^T, \widetilde{\mathbf{x}}^T)_{\mathrm{swap}(S)} \stackrel{d}{=} (\mathbf{x}^T, \widetilde{\mathbf{x}}^T)$$

- $\widetilde{\mathbf{x}} \perp\!\!\!\perp y | \mathbf{x}$

The Knockoffs Statistics

Knockoff statistics $W_j = f_j(\mathbf{y}, [\mathbf{X}, \widetilde{\mathbf{X}}])$ are variable importance measures

- Positive W_j: original more important, strength measured by magnitude
- Null variables: W_j should be symmetric around 0
- Eg: Lasso Coefficient Difference $W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}|$

Set of selected variables

$$\widehat{S} = \{j : W_j > T\}$$

Choice of Threshold

Intuition of FDR control

$$\begin{aligned}
\text{FDR} &= \mathbb{E}\left[\frac{\#\text{selected null variables}}{\#\text{selcted variables}}\right] \\
&= \mathbb{E}\left[\frac{\#\{\text{null } W_j \geq T\}}{\#\{W_j \geq T\}}\right] \\
&\approx \mathbb{E}\left[\frac{\#\{\text{null } -W_j \geq T\}}{\#\{W_j \geq T\}}\right] \\
&\leq \mathbb{E}\left[\frac{\#\{-W_j \geq T\}}{\#\{W_j \geq T\}}\right]
\end{aligned}$$

This suggests to choose the threshold T by examining the ratio

$$\frac{\#\{-W_j \geq T\}}{\#\{W_j \geq T\}}$$

Theorem 2 (Candès, Fan, Janson and L., 18)
Letting

$$T_+ = \min\left\{t > 0 : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\}} \le \tau^*\right\} \quad (\textbf{\textit{Knockoffs+}})$$

and setting $\widehat{S} = \{j : W_j \ge T_+\}$*, controls the usual FDR,*

$$\mathbb{E}\left[\frac{|\widehat{S} \cap S^*|}{|\widehat{S}| \vee 1}\right] \le \tau^*.$$

Time Series Data

Challenges with time series data

$$[\mathbf{y}, \mathbf{X}] = \begin{bmatrix} y_1, & \mathbf{x}_1 \\ y_2, & \mathbf{x}_2 \\ \vdots & \\ y_n, & \mathbf{x}_n \end{bmatrix} \Big\downarrow \text{time}$$

- Model-X knockoffs assumes
  - Row independence (i.e., no serial correlation)
  - Known covariate distribution for $\mathbf{x}_t$
- Too strong for time series data!

- Row independence $\Longrightarrow$ knockoff variables can be generated in a row-wise fashion independent of other rows

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \begin{matrix} \longrightarrow \\ \longrightarrow \\ \vdots \\ \longrightarrow \end{matrix} \begin{bmatrix} \widetilde{\mathbf{x}}_1 \\ \widetilde{\mathbf{x}}_2 \\ \vdots \\ \widetilde{\mathbf{x}}_n \end{bmatrix}$$

Question: is this row-wise construction of knockoff variables still valid for time series data with serial dependence?

- Consider the example where

$$\mathbf{x}_t = (y_{t-1}, \cdots, y_{t-p})^T,$$

  then knowing covariate distribution of $\mathbf{x}_t$ leads to known stationary distribution of the time series, rendering the variable selection invalid!

Question: how to relax the assumption of known covariate distribution?

The TSKI Procedure

Three key ingredients

- Subsample: to overcome the difficulty caused by serial dependence, we consider subsamples

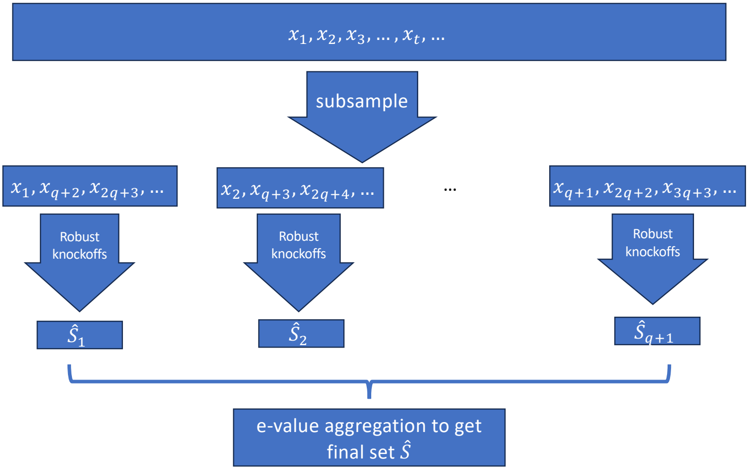
$$H_k = \{k + s(q+1) : s = 0, 1, \cdots, \lfloor \frac{n-k}{q+1} \rfloor\}$$

  for $k \in \{1, \cdots, q+1\}$

- Robust knockoffs: to accommodate unknown covariate distribution, on each subsample H_k, we apply the robust knockoffs inference (Barber, Candès and Samworth, 2020), yielding a set of selected variables $\widehat{S}_k$

- Ensemble: produce an ensemble inference using the e-value method (Wang and Ramdas (2022))

TSKI



$x_1, x_2, x_3, \ldots, x_t, \ldots$
subsample
$x_1, x_{q+2}, x_{2q+3}, \ldots$
$x_2, x_{q+3}, x_{2q+4}, \ldots$
…
$x_{q+1}, x_{2q+2}, x_{3q+3}, \ldots$
Robust knockoffs
Robust knockoffs
Robust knockoffs
$\hat{S}_1$
$\hat{S}_2$
$\hat{S}_{q+1}$
e-value aggregation to get final set $\hat{S}$

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 (2) 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[a]

$$
\begin{aligned}
e_j^k &= \frac{p \times \mathbf{1}_{\{W_j^k \geq T^k\}}}{1 + \sum_{s=1}^{p} \mathbf{1}_{\{W_s^k \leq -T^k\}}}, \\
T^k &= \min\left\{ t \in \mathcal{W}_+^k : \frac{1 + \#\{j : W_j^k \leq -t\}}{\#\{j : W_j^k \geq t\} \vee 1} \leq \tau_1 \right\},
\end{aligned}
\tag{3}
$$

   and $\mathcal{W}_+^k = \{|W_s^k| : |W_s^k| > 0\}$ for each $k \in \{1, \cdots, q+1\}$.
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 $e_{(1)} \geq \cdots \geq e_{(p)}$.

[a] $\min \emptyset$ and $\max \emptyset$ are defined as ∞ and 0, respectively.

Subsampling effectiveness

- **Condition 1** (*h-step β-mixing with exponential decay*). Assume that process $\{\boldsymbol{z}_t\}$ is an m-dimensional stationary Markov chain with a transition kernel $p(\cdot, \cdot)$ and stationary distribution π. There exist a measurable function $V : \mathbb{R}^m \to [0, \infty)$ and some constants $0 \leq \rho < 1$ and $C_0 > 0$ such that for each $\mathbf{x} \in \mathbb{R}^m$,

$$||p^h(\mathbf{x}, \cdot) - \pi(\cdot)||_{TV} \leq C\rho^h V(\mathbf{x}),$$

  where $p^h(\cdot, \cdot)$ denotes the h-step transition kernel, $\| \cdot \|_{TV}$ represents the total variation distance, and $C > 0$ is some constant

- Can hold for many popular time series models

Example: ARX process

Proposition 1

Let $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-k_1}, \boldsymbol{h}_t^T)^T$ *be from* p_h*-dimensional autoregressive models with exogenous variables (ARX)*

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

where $H_t^{(l)} = \epsilon_t^{(l)} + \sum_{j=1}^{k_{3l}} b_j^{(l)} H_{t-j}^{(l)}$, *and* ε_t *and* $\epsilon_t^{(l)}$ *are Gaussian. Assume that for some constant* $C_2 > 0$ *and sufficiently small* $s_2 > 0$,

$$\sup_{h>0}\{p_h \exp(-s_2 h)\} \leq C_2. \tag{1}$$

Then under some regularity conditions of the regression coefficients, $\{\boldsymbol{x}_t^{(h)}\}$ *satisfies Condition 1 with h-step.*

Suggested h: $h \sim (\log p_h)^{1+\delta}$

Robust knockoffs inference

Relaxing the known covariate distribution assumption:

- $\kappa_j : \mathbb{R}^{p-1} \times \mathcal{R}^{p-1} \longmapsto \mathbb{R}$: coordinatewise knockoff generator for each $j \in \{1, \cdots, p\}$ (*approximate the distribution* $X_j | \boldsymbol{x}_{-j}$)

- $\kappa(\mathbf{x}, \cdot)$: knockoff generator used to generate knockoff variables $\widetilde{\mathbf{x}}$

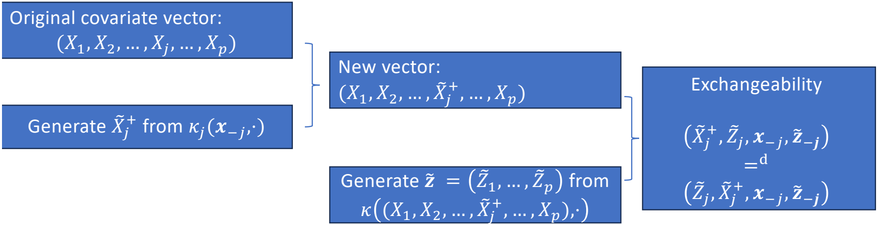
- **Condition 2**. The knockoff generator $\kappa(\cdot, \cdot)$ is independent of training data $\{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$

- **Condition 3** (*pairwise exchangeability*). For each $1 \leq j \leq p$, 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 with $\widetilde{X}_j^\dagger$ sampled from $\kappa_j(\boldsymbol{x}_{-j}, \cdot)$ (a conditional distribution that approximates that of $X_j | \boldsymbol{x}_{-j}$)

Condition 3 illustration

Recall:

- $\kappa_j : \mathbb{R}^{p-1} \times \mathcal{R}^{p-1} \longmapsto \mathbb{R}$: coordinatewise knockoff generator for each $j \in \{1, \cdots, p\}$
- $\kappa(\mathbf{x}, \cdot)$: knockoff generator used to generate knockoff variables $\widetilde{\mathbf{x}}$



Condition 3 illustration

Original covariate vector:
$(X_1, X_2, \dots, X_j, \dots, X_p)$

Generate $\tilde{X}_j^+$ from $\boxed{\kappa_j(\boldsymbol{x}_{-j}, \cdot)}$

Approximate conditional distribution of $X_j | \boldsymbol{x}_{-j}$

New vector:
$(X_1, X_2, \dots, \tilde{X}_j^+, \dots, X_p)$

Generate $\tilde{\boldsymbol{z}} = (\tilde{Z}_1, \dots, \tilde{Z}_p)$ from $\kappa((X_1, X_2, \dots, \tilde{X}_j^+, \dots, X_p), \cdot)$

Exchangeability

$$(\tilde{X}_j^+, \tilde{Z}_j, \boldsymbol{x}_{-j}, \tilde{\boldsymbol{z}}_{-j}) \stackrel{d}{=} (\tilde{Z}_j, \tilde{X}_j^+, \boldsymbol{x}_{-j}, \tilde{\boldsymbol{z}}_{-j})$$

- In implementation, the knockoff variables are generated as $\kappa(\mathbf{x}, \cdot)$
  - Eg: in Gaussian case, $\widetilde{\boldsymbol{x}}|\boldsymbol{x} = (I_p - s\widehat{\Sigma^{-1}})\boldsymbol{x} + (2sI_p - s^2\widehat{\Sigma^{-1}})\mathbf{z}$, $\mathbf{z} \sim N(0, I_p)$

- Only $\kappa(\cdot, \cdot)$ is needed for implementation; κ_j's are only needed for theoretical derivation

- Barber, Candès and Samworth (2020) showed that under the approximate exchangeability, model-X knockoffs achieves the approximate FDR control under the i.i.d. data condition

- Question: how does the serial dependence affect the robust knockoffs procedure?

E-value aggregation

Recall:

- Subsampling yields $q+1$ sets of selected variables
- Directly taking union or intersection does not guarantee FDR control
- We will adopt the idea of e-value aggregation

E-value

- Given a null hypothesis, we call a non-negative random variable E an "e-value" if $\mathbb{E}[E] \le 1$ under the null

- To test a hypothesis at level α, we can reject the null hypothesis when $E \ge 1/\alpha$

- With ideal knockoffs generated from the true covariate distribution, Ren and Barber (2024) showed that

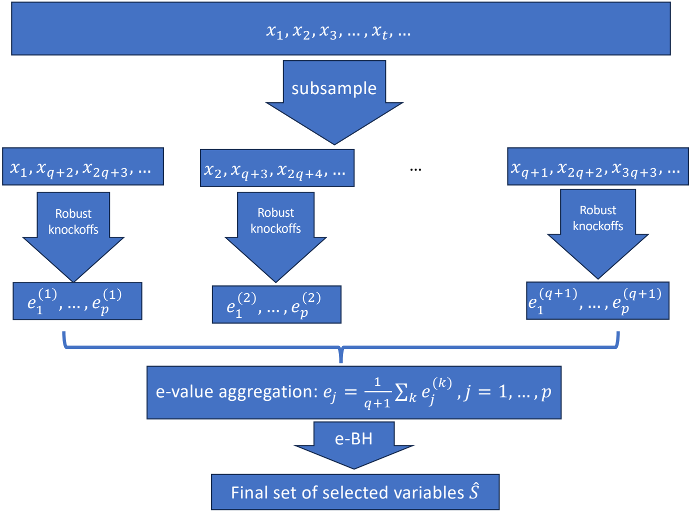
$$e_j = \frac{p \times 1_{\{W_j > T\}}}{1 + \sum_{l=1}^{p} 1_{\{W_l \le -T\}}}$$

are (relaxed) e-values, and that the e-BH procedure (Wang and Ramdas (2022)) achieves FDR control in multiple testing

$$\widehat{S} = \{j : e_j \ge p(\tau^* \widehat{k})^{-1}\} \text{ with } \widehat{k} = \max\{k : e_{(k)} \ge p(\tau^* k)^{-1}\}$$

TSKI using e-value aggregation

The average of multiple e-values is still an e-value



Question: Are these still valid e-values that can guarantee FDR control in the existence of serial dependence?

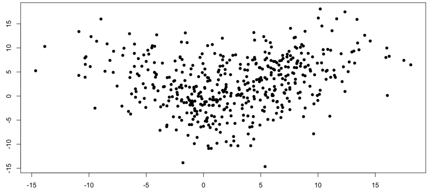
Simulation Studies

SETARX model

For each integer t and $\iota \in \{0, 5\}$, we define

$$Y_t = \begin{cases} \sum_{j=1}^{2} (-0.5)^{j-1} \beta Y_{t-j} + 0.6(\sum_{j=1}^{\iota} H_{t,j} + \sum_{j=\iota+1}^{15} H_{t,j}) + \varepsilon_t, & \text{if } Y_{t-1} > 0.7, \\ \sum_{j=1}^{2} -(-0.5)^{j-1} \beta Y_{t-j} + 0.6(-\sum_{j=1}^{\iota} H_{t,j} + \sum_{j=\iota+1}^{15} H_{t,j}) + \varepsilon_t, & \text{otherwise,} \end{cases}$$

- $\{\varepsilon_t\} \sim_{i.i.d.} N(0, 1)$
- $\beta = 0.7$
- $H_{t,j} = \eta \times H_{t-1,j} + \epsilon_{t,j}$ with $j \in \{1, \cdots, 50\}$ and $\eta = 0.2$
- $\boldsymbol{x}_t = (Y_{t-1}, \cdots, Y_{t-20}, \boldsymbol{h}_t, \boldsymbol{h}_{t-1}, \boldsymbol{h}_{t-2}, \boldsymbol{h}_{t-3}, \boldsymbol{h}_{t-4})$ with $\boldsymbol{h}_t = (H_{t,1}, \cdots, H_{t,50})$, giving rise to $p = 270$.



TSKI performance

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-MDA		0	0.173	0.456		0	0.157	0.718
TSKI-MDA		1	0.026	0.028		1	0.041	0.102
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-MDA		0	0.138	0.215		0	0.196	0.506
TSKI-MDA		1	0.012	0.011		1	0.038	0.036

Method	$n/p/\eta/\iota$	q	FDR	Power
TSKI-LCD	500/270/0.2/0	0	0.176	0.939
TSKI-LCD		1	0.099	0.872
TSKI-MDA		0	0.181	0.922
TSKI-MDA		1	0.092	0.550
TSKI-LCD	500/270/0.2/5	0	0.141	0.634
TSKI-LCD		1	0.086	0.267
TSKI-MDA		0	0.166	0.679
TSKI-MDA		1	0.084	0.216

Comparing with BH and Adaptive Lasso

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

Real Data Application

Real Data Application

- Investigating the temporal relation between the (*one month ahead*) monthly inflation and other macroeconomic time series

- The popular FRED-MD data set (Jurado, Ludvigson and Ng, 2015; McCracken and Ng, 2016; Medeiros and Mendes, 2016)

- Covariates include 127 other monthly macroeconomic variables and their first lags in an AR(2) model (*with feature dimensionality* $p = 254$)

- The one month ahead inflation as response variable

Continued

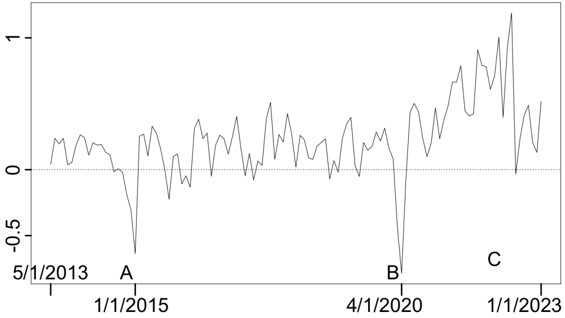


Figure 1: The U.S. inflation from May 2013 to January 2023.

- Inflation series from May 2013 to January 2023 (as response)

- Five-year rolling windows for analysis of versatile time varying patterns (sample size $n = 60$ for each window)

Continued

- Applied TSKI-LCD with target FDR level $\tau^* = 0.2$ and $q = 0, 1$
- Identified some important covariates around 2020 when COVID-19 pandemic began

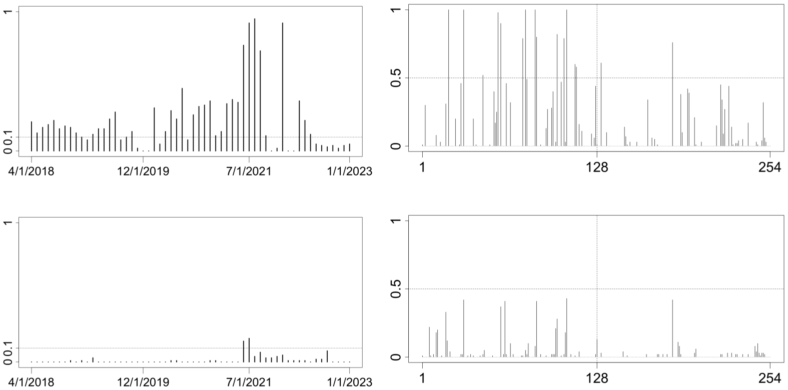


Figure 2: The left panel displays the percentage of times for "having any selections" indicators over 100 repetitions. The right panel shows the percentage of times of being selected for each covariate.

Continued

- Choice of $q > 0$ (*with subsampling*) indeed more conservative in terms of FDR control compared to that of $q = 0$ (*without subsampling*)

- Three frequently selected variables are ACOGNO (*number of new orders for consumer goods*), EXCAUSx (*U.S./Canada exchange rate*), and CLAIMSx (*U.S. initial claims for unemployment benefits*)

- COVID-19 pandemic has much stronger effects on the U.S. economy than the inflation drop in 2015 due to oil supply shock

Continued

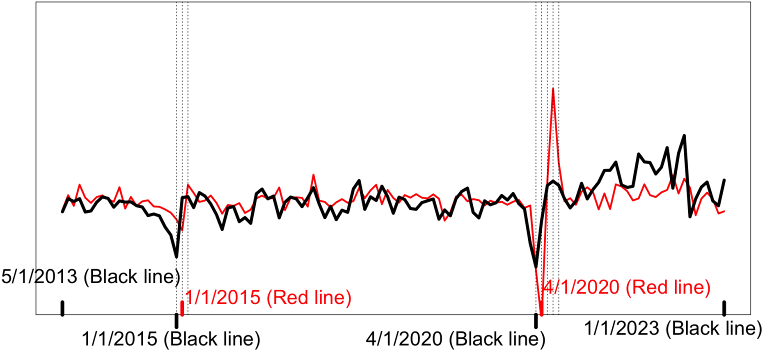


Figure 3: number of new orders for consumer goods

Continued

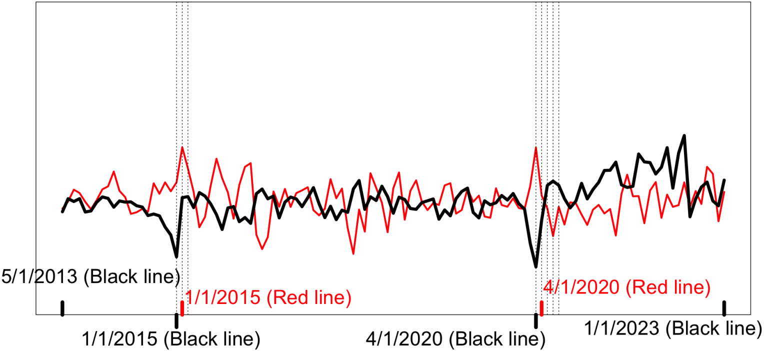


Figure 4: U.S./Canada exchange rate

Continued

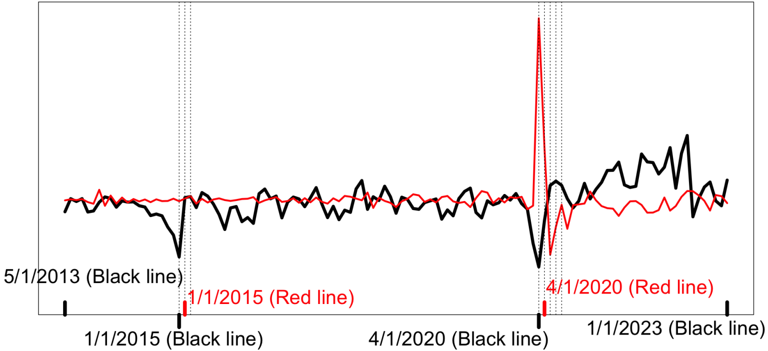


Figure 5: U.S. initial claims for unemployment benefits

Theoretical Justifications

Theorem 1. Under some regularity conditions, we have

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

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

$$\widehat{\mathrm{KL}}_j^{k\pi} = \sum_{i\in H_k} \log \left(\frac{f_{X_j|\boldsymbol{x}_{-j}}(X_{ij}^{\pi}|\boldsymbol{x}_{-ij}^{\pi}) f_{\widetilde{X}_j^{\dagger}|\boldsymbol{x}_{-j}}(\widetilde{X}_{ij}^{\pi}|\boldsymbol{x}_{-ij}^{\pi})}{f_{X_j|\boldsymbol{x}_{-j}}(\widetilde{X}_{ij}^{\pi}|\boldsymbol{x}_{-ij}^{\pi}) f_{\widetilde{X}_j^{\dagger}|\boldsymbol{x}_{-j}}(X_{ij}^{\pi}|\boldsymbol{x}_{-ij}^{\pi})} \right) \tag{3}$$

- $\{\boldsymbol{x}_i^{\pi}, \widetilde{\boldsymbol{x}}_i^{\pi}, Y_i^{\pi}\}_{i=1}^n$ an i.i.d. counterpart of $\{\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i, Y_i\}_{i=1}^n$
- $\mathcal{X}_k = \{\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i, Y_i\}_{i\in H_k}$ and $\mathcal{X}_k^{\pi} = \{\boldsymbol{x}_i^{\pi}, \widetilde{\boldsymbol{x}}_i^{\pi}, Y_i^{\pi}\}_{i\in H_k}$ for each $k \in \{1, \cdots, q+1\}$

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

- The red term is caused by the misspecified conditional distribution $X_j | \mathbf{x}_{-j}$; matches the result in (Barber, Candès and Samworth, 2020)
  - If no misspecification, the red term becomes 0, and $\varepsilon = 0$

- The blue term is caused by the serial dependence after subsampling

Continued

Corollary. If $\{\boldsymbol{x}_i\}_{i \geq 1}$ satisfies Condition 1 with q-step and constants $C_0 > 0$ and $0 \leq \rho < 1$, and Y_i is $\boldsymbol{x}_{i+1}$-measurable, then (2) holds with

$$\sum_{k=1}^{q+1} \sup_{\mathcal{D} \in \mathcal{R}^{\# H_k \times (2p+1)}} |\mathbb{P}(\mathcal{X}_k \in \mathcal{D}) - \mathbb{P}(\mathcal{X}_k^{\pi} \in \mathcal{D})| \leq C_0 \times \rho^q \times n. \tag{4}$$

Moreover, when $(Y_i, \boldsymbol{x}_i)$'s are i.i.d., (4) holds with $\rho = 0$

- The term $C_0 \rho^q n$ reflects the price we pay for serial dependency

Continued

- The first result on FDR control for knockoffs inference under the setting of dependent data

- Enjoys asymptotic FDR control as long as $\log n = o(q)$

- Allows for high-dimensional time series data of feature dimensionality $p_q = O(n^K)$ with some $K > 0$ for the choice of $q = \lfloor (\log n)^{1+\eta} \rfloor$ with some $\eta > 0$

Power Analysis

- Power performance depends on signal strength measure
- Consider GLM with link function $g(\cdot)$:

$$\mathbb{E}(Y_t|\boldsymbol{x}_t) = g(\boldsymbol{x}_t^T \vec{\beta}^o), \tag{5}$$

- Use GLM-Lasso coefficients difference to construct knockoff statistics

$$W_j = |\beta_j| - |\beta_{j+p}|,$$

where

$$\begin{aligned}
&(\widehat{\beta}_1, \cdots, \widehat{\beta}_{2p})^T \\
&= \arg\min_{\vec{\beta}\in\mathbb{R}^{2p}} \left\{ \sum_{i=1}^{n} 2\left(-Y_i \times (\boldsymbol{x}_i^T, \widetilde{\boldsymbol{x}}_i^T)\vec{\beta} + r\left((\boldsymbol{x}_i^T, \widetilde{\boldsymbol{x}}_i^T)\vec{\beta}\right)\right) + n\lambda_n \sum_{j=1}^{2p} |\beta_j| \right\}
\end{aligned}$$

Conditions for power analysis

- **Condition 5.** For $\lim_{n\to\infty} k_{3n} = 0$, it holds that $\mathbb{P}\left(\sum_{j=1}^{2p} |\widehat{\beta}_j - \beta_j^*| \le c_0(\#S^*)\lambda_n\right) \ge 1 - k_{3n}$

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

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

Theoretical guarantee for power analysis

Theorem 3

It holds that for all large n,

$$\mathbb{P}\left(\{\widehat{S}=\emptyset\}\cup\left\{\frac{\#(S^*\cap\widehat{S})}{\#S^*}\geq 1-\frac{4c_0(1+q)}{k_{1n}}\right\}\right)\geq 1-(q+1)(k_{2n}+k_{3n}) \tag{6}$$

If further $\tau_1=\tau^*\times K^{-1}\times(1-4(q+1)c_0k_{1n}^{-1})$ *with some* $K>1$*, then for all large n,*

$$\mathbb{E}\left(\frac{\#(S^*\cap\widehat{S})}{\#S^*}\right)\geq\left(1-\frac{(q+1)(\tau_1+\theta_\varepsilon)K}{K-1}-(q+1)\times(k_{2n}+k_{3n})\right)\times k_{4n}, \tag{7}$$

where $\lim_{n\to\infty}k_{4n}=1$ *and*

$$\theta_\varepsilon=\inf\left\{\theta\geq 0:\max_{1\leq k\leq q+1}\mathbb{E}\left(\frac{\#(\{j:W_j^k\geq T^k\}\cap(S^*)^c)}{\#\{j:W_j^k\geq T^k\}\vee 1}\right)\leq\tau_1+\theta\right\} \tag{8}$$

- (6) shows that with asymptotic probability 1, the set of selected features is either $\emptyset$ or has TDP close to 1

- The $\widehat{S} = \emptyset$ can happen if each and every subsample yields a large number of false positives (i.e., poor FDR control), making the e-values for all features very small so that none can pass the threshold

- (7) ensures that if FDR is controlled with $(\tau_1 + \theta_\varepsilon)q = o(1)$, then the power is asymptotically one

Conclusions

Conclusions

- Suggested the new times series knockoffs inference (TSKI) procedure for feature selection with FDR control
- Developed a general theory showing that TSKI can admit asymptotic FDR control and appealing power in high-dimensional time series setting
- Justified robustness of knockoffs inference for dependent data
- Many interesting yet challenging questions remain for knockoffs inference with time series data (e.g. complicated serial dependency and nonlinear structures)

References

References

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

- Chi, C.-M., Fan, Y., Ing, C.-K. and Lv, J. (2024). High-dimensional knockoffs inference for time series data. *Manuscript*.

- Fan, Y., Gao, L. and Lv, J. (2024). ARK: robust knockoffs inference with coupling. *Manuscript*.